

Wei Gao

AI-based Image and Video Coding

Methods, Standards, and Applications

AI-based Image and Video Coding

Wei Gao

AI-based Image and Video Coding

Methods, Standards, and Applications

Wei Gao 
School of Electronic and Computer
Engineering
Peking University
Shenzhen, China

ISBN 978-981-96-7713-9 ISBN 978-981-96-7714-6 (eBook)
<https://doi.org/10.1007/978-981-96-7714-6>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

If disposing of this product, please recycle the paper.

Preface

From high-resolution imaging sensors to immersive media systems, the rapid evolution of multimedia technologies has intensified the demands for efficient visual data compression for the benefits of human and machine perception. While traditional hybrid image and video coding frameworks have served as the cornerstone for the past over three decades, their diminishing marginal returns in compression efficiency now confront fundamental limitations in the achievable coding gains and the computational complexity. This critical juncture motivates our examination of the deep learning's transformative potential in redefining compression paradigms.

This book systematically investigates the image and video data coding approaches driven by the fast-growing artificial intelligence (AI) technologies, structured through 12 interconnected chapters. The organizational logic progresses from foundational principles to implementation challenges, adhering to three core objectives: (1) To establish theoretical bridges between information theory and neural compression methods. (2) To analyze architectural innovations surpassing the conventional codec performance boundaries. (3) To address practical deployment considerations in standardization and application systems. This book illustrates the theories, methods, and applications of AI-based image and video coding. It provides a comprehensive knowledge of image and video coding powered with AI, including fundamental principles, human visual perception modeling, deep learning-based coding methods, advanced coding problems, technical standards, implementations and open-source projects, and future works. The detailed chapters are organized as follows.

Chapters 1 and 2 begin with an introduction that lays the foundation for the book. In particular, Chap. 1 outlines the core concepts of image and video coding, and Chap. 2 performs a systematic analysis of the deep learning theoretical framework highly relevant to this research field.

Chapter 3 introduces the fundamentals of quality assessment and perceptual modeling, which are essential in evaluating the image and video coding performances, and guiding perceptual optimization in coding processes.

Chapters 4 and 5 systematically investigate the utilization of deep neural networks in image and video coding, respectively, presenting the methodologies of using devised techniques to improve coding efficiency.

Chapters 6–8 discuss several advanced coding problems, focusing on 3D and multimodal coding, human and machine perception oriented coding, and compression artifacts removal, respectively.

Chapter 9 presents the recent standardization efforts on image and video coding from the working groups of different organizations, e.g., IEEE, MPEG, and JPEG, as well as the comparative analysis of emerging standards.

Chapter 10 provides the basic knowledge for implementing AI-based image and video coding techniques, including insights into hardware and software considerations.

Chapter 11 gives a review of valuable open source projects on image and video coding.

Chapter 12 looks to the future, elaborating on potential directions for research and development in AI-based image and video coding technologies.

These chapters constitute the whole book, including the fundamental principles and advanced methods for image and video coding with the aid of deep learning. We aim to provide readers with the indispensable knowledge and tools they need to contribute to this dynamic and evolving field, and also aim to accelerate the transition from the theoretical neural compression models to industrially deployable solutions.

Shenzhen, China
March 2025

Wei Gao

Acknowledgments In the journey of conducting image and video coding research and completing this book, we would like to express great appreciation to many colleagues, collaborators, and graduate students. Without their help, advice, support, and significant technical contributions, this book would not have the possibility to come into being.

The technical chapters and some of the research works presented in this book were developed in collaboration with our students and colleagues at Peking University. Particularly, we would like to thank Kangli Wang, Yuan Li, Huiming Zheng, Chenhao Zhang, Wenxu Gao, Kaiyu Zheng, Shihao Li, Zhaojian Yao, Jingxuan Su, Shunzhou Wang, Wang Liu, Shangkun Sun, Changhao Peng, Xingming Mu, Yuqi Ye, and Zhuozhen Yu. Besides these postdocs and graduate students, some other graduate students also helped edit this book, and spent much time in proofreading and figure drawing, including Qianxi Yi, Wei Cheng, Zijian Zhang, Chengwei Wang, Junhong Lin, and so on.

We also would like to thank many colleagues and friends for working together to promote the image and video coding research and its corresponding standardization efforts, including those from the School of Electronic and Computer Engineering at Peking University, the Audio Video Coding Standard (AVS) Workgroup of China, the IEEE Computer Society Data Compression Standards Committee (DCSC), as well as many other domestic and overseas universities, research institutions and high-tech companies. Many colleagues and friends have collaborated closely to develop advanced image and video coding technologies and applications, and consistently provided their substantial dedication and support.

We are also very grateful to the Springer Nature team for helping us to create this book.

Competing Interests The author has no competing interests to declare that are relevant to the content of this manuscript.

Contents

1	Introduction to Image and Video Coding	1
1.1	Basic Concept of Image and Video Data	1
1.1.1	Image Data	1
1.1.2	Video Data	2
1.2	Typical Image and Video Datasets	4
1.2.1	Image Datasets	4
1.2.2	Video Datasets	6
1.3	AI-based Compression Requirement for Image and Video	8
1.3.1	Why Compression	8
1.3.2	Why AI-based Compression	8
1.4	Image and Video Coding Performance Evaluation	9
1.4.1	Subject Quality Assessment	10
1.4.2	Objective Metrics for Quality Assessment	10
1.5	Organization of This Book	12
1.6	Summary	13
	References	14
2	Fundamentals for Deep Learning-based Image and Video Coding	23
2.1	Introduction	23
2.2	Fundamental Knowledge of Deep Learning	24
2.2.1	Convolutional Neural Networks	24
2.2.2	Transformer	26
2.3	Autoencoder and Variational Autoencoder	28
2.3.1	Autoencoder	29
2.3.2	Variational Autoencoder	30
2.4	Principles and Framework of Learned Image and Video Coding	32
2.4.1	Background Knowledge of Image and Video Compression	32
2.4.2	Bring Learning to Image and Video Compression	33

2.4.3	End-to-End Optimized Image Compression	33
2.4.4	Hyperprior for Variational Image Compression	35
2.4.5	Joint Autoregressive and Hierarchical Priors for Image Compression	36
2.4.6	Generative Adversarial Networks for Image Compression	37
2.5	Summary	38
	References	39
3	Quality Assessment and Perception Models for Image and Video	47
3.1	Introduction	47
3.2	Quality Assessment of Image and Video	48
3.2.1	Subjective Quality Assessment	49
3.2.2	Objective Quality Assessment	51
3.3	Just-Noticeable-Distortion of Image and Video	56
3.3.1	Just-Noticeable-Distortion Datasets	56
3.3.2	Pixel-Domain Just-Noticeable-Distortion Model	58
3.3.3	Frequency-Domain Just-Noticeable-Distortion Model	59
3.3.4	Image and Video Coding with Just-Noticeable-Distortion	60
3.4	Visual Attention Modeling of Image and Video	61
3.4.1	Fixation Prediction Methods	62
3.4.2	Datasets for Salient Object Detection	63
3.4.3	Image Salient Object Detection	64
3.4.4	Video Salient Object Detection	67
3.5	Summary	71
	References	72
4	Deep Learning-based Image Coding	87
4.1	Introduction	87
4.2	Typical Methods of Lossless Image Coding	88
4.3	Typical Methods of Lossy Image Coding	91
4.3.1	Framework	92
4.3.2	Core Components	97
4.4	Typical Methods of Generative Image Coding	108
4.5	Summary	110
	References	111
5	Deep Learning-based Video Coding	123
5.1	Introduction	123
5.2	Framework and Key Components	124
5.2.1	Overall Framework	124
5.2.2	Motion Estimation	125
5.2.3	Motion Compensation	129

5.2.4	Transform and Quantization	130
5.2.5	Entropy Coding	130
5.3	Typical Methods of Video Coding	131
5.4	Summary	138
	References	140
6	Deep Learning-based 3D and Multimodal Coding	151
6.1	Introduction	151
6.2	Overall Framework and Datasets	153
6.2.1	Immersive Media	153
6.2.2	Transformation	154
6.2.3	Entropy Model	155
6.2.4	Related Datasets	155
6.3	Typical Methods of 3D and Multimodal Coding	157
6.3.1	Stereo Image Compression	157
6.3.2	Multi-modality Image Compression	161
6.3.3	Light Field Image Compression	165
6.3.4	360° Image Compression	168
6.4	Summary	170
	References	170
7	Human and Machine Perception Oriented Image and Video Coding	181
7.1	Introduction	181
7.2	Typical Methods of Human Perception-based Coding	182
7.2.1	Loss Function for Human Perception	182
7.2.2	Semantic Coding for Human Perception	183
7.2.3	Regions-of-Interest-based Coding for Human Perception	184
7.2.4	Just-Noticeable-Distortion Modeling	184
7.2.5	Latent Space Masking Network	186
7.3	Machine Vision-Oriented Codec Design and Optimization	187
7.3.1	Encoder-Decoder for Machine Vision Tasks	187
7.3.2	Signal-Semantic Joint Codec	189
7.3.3	Scalable Codec for Collaborative Perception	190
7.4	Summary	192
	References	195
8	Compression Artifacts Removal for Image and Video Coding	203
8.1	Introduction	203
8.2	Compressed Image Artifacts Reduction	204
8.3	Compressed Video Artifacts Reduction	206
8.3.1	Optical Flow-based Methods	206
8.3.2	Motion-based Methods	207
8.3.3	Deformable Convolution-based Methods	209
8.3.4	Attention-based Methods	211

8.3.5	Other Methods	212
8.4	Summary	215
	References	216
9	Standards for AI-based Image and Video Coding	225
9.1	Introduction	225
9.2	Overview of International Standards	226
9.3	IEEE AI-based Image and Video Coding Standard	226
9.3.1	Introduction of IEEE AI Standards	226
9.3.2	The Development of IEEE AI Standards	227
9.3.3	Performance Comparison of Latest Reference Software	246
9.4	JPEG AI-based Image and Video Coding Standard	247
9.4.1	Introduction of JPEG AI Standards	247
9.4.2	The Development of JPEG AI Standards	248
9.4.3	Quality Metrics in JPEG AI	249
9.4.4	Performance Comparison of the Latest VM	249
9.4.5	High Level Encoder Diagram and Decoder Diagram	249
9.5	MPEG Video Coding for Machines Standard	250
9.5.1	Introduction of MPEG VCM Standards	250
9.5.2	VCM Core Experiment 1	250
9.5.3	VCM Core Experiment 2	251
9.5.4	VCM Core Experiment 3	253
9.5.5	VCM Core Experiment 4	254
9.5.6	VCM Core Experiment 5	255
9.6	MPAI End-to-End Video Coding Standard	255
9.7	Summary	256
	References	258
10	Implementations for AI-based Image and Video Coding	271
10.1	Introduction	271
10.2	Basics of Neural Network Compression	272
10.2.1	Network Pruning	272
10.2.2	Quantization	274
10.2.3	Knowledge Distillation	276
10.2.4	Neural Architecture Search	277
10.3	Software and Hardware Platforms for Acceleration	278
10.3.1	Software and Hardware Co-design	278
10.3.2	Graphics Processing Units (GPUs)	279
10.3.3	Application-Specific Integrated Circuits (ASICs)	280
10.3.4	Field-Programmable Gate Arrays (FPGAs)	281

10.3.5	AutoML Platforms	282
10.4	Lightweight Methods for Deep Compression Network	283
10.4.1	Pruning for Deep Image Compression Network	284
10.4.2	Dynamic Neural Network (DNN) for Image Compression	284
10.4.3	Quantization for Deep Image Compression Network	285
10.5	Summary	287
	References	289
11	Open Source Projects for AI-based Image and Video Coding	305
11.1	Introduction	305
11.2	CompressAI	306
11.3	OpenVVC	307
11.4	OpenDIC	307
11.5	OpenDMC	308
11.6	Summary	308
	References	312
12	Future Works for AI-based Image and Video Coding	321
12.1	Introduction	321
12.2	Future Work on Fundamentals for Deep Learning-based Image and Video Coding	322
12.3	Future Work on Quality Assessment and Perception Models for Image and Video	323
12.4	Future Work on Deep Learning-based Image Coding	324
12.5	Future Work on Deep Learning-based Video Coding	324
12.6	Future Work on Deep Learning-based 3D Image Compression	325
12.7	Future Work on Human and Machine Vision Oriented Image and Video Coding	326
12.8	Future Work on Compression Artifacts Removal for Image and Video Coding	326
12.9	Future Work on AI-based Image and Video Coding Standards	327
12.10	Future Work on Implementations for AI-based Image and Video Coding	328
12.11	Future Work on Open Source Projects for AI-based Image and Video Coding	328
12.12	Future Work on Large Language Model-based Image and Video Coding	329
12.13	Summary	330
	References	331
	Index	341

Acronyms

AE	Autoencoder
AI	Artificial Intelligence
API	Application Programming Interface
AutoML	Automated Machine Learning
AVC	Advanced Video Coding
BD-Rate	Bjøntegaard Delta Rate
BM3D	Block-Matching and 3D Filtering
CBAM	Convolutional Block Attention Module
CIFAR	Canadian Institute for Advanced Research (Dataset)
CLIC	Challenge on Learned Image Compression
CNN	Convolutional Neural Networks
COCO	Common Objects in Context
CSV	Comma Separated Values
CTC	Common Test Condition
DCT	Discrete Cosine Transform
DIV2K	Diverse 2K Resolution High-Quality Images
DL	Deep Learning
DNN	Dynamic Neural Network
DVC	Deep Video Compression
FFT	Fast Fourier Transform
FPS	Frames Per Second
GAN	Generative Adversarial Network
H.264	Advanced Video Coding (AVC)
H.265	High-Efficiency Video Coding (HEVC)
H.266	Versatile Video Coding (VVC)
HDR	High Dynamic Range
HEVC	High-Efficiency Video Coding
HVS	Human Visual System
IEEE	Institute of Electrical and Electronics Engineers
IoT	Internet of Things
IoU	Intersection over Union

IQA	Image Quality Assessment
JND	Just Noticeable Distortion
JPEG	Joint Photographic Experts Group
JPEG AI	JPEG Artificial Intelligence Standard
KNN	K-Nearest Neighbors
LF	Light Field
LIC	Learned Image Compression
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MC	Motion Compensation
ML	Machine Learning
MOS	Mean Opinion Score
MPAI	Moving Picture Artificial Intelligence
MPEG	Moving Picture Experts Group
MSE	Mean Squared Error
MS-SSIM	Multiscale Structural Similarity Index Measure
OpenDIC	Open-Source Project for Deep Learning-based Image Coding
OpenDMC	Open-Source Project for Deep Learning-based Multi-frame Coding
PCA	Principal Component Analysis
PPI	Pixels per Inch
PSNR	Peak Signal-to-Noise Ratio
RD	Rate Distortion
ResNet	Residual Network
ROI	Region of Interest
SDK	Software Development Kit
SOTA	State-of-the-Art
SSIM	Structural Similarity Index Measure
SSP	Self-Supervised Pretraining
SVM	Support Vector Machine
TPU	Tensor Processing Unit
UVG	Ultra Video Group
VAE	Variational Autoencoder
VCM	Video Coding for Machines
VGG	Visual Geometry Group (Network)
VMAF	Video Multimethod Assessment Fusion
VQA	Video Quality Assessment
YUV	Luminance and Chrominance Components (Color Encoding System)

Chapter 1

Introduction to Image and Video Coding



1.1 Basic Concept of Image and Video Data

In the era of artificial intelligence (AI), image and video are the two main data formats for visual information acquisition and representation. The multimedia and AI systems strive to provide high quality image and video data for processing and analysis to offer satisfactory visual experience and fulfill complex tasks. In this section, we briefly introduce these two visual data formats by illustrating the basic knowledge in image and video processing.

1.1.1 Image Data

The physical scene presented in the real world is projected onto the sensor array to generate an image. We use a continuous function $f(x, y)$ to represent it, which has a continuous domain and a continuous value range [1]. The domain represents the position, and the range of values represents the signal intensity. To obtain digital images, we need to sample f . Specifically, digitizing a defined domain is dubbed as sampling, and digitizing a range of values is dubbed as quantization. During the sampling process, we can consider the definition domain of f as a rectangle (m, n) , where m and n represent the width and height in pixels, with coordinates $(1, 1)$ in the lower left corner and (m, n) in the upper right corner. Generally, points with integer coordinates are sampled, and sampling points generates image pixels.

In the quantization process, we can divide the signal intensity into L intervals from low to high, and assign a fixed signal intensity value (from 0 to $L - 1$) to each interval. If the function value corresponding to the sampled integer point is in this interval, then the brightness here is quantified as this fixed value [1]. After the above sampling and quantization processes, we obtain a discrete digitalized image.

Digital images have certain requirements for m , n , and L . Generally speaking, m and n only need to be positive integers, and for hardware considerations, the discrete quantization series L is usually taken as an integer power of 2, where $L = 2^k$. An image is also named as a “ k -bit image” if it has 2^k possible signal intensity levels. The image obtained through the above process becomes a single-channel image. The color image is generally made up of three single-channel red, green, and blue (RGB) images, and the values of each single-channel image record the corresponding color intensity values [1]. In addition, there are also some commonly used concepts, such as the resolution $x \times y$ of a certain image, which means that the image has x pixels in each row and y pixels in each column. Moreover, the Pixels Per Inch (PPI) represents the number of pixels per inch on the screen.

1.1.2 Video Data

YUV is a commonly used raw video data format, and most of the bitstreams output by video processing chips are in the format of YUV data. Y represents luminance information, while U and V represent chroma information. Chroma determines two aspects of color, i.e., hue and saturation [2].

Similar to the well-known RGB shown in Fig. 1.1, YUV is another color representation mainly used in television systems. Compared to RGB, YUV has the following advantages: (1) The two kinds of components, i.e., luminance (Y) and chroma (UV), can be effectively separated, and it can also display the image at grayscale without UV information. Therefore, YUV provides the compatibility of colorful and grayscale displays. (2) Differently from RGB, YUV can reduce the transmission and storage overhead by adopting UV sampling. As a result, the decreased bandwidth demand can be seen if using YUV to transport video signals. YUV and RGB are interchangeable for different practical application demands [2].

The key advantage of YUV lies in its capability to maintain perceptual quality while allowing reduced chroma channel sampling rates relative to the luminance (Y) channel. This characteristic is commonly expressed through the A:B:C notation, which specifies the frequency ratios of U and V components to Y [2]. To meet different practical needs, YUV has the following sampling methods:

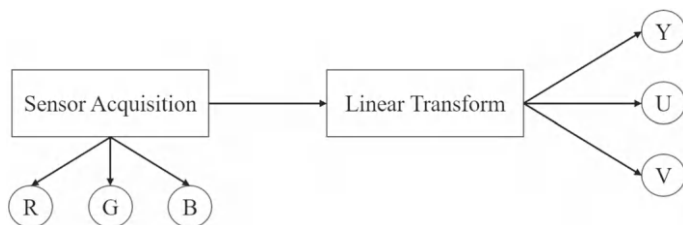


Fig. 1.1 The color space transformation between RGB and YUV. *Source* Author

- **YUV 4:4:4 Sampling.** This method means no downsampling of the chroma channels [3]. YUV 4:4:4 sampling means that the sampling ratios of the Y, U, and V components are equal [2]. Hence, each pixel's information is complete (8 bits of each component, which is one byte) in the generated image.
- **YUV 4:2:2 Sampling.** In 4:2:2 chroma subsampling, a 2:1 horizontal subsampling ratio is applied while maintaining full vertical resolution [3]. This configuration preserves four luminance (Y) components for each pair of chrominance (U/V) components within every scanline. YUV 4:2:2 sampling means that the UV component is the half of the Y component sampling, and the Y and UV components are sampled with 2:1 ratio [2]. If there are 4 pixels in the horizontal direction, then 4 Y components are sampled and only 2 U and V components are sampled. Every pixel sampled will have its Y component sampled, while the U and V components will be alternately sampled. Every two pixels share the same U and V components, and thus the storage space can be saved. Consequently, this representation can also benefit the data transmission.
- **YUV 4:2:0 Sampling.** The 4:2:0 chroma subsampling format implements a 2:1 resolution reduction in both horizontal and vertical dimensions for chrominance components [3]. YUV 4:2:0 sampling does not mean only sampling the U component without sampling the V component. Instead, it refers to scanning only one chroma component (U or V) in each row [2]. For example, we sample Y and U in a 2:1 manner during the scan of the first row, and we sample Y and V in a 2:1 manner during the second row. For each chroma component, its horizontal and vertical sampling is 1:2 compared to the Y component. There are two common variants of YUV 4:2:0 sampling. They were used in the earlier video coding standards, including MPEG-1, MPEG-2, H.261, and H.263. From the mapped pixels, we can see that every four Y components share a set of UV components and are distributed in a 2×2 small grid. Compared to the YUV 4:2:2, YUV 4:2:0 can save more space. It can be seen that YUV 4:2:0 sampled images save half of the storage space compared to RGB model images, so it is also a mainstream sampling method. We can easily find that the storage overhead of YUV 4:2:0 image can have 1/2 of the original cost. Therefore, for various applications, this simplified color signal representation has been adopted.
- **YUV 4:1:1 Sampling.** The 4:1:1 chroma subsampling scheme implements a 4:1 horizontal chroma downsampling ratio while preserving vertical resolution. In this configuration, each scanline retains four Y luminance samples per single U/V chrominance component [3]. Despite being a valid color sampling standard, 4:1:1 remains relatively uncommon compared to other chroma subsampling schemes and is not discussed in detail in this book.

Table 1.1 General information of image and video datasets

Media	Dataset	Year	Institution
Image	ImageNet	2012	Stanford University, USA
	DIV2K	2017	ETH Zurich, Swiss
	COCO	2017	Microsoft inc., USA
	CLIC	2020	Challenge on learned image compression
	Kodak Photo CD	1999	Kodak inc., USA
Video	Vimeo-90K	2019	Google Research, USA
	HEVC CTC	2013	HEVC
	MCL-JCV	2016	University of Southern California, USA
	UVG	2020	Tampere University of Technology, Finland

Source Author

1.2 Typical Image and Video Datasets

In the field of AI-based image and video coding, the quality and diversity of datasets play a pivotal role in training and evaluating the devised algorithms. Datasets serve as the foundation upon which researchers and engineers build models that push the boundaries of compression efficiency and visual fidelity. In this section, we discuss some of the key datasets that have contributed significantly to advance this field. All datasets are categorized in Table 1.1 with general information.

1.2.1 Image Datasets

In the field of AI-based image coding, the image datasets form the bedrock upon which groundbreaking algorithms are honed, tested, and refined. As researchers and practitioners strive for innovation in image compression, they rely on diverse and meticulously curated datasets to propel their work forward. There are several typical image datasets, which are from various original fields and finally can be applied into image compression.¹ These datasets have common characteristics:

¹ Resource: https://en.wikipedia.org/wiki/List_of_datasets_in_computer_vision_and_image_processing.

1. **Diversity.** An excellent dataset with diversity encompasses a wide range of scenes, contexts, and challenges, mirroring the complexity of the real world. The inclusion of varied content ensures that the developed algorithms using these datasets are not only proficient but also adaptable to the unpredictable nuances of different visual scenarios.
2. **Resolution.** Resolution, a critical aspect of image datasets, dictates the level of details captured in each image. High-resolution datasets, such as those images in 2K or 4K resolutions, become crucial in the compression task. Along with the increased demand of human perception, the training of AI-based compression algorithms need high-resolution datasets.

ImageNet Dataset

ImageNet² is one of the largest and most influential image datasets [4]. It is crucial for the development and evaluation of various image-related tasks and models. It is introduced by researchers at Stanford University and Princeton University, and has been widely used in the computer vision and image processing community. With its extensive collection, ImageNet facilitates the development of image compression algorithms that can generalize across a broad spectrum of visual categories.

DIV2K Dataset

The inception of DIV2K,³ the acronym for “DIVERse 2K resolution high quality images”, emerged from the recognition that achieving superior image compression necessitates a diverse and challenging set of images. Released explicitly for the task of single image super-resolution, DIV2K has evolved into a golden standard for evaluating the efficacy of various compression algorithms. Comprising images at a resolution of 2048×1536 pixels, DIV2K is a collection of high-resolution images. It has diverse scenes and encompasses a myriad of content and imaging conditions. Its diversity is instrumental in pushing the boundaries of AI-based compression techniques. From urban landscapes to natural wonders, DIV2K ensures that the compression algorithms under scrutiny are robust and adaptable to real-world scenarios.

Common Objects in Context (COCO) Dataset

The Common Objects in Context (COCO)⁴ dataset emerges as a pivotal resource in the field of AI-based image coding, offering a comprehensive collection of images designed to encapsulate the complexity of real-world scenes [5]. Its development, initiated by Microsoft Research, aimed to propel research in object recognition and scene understanding at first. COCO has become a common dataset for various tasks, including image compression. It is characterized by the presence of multiple objects within intricate scenes, challenging compression algorithms to handle diverse visual elements simultaneously.

² <https://www.image-net.org>.

³ <https://data.vision.ee.ethz.ch/cvl/DIV2K>.

⁴ <https://cocodataset.org>.

Challenge in Learned Image Compression (CLIC) Dataset

The Challenge in Learned Image Compression (CLIC)⁵ has become one of the most well-known image compression challenges around the world. The challenge provides a dataset for lossy image compression, and therefore it is dubbed as CLIC [6]. The CLIC dataset contains a series of professional and mobile images. By focusing on challenges encountered in real-world scenarios, it pushes researchers to develop algorithms that prioritize perceptual quality, aligning with the evolving needs of practical applications.

Kodak PhotoCD Dataset

The Kodak PhotoCD⁶ dataset is a collection of high quality images, and is frequently used as a test dataset to compare the rate distortion performance of image compression algorithms [7]. The dataset typically consists of 24 high-resolution images captured using Kodak PhotoCD technology. These high-resolution images with varying content and high fidelity offer an unique benchmark for assessing the performance of compression algorithms. It is recognized as a common test dataset in image compression.

1.2.2 Video Datasets

The video datasets facilitate researchers and engineers to develop better video compression algorithms, fostering a deeper understanding of the intricacies of moving visual narratives.⁷ The video datasets have the following common characteristics:

1. **Temporal Complexity.** Temporal complexity is an unique characteristic of video datasets, which comes from the interplay of frames. It introduces challenges to motion estimation and compensation, optical flow estimation, and video enhancement. Accurate modeling of temporal corrections is crucial for efficient video coding.
2. **Resolution and Frame Rates.** Similar to image datasets, the resolution of video datasets also plays an important role for compression tasks, as well as frame rates. High-resolution video datasets with varying frame rates can pose great challenges in video coding.

Vimeo-90K Dataset

The Vimeo-90K⁸ dataset is originally established for video enhancement, such as frame interpolation and video denoising, but it can be also used in the realm of video coding to promote the optimization of compression efficiency. It captures diverse

⁵ <https://compression.cc>.

⁶ <https://r0k.us/graphics/kodak>.

⁷ Resource: https://en.wikipedia.org/wiki/List_of_datasets_in_computer_vision_and_image_processing.

⁸ <http://toflow.csail.mit.edu>.

temporal nuances, and its temporal diversity also provide sufficient challenges for video coding optimization [8–37].

HEVC CTC Dataset

High Efficiency Video Coding (HEVC), which is a video coding standard jointly developed by the ITU-T Video Coding Experts Group (VCEG) and the ISO/IEC Moving Picture Experts Group (MPEG). The HEVC common test condition (CTC)⁹ dataset encompasses a wide range of video contents, including diverse scenarios, resolutions, and motion complexities. The dataset contains several types of sequences with different resolutions, and usually the video sequences belonging to Class B, C, D, E are used in AI-based video coding.

MCL-JCV Dataset

Developed by the Media Communications Lab, the MCL-JCV¹⁰ dataset gives an unique challenge for researchers and practitioners to optimize the video coding algorithms. This dataset encapsulates videos with diverse scenarios, resolutions, and motion complexities. Beyond traditional video datasets, this dataset also provides the Just Noticed Distortion (JND) notation for human perception oriented video coding optimization research [28, 38, 39].

UVG Dataset

The Ultra Video Group (UVG)¹¹ dataset has 16 video sequences with the frame rate of either 50 fps or 120 fps. Constructed by the Ultra Video Group in Tampere University, this dataset encapsulates videos from real-world scenarios, including CityAlley, Bosphorus, etc. This dataset distinguishes itself as the premier source of auxiliary 4K@120 fps sequences.

PP8K Dataset

The PP8K¹² dataset is jointly developed by Peking University and Peng Cheng Laboratory, which provides 8 K image and video sources for video coding and image processing research, as well as the saliency annotations for salient object detection research [26]. Therefore, it has two parts, i.e., 8 K Common Raw Format Video (PP8K-CRFV) sequences and 8K Video Salient Object Detection (PP8K-VSOD) sequences. This dataset contributes the ultra high definition 8K images and videos to facilitate the video coding research.

⁹ <https://hevc.hhi.fraunhofer.de>.

¹⁰ <https://mcl.usc.edu/mcl-jcv-dataset>.

¹¹ <https://ultravideo.fi/dataset.html>.

¹² <https://openi.pcl.ac.cn/OpenDatasets/PP8K>.

1.3 AI-based Compression Requirement for Image and Video

1.3.1 Why Compression

The image and video compression task endeavors to minimize the data size and maximize the signal quality [8–75]. Image and video coding plays an indispensable role in modern information systems, addressing the escalating challenges associated with data storage, transmission, and processing [38, 39, 74–80]. Cloud storage services, with their tiered pricing based on data usage, benefit substantially from image and video data compression.

To deliver an uncompressed 4K/60fps video sequence with 8 bits for each color component, the transmission bandwidth should be above $3840 \times 2160 \times 8 \times 3 \times 60 \approx 12$ Gbps. However, the available Internet bandwidth is usually less than 10Mbps. The significant gap between the raw data size and the available bandwidth impels the developments of high efficiency video coding methods. Up to now, the typical video compression standards can achieve over 400:1 compression ratio, which significantly relieves the bandwidth burden [74]. To deliver high definition videos for remote conference, one-hour video call can consume up to several GBs of uncompressed data. By adopting advanced coding technologies, the data size can be reduced significantly to realize high quality visual communications even in bandwidth-limited conditions [35, 81]. These representative examples demonstrate the practical needs of efficient compression algorithms to optimize the bandwidth burden and the visual experience [29, 82].

Although enhancement algorithms [83–112] can also bring quality improvements, we usually deem this process as a relatively independent step in the overall image and video processing system, and only discuss the compression efficiency optimization of the encoder and decoder instead. In addition, high quality visual content after compression and decompression can benefit both human visual experience [82, 113–115] and machine perception performances [56, 81, 116–151].

1.3.2 Why AI-based Compression

As the digital ecosystem expands, the demand for efficient data compression solutions has been strengthened [152–163]. The integration of AI theories and methods into data coding algorithms has demonstrated how AI-driven solutions outperform traditional methods by further effectively optimizing compression efficiency and visual quality of experience. Traditional methods rely on the standardized algorithms following the hybrid coding structure, which may restrict the optimization potentials of rate distortion performance. In contrast, AI-based image and video coding driven by advanced neural networks can adaptively learn intricate patterns with strong fitting power, achieving superior compression ratios while preserving the signal quality.

The current advanced neural network-based image and video compression models can consistently outperform traditional methods significantly. Considering the fast developments of AI technologies and the popularization of imaging sensors, both the improved algorithms and the increase data can bring more performance gains to image and video coding. Therefore, it has been widely recognized that AI-based compression methods have much more potentials of rate distortion optimization than the traditional non-learning-based methods.

The goal of traditional image and video coding optimization efforts is to improve the signal fidelity or the visual experience of reconstructed image and video under the condition of the same bit consumption. With the deepening of investigations for machine vision tasks, the optimization objective can also be the machine perception performances after compression, including classification, segmentation, detection, etc. The quality of compressed images and videos can usually influence the accuracy of machine vision algorithms, and therefore optimization efforts can be focus on the rate-accuracy tradeoff. These new paradigms for image and video coding have been dubbed as image coding for machine and video coding for machine, respectively. Furthermore, in the practical machine vision systems, the extracted features with less data amount are more important than the raw image and video signals for the subsequent machine analysis tasks. Hence, Gao et al. [164] proposed the concept of “digital retina” for AI cameras with the aim of constructing a powerful visual system for smart cities [46], where the dual stream architecture can fulfill both compressions of video data and video features. AI cameras can be deemed as a specific example of the research of coding for machines. Compared with the traditional non-learning-based methods, the adoption of AI-based compression can much better facilitate the joint optimization of multiple objectives, e.g., consumed bit rate, signal fidelity quality, machine analysis task performance, etc.

Considering the potentials and advantages of AI-based image and video coding methods, researcher has invested lots of efforts into deep learning algorithms, software and hardware implementations, and real-time systems. Additionally, the academia and industry communities also have promote the developments of AI-based image and video coding standards, as well as the standards for the emerging technologies of image and video coding for machine.

1.4 Image and Video Coding Performance Evaluation

Quality assessment (QA) is an important research problem [82, 113–115, 138, 165–169] in the field of image and video coding. It provides a way of measuring the quality of distorted image and video content. QA can be broadly categorized into two types, i.e., subjective quality assessment and objective quality assessment. Subjective QA collects human judging scores by conducting practical experiments, while objective QA relies on the designed model. By obtaining the mean opinion scores (MOS) for different samples, the subjective QA is the most accurate approach to evaluate image

and video quality, but laborious. Instead, the objective QA aims to devise accurate objective models to predict the human judging scores.

1.4.1 Subject Quality Assessment

Subjective quality assessment of images and videos focuses on evaluating the visual quality by the means of human judging. In the context of images, it involves assessing factors such as texture and structure, while for videos, motion smoothness and temporal consistency are also considered. Typically, human observers rate the quality to give MOS scores [170] based on pairwise comparison or only subjective visual experience depending on whether the pristine image or video is available or not. Since different images and videos have distinct characteristics, subjective assessments take into account various aspects of visual content. Despite its reliability, subjective assessment is resource-intensive, leading to the development of objective metrics that aim to replicate human judgments in a more efficient and automated manner.

1.4.2 Objective Metrics for Quality Assessment

Objective metrics provide quantitative measurements to evaluate the quality of compressed images and videos. Researchers have developed many different evaluation methods for image and video quality. Herein, three fundamental objective metrics are usually adopted for diverse application scenarios, i.e., Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Metric (SSIM) and its multi-scale version.

Mean Squared Error (MSE). For signal quality evaluation, MSE serves as a foundational criterion. It evaluates image fidelity by measuring the arithmetic mean of squared discrepancies between the original and reconstructed pixel values. It provides a simple but effective measure for the reconstruction error as the following equation:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (x_i - y_i)^2, \quad (1.1)$$

where N indicates the amount of pixels, x_i and y_i characterize the luminance values of the source and processed image matrices.

Peak Signal-to-Noise Ratio (PSNR). As a widely used quality metric, PSNR evaluates the fidelity of encoded data by calculating the logarithmic ratio between the maximum attainable signal intensity and the distortion-induced interference. The PSNR value, with the unit of decibels (dB), reflects the proximity of processed data to its source data, with higher values indicating less perceptual degradation:

$$\text{PSNR} = 10 \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right), \quad (1.2)$$

where MSE is defined as the above, and MAX is the maximum possible pixel value.

Structural Similarity Index Measurement (SSIM). SSIM [171] establishes a perceptual quality framework by analyzing the similarity of luminance, contrast, and structure between the pristine and compressed images. This multi-dimensional approach addresses the limitations of pixel-based metrics like MSE and PSNR in accurately simulating human visual perception. This characteristic makes SSIM a more perception model which can better align with the human visual mechanism. SSIM evaluates the distorted images from three aspects including luminance, contrast, and structure. Luminance is measured by the average intensity of given image, and the luminance component of SSIM is denoted as $l(x, y)$. Contrast reflects the difference in intensity between the images. The contrast value is calculated by the standard deviation σ of the image pixel values, and the contrast component of SSIM can be denoted as $c(x, y)$. The structure component is computed by the covariance of the images, and denoted as $s(x, y)$. The overall SSIM index is computed as a combination of the three components:

$$\text{SSIM}(x, y) = [l(x, y)^\alpha \cdot c(x, y)^\beta \cdot s(x, y)^\gamma]. \quad (1.3)$$

These coefficients (α , β , and γ) regulate the weights of luminance, contrast, and structure, respectively. Typically, they are set to 1 for equal importance, and then SSIM is expressed as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \quad (1.4)$$

where two C parameters are applied with small values.

Multi-Scale Structural Similarity Index Measurement (MS-SSIM). The MS-SSIM [172] is an extension of the original SSIM. It has been specifically designed to capture structural information at multiple scales, thereby offering a more comprehensive and accurate evaluation of the perceptual quality of compressed images. By addressing limitations associated with the single-scale SSIM, MS-SSIM is widely recognized for its enhanced robustness in assessing image quality. MS-SSIM also utilizes the definition of luminance l , contrast c and structure s . Given referenced image x and distorted image y , MS-SSIM is computed by measuring all the M scales:

$$\text{MS-SSIM}(\mathbf{x}, \mathbf{y}) = [l_M(\mathbf{x}, \mathbf{y})]^{\alpha_M} \cdot \prod_{j=1}^M [c_j(\mathbf{x}, \mathbf{y})]^{\beta_j} [s_j(\mathbf{x}, \mathbf{y})]^{\gamma_j}, \quad (1.5)$$

where α_M , β_j , and γ_j can be adjusted for the relative importance of different components. $\alpha_j = \beta_j = \gamma_j$ for all j is usually applied for simplicity.

Rate-Distortion Curve (RD-Curve). RD-Curve [173] is a basic concept in image and video compression, illustrating the tradeoff between the consumed bit rate and the signal distortion. The curve plots the coding results with different bit rates and the corresponding quality measured by PSNR or MSE. Compression efficiency can be quantified through RD-Curve, where the encoder with lower bit rate consumption and higher quality score is a better coding scheme. Plotting RD-Curve is an effective way to compare different coding solutions. It has become the golden criterion for evaluating different image and video coding methods, where we usually plot four, five or even more dots with rate and distortion results. As the development of visual data coding technologies in the past over thirty years, the coding efficiency as indicated by RD-Curve has witnessed a dramatic improvement to reduce the bandwidth demand with the same quality.

Bjontegaard Delta Rate (BD-Rate). The community has widely utilized BD-Rate [174] to measure how many bit rates can be saved when adopting a more advanced coding scheme. BD-Rate is used to compare every two different coding methods. The single rate-distortion data pair can not accurately gauge the encoding strategy. Therefore, the emergence of BD-Rate can solve the problem of comprehensively and fairly evaluating image and video coding methods. This approach is easy to implement by testing four or five different settings to generate rate-distortion data pairs belonging to the commonly-seen bandwidth or quality range. The negative BD-Rate results with unit of % means that the test method is better than the baseline method by saving bit rates. The use of BD-Rate makes the comparison of different coding methods much more easily and fairly, and it has become the indispensable metric in the academia, industry, and standardization communities for image and video coding.

Bjontegaard Delta PSNR (BD-PSNR). Besides the above BD-Rate, BD-PSNR [174] is another similar metric used for comparing performances of image and video coding methods. BD-PSNR and BD-Rate has the relationship of one-to-one correspondence. When comparing two different coding methods, these two metrics can be simultaneously generated. BD-PSNR measures the average PSNR quality gain with the unit of dB between two codecs with the same bit rate. The positive value of BD-PSNR means the test method outperforms the baseline method. At least four or five rate-distortion data pairs should be firstly collected for the test and baseline methods. Similarly with the computing manner of BD-Rate, we can get the BD-PSNR result to indicate how many PSNR quality gains can be obtained with the same bit rate consumption. BD-PSNR has also been widely used as BD-Rate.

1.5 Organization of This Book

The structure of this book is organized with 12 chapters to illustrate the detailed theories and methods of AI-based image and video coding.

Introduction. In this chapter and Chap. 2 begin with an introduction that lays the foundation for the book. In particular, In this chapter outlines the core concepts

of image and video coding, and Chap. 2 performs a systematic analysis of the deep learning theoretical framework highly relevant to this research field.

Human visual perception modeling. Chapter 3 introduces the fundamentals of quality assessment and perceptual modeling, which are essential in evaluating the image and video coding performances, and guiding perceptual optimization in coding processes.

Deep learning-based coding methods. Chapters 4 and 5 systematically investigate the utilization of deep neural networks in image and video coding, respectively, presenting the methodologies of using devised techniques to improve coding efficiency.

Advanced coding problems. Chapters 6, 7, and 8 discuss several advanced coding problems, focusing on 3D and multimodal coding, human and machine perception oriented coding, and compression artifacts removal, respectively.

Technical standards. Chapter 9 presents the recent standardization efforts on image and video coding from the working groups of different organizations, e.g., IEEE, MPEG, JPEG, etc., as well as the comparative analysis of emerging standards.

Implementations and open source projects. Chapter 10 provides the basic knowledge for implementing AI-based image and video coding techniques, including insights into hardware and software considerations. Chapter 11 gives a review of valuable open-source projects on image and video coding.

Future works. Chapter 12 looks to the future, elaborating on potential directions for research and development in AI-based image and video coding technologies.

These chapters constitute the whole book, including the fundamental principles and advanced methods for image and video coding with the aid of deep learning. We aim to provide readers with the indispensable knowledge and tools they need to contribute to this dynamic and evolving field, and also aim to accelerate the transition from the theoretical neural compression models to industrially deployable solutions.

1.6 Summary

This chapter presents a systematic overview of image and video data fundamentals. Beginning with an examination of image and video data formats, we subsequently analyze the distinct characteristics through detailed technical illustrations. A survey of benchmark datasets is provided, acknowledging their pivotal role in advancing deep learning method research for image and video coding. The discussion then focuses on AI-based compression requirements, addressing the persistent challenges in the field and the growing significance of learning-based approaches. Furthermore, we introduce the established quality metrics and evaluation methodologies for compression performance. This chapter concludes with an outline of the book's organizational structure.

Exercises

1. Please answer how many possible grayscale levels does a “5-bit image” have?
2. Please represent the pixel in RGB, if the sampling value of a pixel is $\begin{bmatrix} 100 \\ 100 \\ 100 \end{bmatrix}$ in YUV.
3. Assuming there is a video with 300 frames, each frame has a resolution of 2560×1440 . The first 100 frames are stored using YUV 4:4:4 sampling, the middle 100 frames are stored using YUV 4:2:2 sampling, and the last 100 frames are stored using YUV 4:2:0 sampling. Please answer how much storage space does this video occupy?
4. Please answer what are the five steps of subjective a QA experiment?
5. Suppose you have calculated the MSE of two images, which is equal to 20. The peak value of the pixel is 255. Please answer what is the PSNR of the degraded image?
6. Please describe the difference between lossless and lossy image compression. Give an example of each.
7. Please explain the concept of temporal redundancy in video compression and how it is reduced.
8. Please answer what is the purpose of intra prediction in video coding, and how does it differ from inter prediction?
9. Please describe how does quantization affect the quality of compressed video, and what is the purpose of quantization parameter (QP)?
10. Please explain the role of the discrete cosine transform (DCT) in video compression.
11. Please answer what is Scalable Video Coding, and what are its potential applications?

References

1. R.C. Gonzalez, *Digital Image Processing* (Pearson Education India, 2009)
2. Wikipedia, Y'UV, [Online], <https://en.wikipedia.org/wiki/Y'UV>
3. Wikipedia, “Chroma_subsampling,” [Online], https://en.wikipedia.org/wiki/Chroma_subsampling
4. A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**, 1106–1114 (2012)
5. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in *European Conference on Computer Vision* (2014), pp. 740–755
6. T. George, S. Wenzhe, T. Radu, T. Lucas, B. Johannes, A. Eirikur, J. Nick, M. Fabian, Workshop and challenge on learned image compression (clic2020), CVPR, 2020. [Online]. Available: <http://www.compression.cc>
7. Kodak photocd dataset, 1999. [Online]. Available: <http://r0k.us/graphics/kodak/>

8. Z. Yang, W. Gao, G. Li, Y. Yan, Sur-driven video coding rate control for jointly optimizing perceptual quality and buffer control. *IEEE Trans. Image Process.* (2023)
9. F. Shen, Z. Cai, W. Gao, An efficient rate control algorithm for intra frame coding in avs3, in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2021), pp. 3164–3169
10. H. Yuan, W. Gao, J. Wang, Dynamic computational resource allocation for fast inter frame coding in video conferencing applications, in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6
11. W. Gao, Q. Jiang, R. Wang, S. Ma, G. Li, S. Kwong, Consistent quality oriented rate control in hevc via balancing intra and inter frame coding. *IEEE Trans. Ind. Inform.* **18**(3), 1594–1604 (2021)
12. H. Yuan, W. Gao, A new coding unit partitioning mode for screen content video coding, in *Proceedings of the 2021 5th International Conference on Digital Signal Processing* (2021), pp. 66–72
13. W. Gao, On the performance evaluation of state-of-the-art rate control algorithms for practical video coding and transmission systems, in *Proceedings of the 2020 4th International Conference on Video and Image Processing* (2020), pp. 179–185
14. W. Gao, S. Kwong, Q. Jiang, C.-K. Fong, P.H. Wong, W.Y. Yuen, Data-driven rate control for rate-distortion optimization in hevc based on simplified effective initial qp learning. *IEEE Trans. Broadcast.* **65**(1), 94–108 (2018)
15. W. Gao, A multi-objective optimization perspective for joint consideration of video coding quality, in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (IEEE, 2019), pp. 986–991
16. W. Gao, S. Kwong, Y. Jia, Joint machine learning and game theory for rate control in high efficiency video coding. *IEEE Trans. Image Process.* **26**(12), 6074–6089 (2017)
17. W. Gao, S. Kwong, Y. Zhou, H. Yuan, Ssim-based game theory approach for rate-distortion optimized intra frame ctu-level bit allocation. *IEEE Trans. Multimedia* **18**(6), 988–999 (2016)
18. W. Gao, S. Kwong, H. Yuan, X. Wang, Dct coefficient distribution modeling and quality dependency analysis based frame-level bit allocation for hevc. *IEEE Trans. Circuits Syst. Video Technol.* **26**(1), 139–153 (2015)
19. W. Gao, S. Kwong, Phase congruency based edge saliency detection and rate control for perceptual image and video coding, in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2016), pp. 000 264–000 269
20. H. Yuan, W. Gao, Openfastvc: An open source library for video coding fast algorithm implementation, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9660–9663
21. H. Yuan, W. Gao, S. Ma, Y. Yan, Divide-and-conquer-based rdo-free cu partitioning for 8k video compression. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(4), 1–20 (2024)
22. L. Tao, W. Gao, A hardware implementation of entropy encoder for 8k video coding, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
23. Y. Guo, W. Gao, S. Ma, G. Li, Accelerating transform algorithm implementation for efficient intra coding of 8k uhd videos. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* **18**(4), 1–20 (2022)
24. Z. Cai, W. Gao, Efficient fast algorithm and parallel hardware architecture for intra prediction of avs3, in *IEEE International Symposium on Circuits and Systems (ISCAS)* (IEEE, 2021), pp. 1–5
25. W. Gao, H. Yuan, Y. Guo, L. Tao, Z. Cai, G. Li, Openhardwarevc: an open source library for 8k uhd video coding hardware implementation, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 7339–7342
26. W. Gao, H. Yuan, G. Liao, Z. Guo, J. Chen, Pp8k: A new dataset for 8k uhd video compression and processing. *IEEE MultiMedia* (2023)
27. W. Liu, W. Gao, G. Li, S. Ma, T. Zhao, H. Yuan, Enlarged motion-aware and frequency-aware network for compressed video artifact reduction. *IEEE Trans. Circuits Syst. Video Technol.* (2024)

28. Y. Wu, L. Xie, S. Sun, W. Gao, Y. Yan, Adaptive intra period size for deep learning-based screen content video coding, in *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2024), pp. 1–6
29. H. Zheng, W. Gao, End-to-end rgb-d image compression via exploiting channel-modality redundancy. *Proc. AAAI Conf. Artif. Intell.* **38**(7), 7562–7570 (2024)
30. L. Tao, W. Gao, G. Li, C. Zhang, Adanic: towards practical neural image compression via dynamic transform routing, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 16 879–16 888
31. Y. Wu, W. Gao, End-to-end lossless compression of high precision depth maps guided by pseudo-residual (2022). [arXiv:2201.03195](https://arxiv.org/abs/2201.03195)
32. Y. Wu, Z. Qi, H. Zheng, L. Tao, W. Gao, Deep image compression with latent optimization and piece-wise quantization approximation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 1926–1930
33. W. Gao, L. Tao, L. Zhou, D. Yang, X. Zhang, Z. Guo, Low-rate image compression with super-resolution learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2020), pp. 154–155
34. W. Gao, S. Sun, H. Zheng, Y. Wu, H. Ye, Y. Zhang, Opendmc: an open-source library and performance evaluation for deep-learning-based multi-frame compression, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9685–9688
35. W. Gao, H. Zheng, C. Zhang, K. Zheng, Z. Yu, Y. Li, H. Ye, Y. Zhang, Opendic: an open-source library and performance evaluation for deep-learning-based image compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 202–11 205
36. Y. Li, W. Gao, G. Li, S. Ma, Saliency segmentation oriented deep image compression with novel bit allocation. *IEEE Trans. Image Process.* **34**, 16–29 (2025)
37. C. Zhang, W. Gao, Learned rate control for frame-level adaptive neural video compression via dynamic neural network, in *European Conference on Computer Vision* (Springer, Berlin, 2024)
38. L. Xie, W. Gao, Learningpcc: a pytorch library for learning-based point cloud compression, *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 234–11 238
39. L. Xie, W. Gao, Pchmvision: An open-source library of point cloud compression for human and machine vision,” in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 239–11 243
40. W. Gao, S. Kwong, H. Sang, Low-cost memory data scheduling method for reconfigurable fft bit-reversal circuits. *Electron. Lett.* **51**(3), 217–219 (2015)
41. T. Qin, G. Li, W. Gao, S. Liu, Multi-grained point cloud geometry compression via dual-model prediction with extended octree. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(9), 1–30 (2024)
42. Y. Shao, W. Gao, S. Liu, G. Li, Advanced patch-based affine motion estimation for dynamic point cloud geometry compression. *Sensors* **24**(10), 3142 (2024)
43. Y. Shao, F. Song, W. Gao, S. Liu, G. Li, Texture-guided graph transform optimization for point cloud attribute compression. *Appl. Sci.* **14**(10), 4094 (2024)
44. Y. Shao, X. Yang, W. Gao, S. Liu, G. Li, 3d point cloud attribute compression using diffusion-based texture-aware intra prediction. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
45. J. Zhang, Y. Chen, G. Liu, W. Gao, G. Li, Efficient point cloud attribute compression framework using attribute-guided graph fourier transform, in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2024), pp. 8426–8430
46. W. Gao, H. Yuan, G. Li, Z. Li, H. Yuan, Low complexity coding unit decision for video-based point cloud compression. *IEEE Trans. Image Process.* **33**, 149–162 (2023)
47. Y. Shao, G. Li, Q. Zhang, W. Gao, S. Liu, Nonrigid registration-based progressive motion compensation for point cloud geometry compression. *IEEE Trans. Geosci. Remote Sens.* (2023)

48. F. Song, G. Li, X. Yang, W. Gao, S. Liu, Block-adaptive point cloud attribute coding with region-aware optimized transform. *IEEE Trans. Circuits Syst. Video Technol.* (2023)
49. Y. An, Y. Shao, G. Li, W. Gao, S. Liu, A fast motion estimation method with hamming distance for lidar point cloud compression, in *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2022), pp. 1–5
50. H. Yuan, W. Gao, G. Li, Z. Li, Rate-distortion-guided learning approach with cross-projection information for v-pcc fast cu decision, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 3085–3093
51. F. Song, G. Li, W. Gao, T.H. Li, Rate-distortion optimized graph for point cloud attribute coding. *IEEE Signal Process. Lett.* **29**, 922–926 (2022)
52. F. Song, G. Li, X. Yang, W. Gao, T. H. Li, Fine-grained correlation representation for graph-based point cloud attribute compression, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
53. F. Shen, W. Gao, A rate control algorithm for video-based point cloud compression, in *2021 International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2021), pp. 1–5
54. F. Song, Y. Shao, W. Gao, H. Wang, T. Li, Layer-wise geometry aggregation framework for lossless lidar point cloud compression. *IEEE Trans. Circuits Syst. Video Technol.* **31**(12), 4603–4616 (2021)
55. Z. Cai, W. Gao, G. Li, W. Gao, Distortion propagation model-based v-pcc rate control for 3d point cloud broadcasting. *IEEE Trans. Broadcast.* 1–13 (2024)
56. Y. Li, Z. Yu, X. Chen, S. Wang, H. Yuan, W. Gao, Zero-shot quantization for large-kernels via shape-based distribution and diversity self-distillation, in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2025)
57. X. Chen, Y. Wang, Y. Li, X. Ling, M. Li, R. Liu, M. Ouyang, K. Zhao, T. Guan, Y. He, Low bit-width zero-shot quantization with soft feature-infused hints for iot systems. *IEEE Internet Things J.* (2024)
58. C. Peng, W. Gao, Laplacian matrix learning for point cloud attribute compression with ternary search-based adaptive block partition, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 10 412–10 420
59. Y. Wang, W. Gao, X. Mu, H. Yuan, Rate control optimization for joint geometry and attribute coding of lidar point clouds, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
60. T. Guo, H. Yuan, T. Wang, W. Gao, Graph filter-based fast motion matching for inter frame coding of mpeg g-pcc, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 1151–1155
61. L. Xie, W. Gao, H. Zheng, G. Li, Spcgc: scalable point cloud geometry compression for machine vision, in *2024 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2024), pp. 17 272–17 278
62. L. Xie, W. Gao, H. Zheng, H. Ye, Semantic-aware visual decomposition for point cloud geometry compression, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 595–595
63. Z. Qi, W. Gao, Variable-rate point cloud geometry compression based on feature adjustment and interpolation, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 63–72
64. Z. Yu, W. Gao, When dynamic neural network meets point cloud compression: Computation-aware variable rate and checkerboard context, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 600–600
65. L. Xie, W. Gao, S. Fan, Z. Yao, Pdnet: Parallel dual-branch network for point cloud geometry compression and analysis, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 596–596
66. L. Xie, W. Gao, H. Zheng, End-to-end point cloud geometry compression and analysis with sparse tensor, in *International Workshop on Advances in Point Cloud Compression, Processing and Analysis* (2022), pp. 27–32
67. C. Fu, G. Li, R. Song, W. Gao, S. Liu, Octattention: octree-based large-scale contexts model for point cloud compression. *Proc. AAAI Conf. Artif. Intell.* **36**(1), 625–633 (2022)

68. C. Zhang, W. Gao, Adadpcc: Adaptive rate control and rate-distortion-complexity optimization for dynamic point cloud compression, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
69. K. Wang, W. Gao, Unipcgc: Towards practical point cloud geometry compression via an efficient unified approach, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
70. K. Zheng, W. Gao, H. Zheng, Octree-based learned point cloud geometry compression: a lossy perspective, in *2025 Data Compression Conference (DCC)* (IEEE, 2025)
71. Z. Qi, H. Liu, W. Gao, Variable-rate point cloud compression with channel-wise entropy modeling, in *2025 Data Compression Conference (DCC)* (IEEE, 2025)
72. H. Zheng, W. Gao, Z. Yu, T. Zhao, G. Li, Viewpcgc: view-guided learned point cloud geometry compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 7152–7161
73. L. Xie, W. Gao, H. Zheng, G. Li, Roi-guided point cloud geometry compression towards human and machine vision, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 3741–3750
74. W. Gao, H. Ye, G. Li, H. Zheng, Y. Wu, L. Xie, OpenPointCloud: an open-source algorithm library of deep learning based point cloud compression, in *ACM International Conference on Multimedia* (2022), pp. 7347–7350
75. Y. Zhang, W. Gao, G. Li, Openpointcloud-v2: a deep learning based open-source algorithm library of point cloud processing, in *Proceedings of the 1st International Workshop on Advances in Point Cloud Compression, Processing and Analysis* (2022), pp. 51–55
76. W. Gao, S. Wang, X. Zhang, S. Kwong, Editorial for special issue on deep learning-based data compression. *APSIPA Trans. Signal Inf. Process.* **13**(6) (2024)
77. W. Gao, W. Gao, X. Mu, C. Peng, G. Li, Overview and comparison of avs point cloud compression standard. *APSIPA Trans. Signal Inf. Process.* (2024)
78. H. Sun, G. Li, S. Chen, L. Li, W. Gao, Special topic on 3d point cloud processing and applications. *ZTE Commun.* **21**(4), 1 (2023)
79. W. Gao, G. Li, H. Yuan, R. Hamzaoui, Z. Li, S. Liu, Apccpa'22: 1st international workshop on advances in point cloud compression, processing and analysis, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 7392–7393
80. W. Gao, G. Li, Point cloud compression, enhancement and applications: from 3d perception to large models, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 292–11 293
81. Z. Yao, W. Gao, Iterative saliency aggregation and assignment network for efficient salient object detection in optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **62**, 1–13 (2024)
82. B. Qu, H. Li, W. Gao, Bringing textual prompt to ai-generated image quality assessment, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024)
83. Z. Li, G. Li, T. Li, S. Liu, W. Gao, Information-growth attention network for image super-resolution, in *Proceedings of the 29th ACM International Conference on Multimedia* (2021), pp. 544–552
84. L. Zhou, W. Gao, G. Li, H. Yuan, T. Zhao, G. Yue, Disentangled feature distillation for light field super-resolution with degradations, in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2023), pp. 116–121
85. L. Zhou, W. Gao, G. Li, End-to-end spatial-angular light field super-resolution using parallax structure preservation strategy, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 3396–3400
86. W. Gao, L. Zhou, L. Tao, A fast view synthesis implementation method for light field applications *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* **17**(4), 1–20 (2021)
87. X. Zhang, W. Gao, G. Li, Q. Jiang, R. Cong, Image quality assessment-driven reinforcement learning for mixed distorted image restoration. *ACM Trans. Multimedia Comput. Commun. Appl.* **19**(1s), 1–23 (2023)

88. X. Zhang, W. Gao, H. Yuan, G. Li, Je 2 net: Joint exploitation and exploration in reinforcement learning based image restoration, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2090–2094
89. X. Zhang, W. Gao, Hirl: Hybrid image restoration based on hierarchical deep reinforcement learning via two-step analysis, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2445–2449
90. S. Wang, Y. Lu, W. Xia, P. Xia, Z. Wang, W. Gao, Light field angular super-resolution by view-specific queries. *Visual Comput.* 1–16 (2024)
91. W. Liu, W. Gao, X. Mu, Fast inter-frame motion prediction for compressed dynamic point cloud attribute enhancement. *Proc. AAAI Conf. Artif. Intell.* **38**(4), 3720–3728 (2024)
92. Z. Yang, W. Gao, X. Lu, Danet: density-adaptive network for geometry-based point cloud compression artifacts removal, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
93. X. Fan, G. Li, D. Li, Y. Ren, W. Gao, T. H. Li, Deep geometry post-processing for decompressed point clouds, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
94. X. Zhang, G. Liao, W. Gao, G. Li, Tdrnet: transformer-based dual-branch restoration network for geometry based point cloud compression artifacts, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
95. Z. Li, G. Li, T. H. Li, S. Liu, W. Gao, Semantic point cloud upsampling. *IEEE Trans. Multimedia* (2022)
96. Z. Li, S. Liu, W. Gao, G. Li, G. Li, S4r: rethinking point cloud sampling via guiding upsampling-aware perception. *IEEE Trans. Multimedia* (2025)
97. R. Zhang, W. Gao, G. Li, T.H. Li, Qinet: decision surface learning and adversarial enhancement for quasi-immune completion of diverse corrupted point clouds. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–14 (2022)
98. R. Bao, Y. Ren, G. Li, W. Gao, S. Liu, Flow-based point cloud completion network with adversarial refinement, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2559–2563
99. J. Chen, G. Li, R. Zhang, T. H. Li, W. Gao, Pointvae: invertible variational autoencoder framework for 3d point cloud generation, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 3216–3220
100. R. Zhang, J. Chen, W. Gao, G. Li, T.H. Li, Pointot: interpretable geometry-inspired point cloud generative model via optimal transport. *IEEE Trans. Circuits Syst. Video Technol.* **32**(10), 6792–6806 (2022)
101. W. Gao, G. Li, *Deep Learning for 3D Point Clouds* (Springer Nature, 2024)
102. W. Gao, G. Li, Introduction to 3d point clouds: Datasets and perception, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 1–27
103. W. Gao, G. Li, Learning basics for 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 29–70
104. W. Gao, G. Li, Deep-learning-based point cloud enhancement i, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 71–97
105. W. Gao, G. Li, Deep-learning-based point cloud enhancement ii, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 99–130
106. W. Gao, G. Li, Deep-learning-based point cloud analysis i, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 131–162
107. W. Gao, G. Li, Deep-learning-based point cloud analysis ii, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 163–193
108. W. Gao, G. Li, Point cloud pre-trained models and large models, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 195–225
109. W. Gao, G. Li, Point cloud-language multi-modal learning, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 227–254
110. W. Gao, G. Li, Open-source projects for 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 255–272

111. W. Gao, G. Li, Typical engineering applications of 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2025), pp. 273–299
112. W. Gao, G. Li, Future work on deep learning-based point cloud technologies, in *Deep Learning for 3D Point Clouds* (Springer, 2025), pp. 301–315
113. H. Yuan, W. Gao, W. Gao, Opensep: an open source subjective experiment platform, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 248–11 251
114. B. Qu, X. Liang, S. Sun, W. Gao, Exploring aigc video quality: a focus on visual harmony, video-text consistency and domain distribution gap, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2024)
115. S. Sun, X. Liang, S. Fan, W. Gao, W. Gao, Ve-bench: Subjective-aligned benchmark suite for text-driven video editing quality assessment, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
116. X. Zang, W. Gao, G. Li, H. Fang, C. Ban, Z. He, H. Sun, A baseline investigation: Transformer-based cross-view baseline for text-based person search, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 7737–7746
117. G. Liao, W. Gao, G. Li, J. Wang, S. Kwong, Cross-collaborative fusion-encoder network for robust rgb-thermal salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **32**(11), 7646–7661 (2022)
118. W. Gao, G. Liao, S. Ma, G. Li, Y. Liang, W. Lin, Unified information fusion network for multi-modal rgb-d and rgb-t salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **32**(4), 2091–2106 (2021)
119. Y. Chen, S. Sun, G. Li, W. Gao, T. H. Li, Closing the gap between theory and practice during alternating optimization for gans. *IEEE Trans. Neural Netw. Learn. Syst.* (2023)
120. Y. Chen, C. Jin, G. Li, T. H. Li, W. Gao, Mitigating label noise in gans via enhanced spectral normalization. *IEEE Trans. Circuits Syst. Video Technol.* (2023)
121. X. Zang, G. Li, W. Gao, Multidirection and multiscale pyramid in transformer for video-based pedestrian retrieval. *IEEE Trans. Ind. Inform.* **18**(12), 8776–8785 (2022)
122. X. Zang, G. Li, W. Gao, X. Shu, Learning to disentangle scenes for person re-identification. *Image Vision Comput.* **116**, 104330 (2021)
123. X. Zang, G. Li, W. Gao, X. Shu, Exploiting robust unsupervised video person re-identification. *IET Image Process.* **16**(3), 729–741 (2022)
124. Z. Yue, G. Li, W. Gao, Cross-level guided attention for human-object interaction detection, in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2023), pp. 284–289
125. S. Sun, J. Liu, H. Li, G. Liu, T. H. Li, W. Gao, Streamflow: Streamlined multi-frame optical flow estimation for video sequences, in *Advances in Neural Information Processing Systems* (2024)
126. G. Liao, W. Gao, Rethinking feature mining for light field salient object detection. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(10), 1–24 (2024)
127. B. Yang, W. Gao, G. Li, Focus and adjust: Progressive refinement network for human object interaction detection, in *2022 26th International Conference on Pattern Recognition (ICPR)* (IEEE, 2022), pp. 2546–2552
128. G. Liao, W. Gao, Q. Jiang, R. Wang, G. Li, Mmnet: multi-stage and multi-scale fusion network for rgb-d salient object detection, in *Proceedings of the 28th ACM International Conference on Multimedia* (2020), pp. 2436–2444
129. W. Gao, S. Kwong, Y. Zhou, Y. Jia, J. Zhang, W. Wu, Multiscale phase congruency analysis for image edge visual saliency detection, in *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 1 (IEEE, 2016), pp. 75–80
130. W. Gao, S. Kwong, Y. Zhou, X. Wang, Phase congruency analysis of down-sampled and blurring images for foreground extraction, in *2015 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)* (IEEE, 2015), pp. 152–157
131. Y. Guo, W. Gao, G. Li, Interpretable task-inspired adaptive filter pruning for neural networks under multiple constraints. *Int. J. Comput. Vision* 1–17 (2024)

132. W. Gao, Y. Guo, S. Ma, G. Li, S. Kwong, Efficient neural network compression inspired by compressive sensing. *IEEE Trans. Neural Netw. Learn. Syst.* (2022)
133. Y. Guo, W. Gao, Semantic-driven automatic filter pruning for neural networks, in *IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
134. L. Tao, W. Gao, Efficient channel pruning based on architecture alignment and probability model bypassing, in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2021), pp. 3232–3237
135. Y. Sun, Z. Li, S. Wang, W. Gao, Depth-assisted calibration on learning-based factorization for a compressive light field display. *Opt. Express* **31**(4), 5399–5413 (2023)
136. Y. Sun, Z. Li, L. Li, S. Wang, W. Gao, Optimization of compressive light field display in dual-guided learning, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2075–2079
137. W. Gao, S. Fan, G. Li, W. Lin, A thorough benchmark and a new model for light field saliency detection. *IEEE Trans. Pattern Anal. Mach. Intell.* (2023)
138. Z. Guo, W. Gao, H. Wang, J. Wang, S. Fan, No-reference deep quality assessment of compressed light field images, in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6
139. S. Fan, W. Gao, G. Li, Salient object detection for point clouds, in *European Conference on Computer Vision* (Springer, 2022), pp. 1–19
140. S. Luo, W. Gao, A general framework for rotation invariant point cloud analysis, in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2024), pp. 3665–3669
141. X. Lu, W. Gao, Attentivenet: Detecting small objects for lidar point clouds by attending to important points, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
142. Z. Pan, N. Zhang, W. Gao, S. Liu, G. Li, Less is more: label recommendation for weakly supervised point cloud semantic segmentation. *Proc. AAAI Conf. Artif. Intell.* **38**(5), 4397–4405 (2024)
143. Z. Pan, G. Liu, W. Gao, T.H. Li, Epcontrast: effective point-level contrastive learning for large-scale point cloud understanding, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024), pp. 1–6
144. N. Zhang, Z. Pan, T.H. Li, W. Gao, G. Li, Improving graph representation for point cloud segmentation via attentive filtering, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 1244–1254
145. K. Wen, N. Zhang, G. Li, W. Gao, Mpvnn: multi-resolution point-voxel non-parametric network for 3d point cloud processing, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024), pp. 1–6
146. Z. Pan, N. Zhang, W. Gao, S. Liu, G. Li, Point cloud semantic segmentation with sparse and inhomogeneous annotations, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
147. R. Zhang, G. Li, W. Gao, S. Liu, A quantum-inspired framework in master-servant mode for large-scale multi-modal place recognition. *IEEE Trans. Intell. Transp. Syst.* (2024)
148. D. Yang, W. Gao, Pointchd: a point cloud benchmark for congenital heart disease classification and segmentation. *IEEE J. Biomed. Health Inform.* (2024)
149. S. Fan, W. Gao, G. Li, Point-mpp: Point cloud self-supervised learning from masked position prediction. *IEEE Trans. Neural Netw. Learn. Syst.* (2024)
150. Z. Pan, W. Gao, G. Li, Distribution guidance network for weakly supervised point cloud semantic segmentation, in *Advances in Neural Information Processing Systems* (2024)
151. R. Zhang, G. Li, W. Gao, T. H. Li, Compoinet: can complex-valued representation benefit point cloud place recognition? *IEEE Trans. Intell. Transp. Syst.* (2024)
152. G. Li, W. Gao, W. Gao, *Point Cloud Compression: Technologies and Standardization* (Springer Nature, 2024)
153. G. Li, W. Gao, W. Gao, Introduction, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 1–28

154. G. Li, W. Gao, W. Gao, Background knowledge, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 29–51
155. G. Li, W. Gao, W. Gao, Predictive coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 53–70
156. G. Li, W. Gao, W. Gao, Transform coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 71–96
157. G. Li, W. Gao, W. Gao, Quantization techniques, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 97–112
158. G. Li, W. Gao, W. Gao, Entropy coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 113–133
159. G. Li, W. Gao, W. Gao, Mpeg geometry-based point cloud compression (g-pcc) standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 135–165
160. G. Li, W. Gao, W. Gao, Avs point cloud compression standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 167–197
161. G. Li, W. Gao, W. Gao, Mpeg video-based point cloud compression (v-pcc) standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 199–218
162. G. Li, W. Gao, W. Gao, Mpeg ai-based 3d graphics coding standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 219–241
163. G. Li, W. Gao, W. Gao, Future work, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 243–250
164. W. Gao, S. Ma, L. Duan, Y. Tian, P. Xing, Y. Wang, S. Wang, H. Jia, T. Huang, Digital retina: a way to make the city brain more efficient by visual coding. *IEEE Trans. Circuits Syst. Video Technol.* **31**(11), 4147–4161 (2021)
165. L. Xie, X. Mu, G. Li, W. Gao et al., Pku-dpcc: a new dataset for dynamic point cloud compression. *APSIPA Trans. Signal Inf. Process.* **13**(6) (2024)
166. S. Fan, W. Gao, Screen-based 3d subjective experiment software, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9672–9675
167. J. Wang, W. Gao, G. Li, Zoom to perceive better: No-reference point cloud quality assessment via exploring effective multiscale feature. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
168. J. Wang, W. Gao, G. Li, Applying collaborative adversarial learning to blind point cloud quality measurement. *IEEE Trans. Instrum. Meas.* (2023)
169. H. Li, W. Gao, Information exploration of projected views for point cloud quality measurement. *IEEE Trans. Instrum. Meas.* (2024)
170. Wikipedia, Mean opinion score. [Online], https://en.wikipedia.org/wiki/Mean_opinion_score
171. Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**(4), 600–612 (2004)
172. Z. Wang, E.P. Simoncelli, A.C. Bovik, Multiscale structural similarity for image quality assessment, in *The Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, vol. 2 (IEEE, 2003), pp. 1398–1402
173. Wikipedia, Rate-distortion theory, [Online], <https://en.wikipedia.org/wiki/Rate-distortiontheory>
174. N. Barman, M.G. Martini, Y. Reznik, Bjontegaard delta (bd): a tutorial overview of the metric, evolution, challenges, and recommendations (2024). [arXiv:2401.04039](https://arxiv.org/abs/2401.04039)

Chapter 2

Fundamentals for Deep Learning-based Image and Video Coding



2.1 Introduction

Traditional data compression leverages the data redundancy analysis and removal methods to implement the optimization goal [1–73]. However, deep learning data compression is built upon the foundation of deep learning techniques and data compression theory [6–9, 29–50, 53]. The key theoretical aspects of deep learning data compression are as follows.

Deep Learning Architectures: Deep learning architectures for data compression can be categorized into four main types. First, autoencoders form fundamental compression architectures, consisting of two primary components that establish the basis for deep learning-based compression methods. Second, variational autoencoders extend this framework by learning probability distributions, enabling both efficient compression and the generation of new samples. Third, generative adversarial networks (GANs) [74] employ adversarial training between generator and discriminator networks to produce high-quality compressed representations and reconstructions. Finally, convolutional neural networks (CNNs) [75] leverage their ability to learn local features, making them particularly effective for image compression tasks. Each architecture offers distinct advantages for different compression scenarios.

Lossless and Lossy Compression: Lossless compression methods retain all the original data's information, while lossy compression methods sacrifice some information but often achieve higher compression ratios. Deep learning methods can be applied to both types of data compression, with particular utility in lossy compression.

Feature Extraction through Learning: These methods can learn useful features, resulting in more effective data representation. These learned features can be used in the compression and decompression processes.

Efficient Encoding Techniques: Deep learning methods are often combined with efficient encoding techniques, such as entropy coding and arithmetic coding, to achieve higher compression rates.

In summary, the theoretical foundation of deep learning data compression encompasses various aspects, including deep learning models, autoencoders, generative adversarial networks, convolutional neural networks, and more. These technologies and theories provide more efficient and intelligent approaches to data compression, applicable to various data types, including images, audio, video, and point clouds.

2.2 Fundamental Knowledge of Deep Learning

2.2.1 *Convolutional Neural Networks*

Deep Neural Networks (DNN): As an extension of perceptrons, DNN can be viewed as neural networks with many hidden layers. When a deep neural network (DNN) is segmented according to the arrangement of its layers, the layers can be classified into three categories: input, output and hidden modules.

Convolutional Neural Networks (CNN): CNNs are mainly composed of several types of layers. In practice, the conv layer and the ReLU activation layer are frequently grouped together and referred to simply as the convolutional layer. Specifically, both these layers utilize activation functions during their transformation processes, incorporating various parameters like weights (w) and biases (b). In contrast, the activation layer and pooling layer execute fixed, invariant operations. The parameters in the convolutional and fully connected layers are adjusted through gradient descent until the loss is minimized, leading to the final output. The neurons in the layers of a convolutional neural network are organized in three dimensions: width, height, and depth. Width and height are straightforward, usually corresponding to the dimensions of an image. In deep learning, depth generally refers to the number of channels.

It is well known that channel variations are achieved through convolution, and the purpose of convolving with kernels is to extract image features. However, using just one convolutional kernel is far from sufficient because a single kernel can only capture certain features of the image. Features can be understood as parts of the image. For example, after convolution and activation functions are applied to a colored beach image in RGB format, the resulting feature map might highlight only the contours of the sea. Another convolution might reveal the contours of the beach. Thus, to fully interpret the image, multiple convolutional kernels are needed to extract diverse features. The conceptual framework maintains fixed input image pixels while the feature extraction process dynamically evolves through continuous training of convolutional operations in the CNN architecture. Through loss computation and backpropagation (using the chain rule to calculate partial derivatives), the convolutional kernel parameters are updated. In fact, this gradient-based optimization process can be viewed as a form of reverse convolution.

The original image undergoes progressive abstraction as different convolutional kernels extract various features, collectively working to preserve the most essential characteristics of the input. Through successive convolutional layers, higher-level

semantic information becomes increasingly distinct while spatial details gradually diminish. As the network depth increases, the original image becomes less visually recognizable as features are transformed into increasingly abstract representations that only the model can effectively interpret. Furthermore, deeper layers develop larger receptive fields over the input image, resulting in reduced sensitivity to precise local spatial relationships. Each component is examined below.

Input Layer: This layer typically receives raw data input into the convolutional neural network or preprocessed data with operations such as scaling, mean subtraction, and normalization. For example, in image recognition, the input to the input layer is often a three-dimensional colored image. The input image usually consists of RGB channels, forming a 3D pixel value matrix with dimensions $H \times W \times C$, where H and W represent height and width. The convolutional network processes the data through convolution and pooling operations for feature extraction and transformation. The final summary and result output of the features are produced by the fully connected layer. If the number of images specified for the input layer is N , the output data of the input layer is $N \times H \times W \times 3$.

Convolutional Layer: This layer is primarily utilized to extract features using a convolutional kernel matrix, which reveals hidden relationships within the original data. The convolution operation fundamentally consists of performing a dot product and summation between two matrices: one representing the input data and the other representing the convolutional kernel (also known as a filter or feature matrix). The resulting output highlights specific local features derived from the original image. As previously mentioned, convolution operations at various levels extract different types of features—shallow layers tend to capture edge features, mid-level convolutions emphasize local features, and deeper convolutions are more aligned with global features. Convolution is inseparable from the convolutional kernel, and the kernel corresponds to some basic parameters: kernel size, stride, padding method, input channels, and output channels. With the continuous development of deep learning, convolutional kernels also come in various types, such as standard convolution, transpose convolution, separable convolution, and dilated convolution. In early networks, due to limited computational power and shallower network architectures, larger convolutional kernels were used to achieve larger receptive fields. The 1×1 kernel is primarily used for cross-channel interaction and fusion of information, as well as for channel dimensionality reduction and expansion, thereby reducing the number of parameters. Nevertheless, larger convolutional kernels still play a unique role in certain domains, and each size of convolutional kernel has its advantages. Additionally, in architectures like GoogLeNet, Inception series, or SPP modules, different-sized convolutional kernels are jointly used within the same layer. The features obtained from each kernel are either added or concatenated, resulting in more expressive new features.

Activation Layer: The non-linearity in the activation layer can address the linear transformation relationship between the input matrix and the convolutional kernel matrix. Further details about activation functions will be discussed below.

Pooling Layer: The pooling layer filters features within a receptive field, extracting the most representative features from a given region. Max pooling is the most

widely used method, selecting the maximum activation within each sub-region. Additionally, pooling serves as a subsampling operation, reducing the spatial dimensions of feature maps and decreasing the number of parameters in subsequent layers.

Fully Connected Layer: The fully connected layer serves as the final feature integrator in convolutional neural networks, transforming high-dimensional feature maps into task-specific outputs. It achieves this by flattening spatial features into vectors and applying learned weight matrices. While crucial for classification tasks, this layer typically contains the majority of network parameters, often leading to computational bottlenecks. Unlike convolutional layers which focus on local feature extraction, fully connected layers establish global relationships across all features, making them particularly parameter-intensive despite their straightforward operation.

2.2.2 *Transformer*

The Transformer model [76], created by Google, is an architecture of neural networks that utilizes the self-attention mechanism to handle sequential data. This is illustrated in Figs. 2.1 and 2.2. In comparison to traditional recurrent neural network (RNN) models, the Transformer offers improved parallel performance and reduced training times, which has led to its widespread use in natural language processing. The input sequential data can include various forms such as text, speech signals, images, or videos. While traditional RNN models have demonstrated effective performance in these tasks, they encounter two significant challenges: limitations in parallel computation and difficulties in capturing long-range dependencies. The introduction of the Transformer model addresses these issues effectively.

Transformer is capable of globally modeling each element within a sequence and establishing relationships between those elements. In contrast to recurrent neural network models, Transformer exhibits enhanced parallel performance and reduced training times. Transformer consists of encoders and decoders, with each layer comprising several modules for the attention mechanism and feedforward neural networks. The encoder transforms input sequence into a high-dimensional feature vector, while the decoder converts this vector representation back into target sequence. To improve model convergence and overall performance, techniques like residual connections and layer normalization are utilized within Transformer model.

Transformer involves three steps: (1) Compute Attention Weights: Determine the attention weights for each position in relation to others, reflecting the significance of each position to the others. (2) Compute Weighted Sum: Multiply each position vector by its corresponding attention weights, then sum these products. (3) Linear Transformation: Apply it to the weighted sum vector to generate the final output vector. Transformer model is usually pre-trained by unsupervised methods, followed by fine-tuning through supervised techniques. During pre-training, methods like autoencoders or masked language models are often employed to learn representations of input sequences. In the fine-tuning process, supervised training is used, such as using

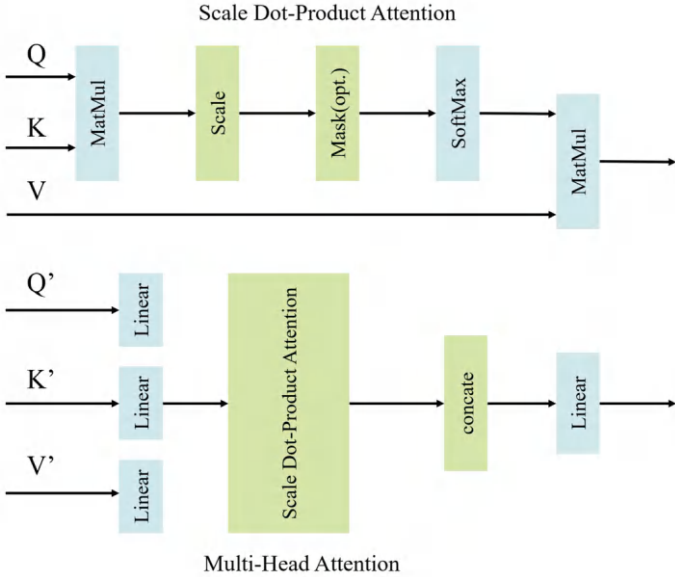


Fig. 2.1 The structure of transformer. Q, K, V in the scale dot-product attention are obtained through linear layer. Public domain open access image [77]

parallel corpora for machine translation tasks, with the goal of learning the mapping relationship from input sequences to target sequences.

Multi-Head Self-Attention. First, the multi-head mechanism utilizes several self-attention heads, with each head learning distinct attention weights. This approach enables the model to recognize a variety of relationships within the data. The head combination can combine the outputs of multiple attention heads through linear projection and concatenation, followed by another linear transformation to produce the final attention output.

Positional Encoding. Transformer lacks inherent sequential information. Positional encoding is introduced to embed positional information into the input embedding matrix.

Feedforward Neural Network. Each Transformer layer contains a feedforward neural network, which further nonlinearly transforms the representation at each position, building on the self-attention mechanism.

Residual Connections and Layer Normalization. First, for the residual connections, the input and output are added via residual connections. The layer normalization can apply layer-wise normalization to the output of each sub-layer for improved training stability.

Encoder-Decoder Architecture. For applications such as machine translation, the Transformer utilizes an encoder-decoder architecture. The decoder component incorporates additional self-attention mechanisms, which enable it to effectively

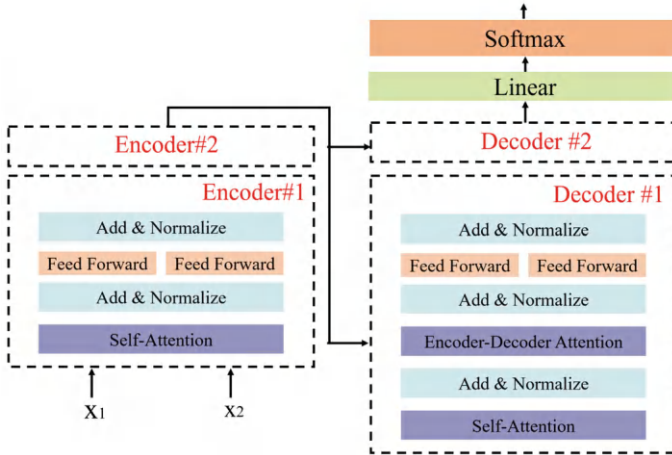


Fig. 2.2 The structure of encoder and decoder in transformer. Public domain open access image [77]

process and generate sequential outputs while maintaining awareness of previously generated tokens.

Moreover, Transformers also have many applications in visual tasks: (1) Image Classification. Transformers have been demonstrated competitive performance with convolutional neural networks (CNNs). (2) Object Detection. Vision Transformer (ViT) is a transformer-based model that has shown effectiveness in object detection tasks, outperforming traditional CNN-based approaches. (3) Image Segmentation. Transformers have been adapted for image segmentation tasks, showcasing their capability to capture long-range dependencies in images. The Transformer’s versatility extends beyond NLP, making it a fundamental architecture in various sequence processing tasks, including vision-related applications.

2.3 Autoencoder and Variational Autoencoder

Data compression allows representing the same amount of information using fewer dimensions. This helps address the curse of dimensionality and can partially mitigate the phenomenon of overfitting. In general, dimensionality reduction methods are often applied to the dataset before it is used for training. This is where autoencoders [78] (AE) and variational autoencoders [79] (VAE) come into play. They are end-to-end networks used for compressing input data. Both Autoencoders and Variational Autoencoders are used to convert data to a lower-dimensional representation, effectively achieving data compression.

2.3.1 Autoencoder

An autoencoder (AE)¹ is designed to learn meaningful embeddings of unlabeled data within a specific neural network architecture, as illustrated in Fig. 2.3. The decoder plays a critical role for guaranteeing that the latent space captures most of the information from the original dataset, while also leveraging the output from the latent space to assess the accuracy of the data reconstruction.

Key Components. The encoder component transforms the input data into a lower-dimensional representation, effectively compressing the information into a latent space. Subsequently, the decoder reconstructs the original input data from this compressed representation, completing the encoding-decoding process.

Basic Workflow. The encoding process begins by feeding input data into the encoder network, which transforms it into a compressed representation within a latent space. This compressed representation then serves as input to the decoder network, which performs the inverse operation to reconstruct the original data from its lower-dimensional encoding.

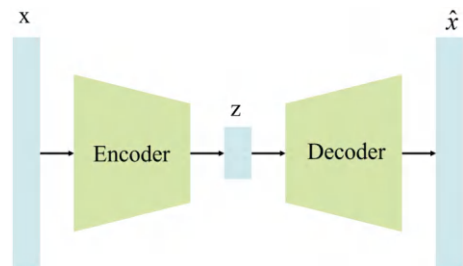
Loss Function. The loss function usually incorporates a reconstruction loss (such as mean squared error) to quantify the disparity between the input and output, as illustrated below:

$$\text{Loss} = \|x - \hat{x}\|_2 = \|x - d_\phi(z)\|_2 = \|x - d_\phi(e_\theta(x))\|_2. \quad (2.1)$$

During the training process, the input data, denoted as x , is fed into the encoder function, $e_\theta(x)$. The input undergoes a series of layers parameterized by the variable θ to reduce its dimensionality, resulting in a compressed latent vector, z . The number of layers, their types and sizes, as well as the dimensionality of the latent space, are configurable hyperparameters. The decoder, denoted as $d_\phi(z)$, typically (but not necessarily) consists of approximately complementary layers to those used in the encoder, arranged in reverse order.

Within a single network architecture, complementary layers can be implemented to partially reverse the operations performed by preceding layers. This symmetric design principle manifests in several layer pairings: convolutional layers may be paired with transposed convolutional layers, pooling operations with unpooling

Fig. 2.3 The brief structure of autoencoder. x is the original data and $\hat{x}$ is the reconstructed data. Public domain open access image (<https://en.wikipedia.org/wiki/Autoencoder>)



¹ <https://en.wikipedia.org/wiki/Autoencoder>.

operations, and fully connected layers with corresponding inverse fully connected layers. Such architectural symmetry enables more flexible feature transformation pathways throughout the network. When these regions are used as inputs to the encoder, the resulting outputs do not resemble any digits in the MNIST² dataset. This phenomenon illustrates the lack of regularization in the latent space.

In such a latent space, only a few regions or clusters exhibit generative capabilities, meaning that sampling any point within a cluster will produce outputs related to that cluster. However, the entire latent space does not possess generative capabilities; regions that do not belong to any cluster will yield spurious outputs. Once the network is trained and the training data is removed, it becomes impossible to determine whether the decoder's output from a randomly sampled latent vector is valid. Consequently, autoencoders (AEs) are primarily utilized for compression. For valid inputs, AEs can reduce the dimensionality, effectively eliminating redundancy. However, because the latent space in autoencoders is not regularized, the decoder may struggle to produce valid input data from vectors sampled within that space.

2.3.2 Variational Autoencoder

Unlike the encoder in a standard autoencoder, which produces a latent vector, the Variational Autoencoder (VAE)'s encoder outputs parameters for a predefined distribution in the latent space for each input. The VAE³ enforces constraints on this latent distribution, requiring it to approximate a normal distribution. This constraint effectively regularizes the latent space. Like a traditional autoencoder, the input passes through a series of layers (defined by hyperparameters) to reduce its dimensionality and generate a compressed latent vector z . The latent vector is then sampled from these mean and standard deviation values, which are subsequently inputted into the decoder to reconstruct the original input. As illustrated in Fig. 2.4, the decoder in a VAE operates similarly to that in a traditional autoencoder.

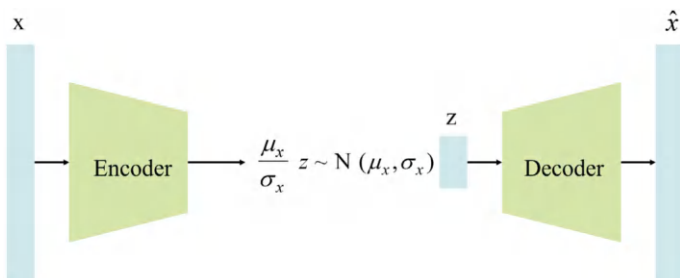


Fig. 2.4 The structure of variational autoencoder. Public domain open access image (https://en.wikipedia.org/wiki/Variational_autoencoder)

² <http://yann.lecun.com/exdb/mnist>.

³ https://en.wikipedia.org/wiki/Variational_autoencoder.

The loss function for the VAE is characterized by two primary objectives: first, to accurately reconstruct the original input data, and second, to ensure that the latent space conforms to a normal distribution. The reconstruction error, similar to that in traditional autoencoders, is calculated as the mean squared error between the input-output pair and the reconstructed output. Thus, the overall loss function is the sum of these two components, as illustrated below:

$$\begin{aligned}
 L_{reconstruction} &= \|x - \hat{x}\|_2 = \|x - d_\phi(z)\|_2 = \|x - d_\phi(\mu_x + \sigma_x \epsilon)\|_2, \\
 \mu_x, \sigma_x &= e_\theta(x), \\
 \epsilon &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \\
 L_{similarity} &= D_{KL}(\mathcal{N}(\mu_x, \sigma_x) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I})), \\
 L &= L_{reconstruction} + L_{similarity},
 \end{aligned} \tag{2.2}$$

where x is the original input data, $\hat{x}$ is the reconstructed output, $d_\phi(z)$ is the decoder function, z is the latent variable, $L_{reconstruction}$ is the reconstruction loss measuring the difference between x and $\hat{x}$, $L_{similarity}$ is the KL divergence ensuring the latent distribution matches a normal distribution, and L is the total loss combining both components.

Key Features. The VAE fundamentally differs from standard autoencoders by representing the latent space probabilistically, typically modeled as a multivariate Gaussian distribution. This probabilistic framework enables the model to sample latent vectors from the learned distribution during both training and inference, rather than deterministically encoding inputs to fixed points in latent space. The stochastic sampling mechanism introduces beneficial regularization effects while maintaining the continuous and complete property of the latent space representation.

Workflow. During training, samples from this distribution are drawn to create the latent representations. The decoder then reconstructs the input using these sampled latent vectors.

Loss Function. The loss function in a VAE consists of two components: the reconstruction loss, which measures how accurately the output matches the input, and the regularization term (KL divergence), which keeps the latent space distribution close to a standard Gaussian.

This approach effectively clusters similar input data points together. The unit norm condition ensures a uniform distribution of the latent space with no significant gaps between clusters. In fact, clusters of similar input data often overlap in certain regions.

Differences Between AE and VAE. The VAE differs fundamentally from traditional AEs in several key aspects. While AEs employ deterministic latent space representations, VAEs model the latent space probabilistically. This probabilistic approach introduces stochasticity through sampling from the latent distribution during training. The VAE's regularization term further structures the latent space to be continuous and well-organized, enhancing its suitability for generative tasks. Unlike standard autoencoders primarily used for feature learning, dimensionality reduction,

denoising, and compression, VAEs are explicitly designed as generative models capable of synthesizing new data samples through latent space sampling, making them particularly valuable for image generation and data synthesis applications.

In summary, autoencoders and variational autoencoders offer different approaches to representation learning, with VAE incorporating probabilistic and generative aspects for more structured latent space and generative capabilities.

2.4 Principles and Framework of Learned Image and Video Coding

2.4.1 *Background Knowledge of Image and Video Compression*

When network bandwidth is constrained, data compression has gained immense significance. A notable example is the innovative data compression algorithm featured in the first season of the TV series *Silicon Valley*, which catapulted the Pied Piper company to prominence. Effective data compression has enabled large-scale network applications to flourish on mobile platforms, paving the way for a promising future. In scenarios where transmission networks and storage capacities are limited, the importance of data compression technology cannot be overstated. Shannon's source coding theorem, also known as Shannon-Fano Coding, can be viewed as a fundamental theorem describing this idea. Classic data compression algorithms based on statistical models, such as Huffman Coding, Arithmetic Coding, and the more recently developed Asymmetric Numeral Systems (ANS), are all designed based on the frequency of individual characters in the information. Beyond methods grounded in statistical probabilities, classic data compression techniques also include dictionary-based compression technologies like LZ77, LZ78, LZW, as well as Entropy Encoding, Run-Length Encoding, etc.

The commonly-used data compression tools are generally variations, combinations, or optimizations of these classic methods, rarely employing a single technique in isolation. For example, gzip compresses data first using a variant of the LZ77 algorithm, then compresses the result using either static or dynamic Huffman coding. bzip2 works by initially encoding with a run-length encoder, followed by block-sorting compression and Move-to-Front (MTF) transformation to generate a sequence of identical symbols, and then encoding again with another run-length encoder. The final result is packed with a message header using Huffman coding. LZMA is an enhanced and optimized compression algorithm derived from Deflate and LZ77, with Deflate itself utilizing both LZ77 and Huffman coding in a lossless data compression scheme. However, in the big data era, traditional data compression techniques are proving inadequate to meet the challenges posed by vast volumes, rapid growth, and intricate data structures, especially when it comes to real-time data compression.

Some researchers have attained promising results in data compression by utilizing traditional neural network models. By processing original images through neural networks, it is possible to store only the network weights instead of the image information itself, achieving a higher compression ratio without compromising image quality. Building on this, combining neural networks with other compression techniques, improving neural network structures, and enhancing training algorithms have further advanced the application of neural networks in data compression. However, neural networks are shallow networks, and there are deficiencies in the model's convergence speed, stability, and the effectiveness of weight updates. Additionally, neural networks rely on labeled pre-training data, which can be challenging to obtain in practical applications.

2.4.2 Bring Learning to Image and Video Compression

The introduction of deep learning has effectively addressed issues inherent in traditional methods. Compared to conventional compression techniques, deep learning-based methods possess the following natural advantages.

Deep learning parameters are generated from extensive datasets derived from the real world, while traditional data compression encoding methods are primarily crafted manually using established knowledge and heuristics. Hence, deep learning demonstrates superior content adaptability compared to signal processing-based models. Deep learning methodologies efficiently utilize larger receptive fields. This not only allows the use of adjacent information but also leverages distant samples to improve encoding efficiency, a contrast to traditional encoding tools which only utilize neighboring samples and struggle to exploit distant samples. The internal representations of deep learning are suitable for the modern data processing.

Deep learning-based compression methods exhibit advantages over traditional neural network approaches. First, their multi-layer architectures provide enhanced nonlinear mapping capabilities, enabling more effective learning of deep data features, particularly for image data. Second, the multi-layer learning process with weight fine-tuning improves training efficiency, making these methods suitable for large-scale data compression tasks. Following this discussion of advantages, we can present the general framework for deep learning-based image and video coding.

2.4.3 End-to-End Optimized Image Compression

The study detailed in [80], published at ICLR 2017, introduces an algorithm that integrates analysis transform, uniform quantizer, and synthesis transform. The analysis transform process entails three repetitive stages, each comprising a convolutional linear filter paired with a nonlinear activation function. The purpose of this joint nonlinearity is to achieve local gain control. The introduction of a proxy function is

necessitated due to the non-differentiable nature of the uniform quantizer, enabling end-to-end training of the network. The relaxed loss function can be interpreted as the log-likelihood through a VAE-based generative model. The algorithm outperforms JPEG and JPEG-2000 methods in terms of metrics like the Peak Signal-to-Noise Ratio (PSNR) and the Multi-Scale Structural Similarity (MS-SSIM).

As shown in Fig. 2.5, the algorithm encodes the image in the coding space to obtain a compressed representation, denoted as R . It then transforms into the perceptual space to compute the distortion rate of the image, referred to as D . Figure 2.6 showcases the analysis (encoding) and synthesis (decoding) modules of the algorithm. These modules employ upsampling and downsampling techniques, along with Generalized Divisive Normalization (GDN) and Inverse GDN (IGDN) structures.

Within the Autoencoder module, GDN is employed for normalization. This is akin to the role of Batch Normalization in conventional CNN networks, effectively capturing the statistical characteristics of images and transforming them into a Gaussian distribution.

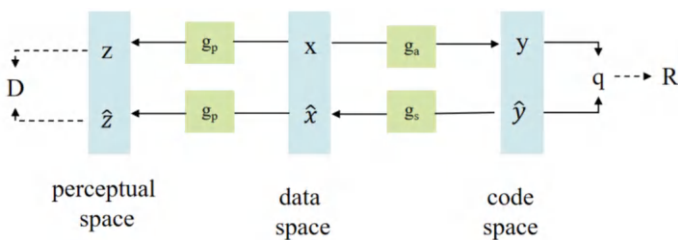


Fig. 2.5 The framework of end-to-end image compression. Public domain open access image [80]

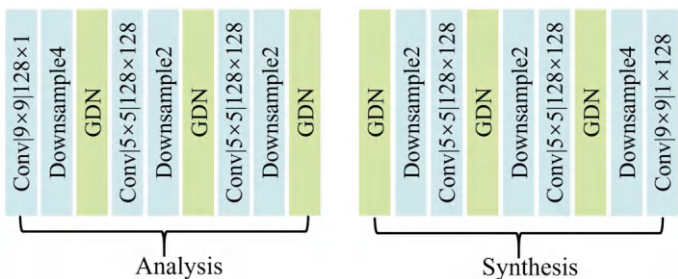


Fig. 2.6 The specific architecture of end-to-end image compression. Public domain open access image [80]

2.4.4 Hyperprior for Variational Image Compression

The trainable image compression model is presented in [81] built on VAE, as shown in Fig. 2.7. This model utilizes the hyperprior to for removing the spatial dependencies within the latent representations. The hyperprior concept, which relates to auxiliary information, is widely applicable in contemporary image codecs but has been largely overlooked in image compression using Artificial Neural Networks (ANN). Unlike existing autoencoder-based compression techniques, this model is jointly trained with the underlying autoencoder to develop intricate priors. It has been shown that this model achieves state-of-the-art performance in image compression when evaluated for visual quality using the widely recognized MS-SSIM index [82–91]. Moreover, its rate-distortion performance exceeds that of previously published ANN-based methods when assessed with more conventional metrics such as the PSNR. Additionally, a qualitative comparison is provided for models trained using different distortion metrics. The bitrate estimation for this model is computed as:

$$R = \mathbb{E}_{\hat{y} \sim m} [-\log_2 p_{\hat{y}}(\hat{y})], \quad (2.3)$$

where $p_{\hat{y}}$ represents the actual distribution of the latent representation, estimated by the entropy model. The entropy model is a shared prior probability model between sender and receiver, utilized to estimate the true distribution of latent representation. The formula indicates that the encoding length is minimized when the entropy model's estimation of $p_{\hat{y}}$ perfectly matches the actual distribution. This reveals that, on one hand, when entropy model uses a fully factorized probability distribution, it naturally fails to fit the actual distribution if statistical dependencies exist in reality.

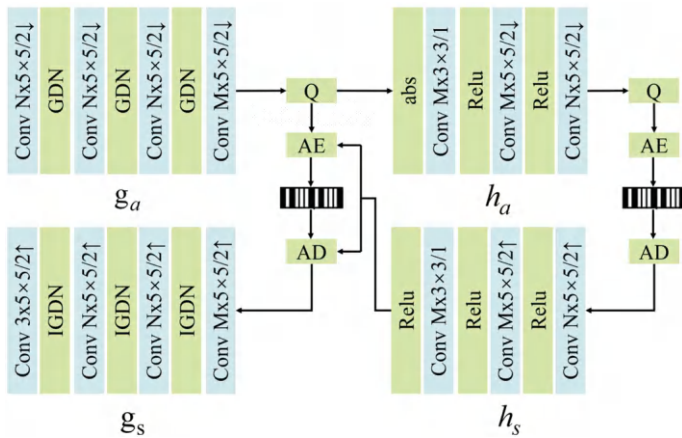


Fig. 2.7 The framework of variational image compression with a scale hyperprior. Public domain open access image [81]

On the other hand, since $y = \text{encoder}(x)$ is a deterministic process, improving the outcome under a fully factorized probability distribution requires encoder to remove as much statistical dependency as possible from input natural images.

The key aspect is the introduction of side information to capture hidden information in the latent representation, and thereby addressing the mismatch between the estimated and actual distributions of the latent layer. By incorporating side information into the bitstream, the decoder can also share the entropy model. During decompression, the decoder first unpacks the side information to construct the entropy model, and then decompresses the latent information based on the correct entropy model.

The optimization problem of the algorithm can be modeled using a VAE. In the context of compression, the analysis model g_a is analogous to the VAE's inference model, while the synthesis model g_s is likened to the generative model. Within the inference model of VAE, the true posterior probability is estimated $p_{\tilde{y}|x}(\tilde{y} | x)$, which is often impractical. Therefore, VAE employs a parameterized variational density $q(\tilde{y} | x)$ to approximate the true posterior probability $p_{\tilde{y}|x}(\tilde{y} | x)$. This is achieved by minimizing the optimization target—the Kullback-Leibler (KL) divergence between the true distribution of the latent variables and the simulated distribution. This process is equivalent to minimizing the rate-distortion performance in image compression:

$$\mathbb{E}_{x \sim p_x} D_{KL} [q \| p_{\tilde{y}|x}] = \mathbb{E}_{x \sim p_x} \mathbb{E}_{\tilde{y} \sim q} \left[\underbrace{-\log q(\tilde{y} | x)}_{\text{weighted distortion}} + \underbrace{-\log p_{x|\tilde{y}}(x | \tilde{y}) - \log p_{\tilde{y}}(\tilde{y})}_{\text{rate}} \right] + \text{const}, \quad (2.4)$$

where x is the input variable sampled from the distribution p_x , $\tilde{y}$ is the latent variable sampled from the distribution q , D_{KL} is the Kullback-Leibler divergence measuring the difference between the two distributions, $p_{x|\tilde{y}}$ is the conditional probability of x given $\tilde{y}$, and $p_{\tilde{y}}$ is the prior distribution of $\tilde{y}$.

2.4.5 Joint Autoregressive and Hierarchical Priors for Image Compression

Minnen et al. [92] proposed the hyperprior model enhances compression performance by constructing a Gaussian Scale Mixture (GSM) as side information, incorporating the variance σ of Gaussian distributions. Building upon this, the mean μ of Gaussian distributions to construct a Gaussian Mixture Model (GMM) is introduced in [92], further improving compression performance, as shown in Fig. 2.8. The left part of the figure below, when considered in isolation, represents the simplest form of an autoencoder. Here, the input image is processed through an encoder, quantized, and then passed through an arithmetic encoder to generate a bitstream for storage and transmission. This bitstream is subsequently decoded through an arithmetic decoder

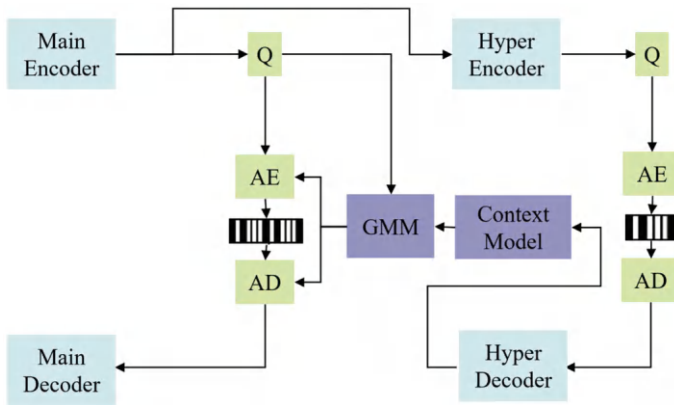


Fig. 2.8 The framework of joint autoregressive and hierarchical priors for learned image compression. Public domain open access image [92]

and then reconstructed by the decoder to retrieve the image. To enhance compression performance, the authors have incorporated side information to aid in decoding. This is done by storing the variance information of the Gaussian distribution in the bitstream through a Hyper Encoder, a quantizer, and an arithmetic encoder. During image reconstruction, this stored variance is retrieved through an arithmetic decoder and a Hyper Decoder. Simultaneously, the mean information is reconstructed through a Context Model. The variance and mean are then combined in the Entropy Parameters to form a Gaussian distribution. This Gaussian distribution is utilized to assist the decoder in reconstructing the image.

2.4.6 Generative Adversarial Networks for Image Compression

A principled full-resolution GAN compression framework is proposed in [93], utilizing it to develop an extreme image compression method, as shown in Fig. 2.9. Additionally, they present the first extensive user study related to full-resolution image compression using this framework, which sets it apart from previous works. Their approach considers two operational modes: (1) The Generative Compression (GC) mode aims to preserve the image signals and generate structures at various scales. (2) The Selective Generative Compression (SC) mode involves fully producing the parts of image based on the semantic label mappings, and retaining user-defined areas with desirable fidelity.

The extreme image compression GAN framework operates by first encoding an input image x through the encoder E and the quantizer q to obtain a compressed representation $\hat{w}$, which is then concatenated with random noise v to form a latent

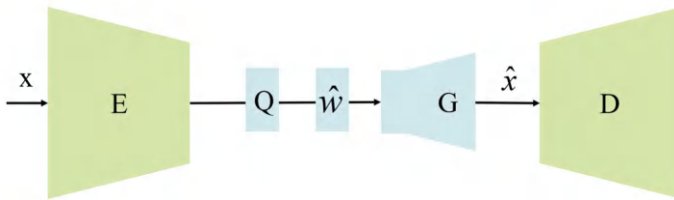


Fig. 2.9 The framework of generative adversarial networks for extreme learned image compression. ©2019 IEEE. Reprinted, with permission, from Ref. [93]

representation z . The framework optimizes a composite loss function that balances three key components: a GAN-based adversarial loss ensuring the generated images match the real data distribution, a reconstruction loss measuring how well the specific input image is preserved, and an entropy term controlling the compression rate. When extended to its conditional variant, the framework additionally incorporates semantic information s (like scene label maps) solely into the discriminator during training, while keeping the encoder, generator, and inference process unchanged.

2.5 Summary

This chapter provides a foundational overview of modern image and video coding through the integration of deep learning approaches. It bridges the gap between traditional coding methods and contemporary AI-driven solutions, with a particular focus on deep learning architectures such as autoencoders and variational autoencoders. The discussion encompasses both theoretical frameworks and practical implementations, demonstrating how AI-based coding systems can enhance compression efficiency and visual quality. By examining the entire pipeline from data acquisition to compression, the chapter establishes essential groundwork for understanding advanced coding techniques, making it a crucial reference for both newcomers and practitioners in the field of visual data compression. Moreover, the image and video enhancement and restoration algorithms [91, 94–126] are also very helpful for quality improvement of image and video content, and thus elevate the overall rate-distortion performances. Additionally, the current image and video coding optimization efforts not only focus on the visual quality improvements, but also concern with the machine analysis performances.

Exercises

1. Please describe the functions of the encoder and decoder of the autoencoder.
2. Please describe how variational autoencoders are used for data compression.
3. Please describe why we need multiple convolution kernels in convolutional neural networks.
4. Please describe the basic workflow of the autoencoder.
5. Please describe the basic workflow of a variational autoencoder.
6. Please explain the importance of image and video data compression.
7. Please describe the classical data compression algorithms based on statistical models.
8. Please describe some classical data compression techniques besides methods based on statistical probability.
9. Please explain what are the advantages of deep learning-based methods compared to traditional neural network compression techniques?
10. Please describe what types of deep learning-based methods can be used for data compression currently.

References

1. W. Gao, G. Li, H. Yuan, R. Hamzaoui, Z. Li, S. Liu, Apccpa'22: 1st international workshop on advances in point cloud compression, processing and analysis, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 7392–7393
2. W. Gao, G. Li, Point cloud compression, enhancement and applications: from 3d perception to large models, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 292–11 293
3. H. Sun, G. Li, S. Chen, L. Li, W. Gao, Special topic on 3d point cloud processing and applications. *ZTE Commun.* **21**(4), 1 (2023)
4. W. Gao, W. Gao, X. Mu, C. Peng, G. Li, Overview and comparison of avs point cloud compression standard. *APSIPA Trans. Signal Inf. Process.* (2024)
5. W. Gao, S. Wang, X. Zhang, S. Kwong, Editorial for special issue on deep learning-based data compression. *APSIPA Trans. Signal Inf. Process.* 13(6) (2024)
6. W. Gao, H. Ye, G. Li, H. Zheng, Y. Wu, L. Xie, OpenPointCloud: an open-source algorithm library of deep learning based point cloud compression, in *ACM International Conference on Multimedia* (2022), pp. 7347–7350
7. Y. Zhang, W. Gao, G. Li, Openpointcloud-v2: a deep learning based open-source algorithm library of point cloud processing, in *Proceedings of the 1st International Workshop on Advances in Point Cloud Compression, Processing and Analysis* (2022), pp. 51–55
8. L. Xie, W. Gao, Learningpcc: a pytorch library for learning-based point cloud compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024)
9. L. Xie, W. Gao, Pchmvision: an open-source library of point cloud compression for human and machine vision, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024)
10. T. Qin, G. Li, W. Gao, S. Liu, Multi-grained point cloud geometry compression via dual-model prediction with extended octree. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(9), 1–30 (2024)

11. Y. Shao, W. Gao, S. Liu, G. Li, Advanced patch-based affine motion estimation for dynamic point cloud geometry compression. *Sensors* **24**(10), 3142 (2024)
12. Y. Shao, F. Song, W. Gao, S. Liu, G. Li, Texture-guided graph transform optimization for point cloud attribute compression. *Appl. Sci.* **14**(10), 4094 (2024)
13. Y. Shao, X. Yang, W. Gao, S. Liu, G. Li, 3d point cloud attribute compression using diffusion-based texture-aware intra prediction. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
14. J. Zhang, Y. Chen, G. Liu, W. Gao, G. Li, Efficient point cloud attribute compression framework using attribute-guided graph fourier transform, in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2024), pp. 8426–8430
15. W. Gao, H. Yuan, G. Li, Z. Li, H. Yuan, Low complexity coding unit decision for video-based point cloud compression. *IEEE Trans. Image Process.* **33**, 149–162 (2023)
16. Y. Shao, G. Li, Q. Zhang, W. Gao, S. Liu, Nonrigid registration-based progressive motion compensation for point cloud geometry compression. *IEEE Trans. Geosci. Remote Sens.* (2023)
17. F. Song, G. Li, X. Yang, W. Gao, S. Liu, Block-adaptive point cloud attribute coding with region-aware optimized transform. *IEEE Trans. Circuits Syst. Video Technol.* (2023)
18. Y. An, Y. Shao, G. Li, W. Gao, S. Liu, A fast motion estimation method with hamming distance for lidar point cloud compression, in *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2022), pp. 1–5
19. H. Yuan, W. Gao, G. Li, Z. Li, Rate-distortion-guided learning approach with cross-projection information for v-pcc fast cu decision, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 3085–3093
20. F. Song, G. Li, W. Gao, T.H. Li, Rate-distortion optimized graph for point cloud attribute coding. *IEEE Signal Process. Lett.* **29**, 922–926 (2022)
21. F. Song, G. Li, X. Yang, W. Gao, T.H. Li, Fine-grained correlation representation for graph-based point cloud attribute compression, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
22. F. Shen, W. Gao, A rate control algorithm for video-based point cloud compression, in *2021 International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2021), pp. 1–5
23. F. Song, Y. Shao, W. Gao, H. Wang, T. Li, Layer-wise geometry aggregation framework for lossless lidar point cloud compression. *IEEE Trans. Circuits Syst. Video Technol.* **31**(12), 4603–4616 (2021)
24. Z. Cai, W. Gao, G. Li, W. Gao, Distortion propagation model-based v-pcc rate control for 3d point cloud broadcasting. *IEEE Trans. Broadcast.* 1–13 (2024)
25. X. Mu, W. Gao, H. Yuan, S. Wang, G. Li, Adaptive lpu decision for dynamic point cloud compression. *IEEE Signal Process. Lett.* (2024)
26. C. Peng, W. Gao, Laplacian matrix learning for point cloud attribute compression with ternary search-based adaptive block partition, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 10 412–10 420
27. Y. Wang, W. Gao, X. Mu, H. Yuan, Rate control optimization for joint geometry and attribute coding of lidar point clouds, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
28. T. Guo, H. Yuan, T. Wang, W. Gao, Graph filter-based fast motion matching for inter frame coding of mpeg g-pcc, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 1151–1155
29. L. Xie, W. Gao, H. Zheng, G. Li, Spcgc: scalable point cloud geometry compression for machine vision, in *2024 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2024), pp. 17 272–17 278
30. L. Xie, W. Gao, H. Zheng, H. Ye, Semantic-aware visual decomposition for point cloud geometry compression, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 595–595
31. Z. Qi, W. Gao, Variable-rate point cloud geometry compression based on feature adjustment and interpolation, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 63–72

32. Z. Yu, W. Gao, When dynamic neural network meets point cloud compression: computation-aware variable rate and checkerboard context, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 600–600
33. L. Xie, W. Gao, S. Fan, Z. Yao, Pdnet: parallel dual-branch network for point cloud geometry compression and analysis, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 596–596
34. L. Xie, W. Gao, H. Zheng, End-to-end point cloud geometry compression and analysis with sparse tensor, in *International Workshop on Advances in Point Cloud Compression, Processing and Analysis* (2022), pp. 27–32
35. C. Fu, G. Li, R. Song, W. Gao, S. Liu, Octattention: octree-based large-scale contexts model for point cloud compression. *Proc. AAAI Conf. Artif. Intell.* **36**(1), 625–633 (2022)
36. C. Zhang, W. Gao, Adapcc: adaptive rate control and rate-distortion-complexity optimization for dynamic point cloud compression, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
37. K. Wang, W. Gao, Unipcgc: towards practical point cloud geometry compression via an efficient unified approach, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
38. K. Zheng, W. Gao, H. Zheng, Octree-based learned point cloud geometry compression: a lossy perspective, in *2025 Data Compression Conference (DCC)* (IEEE, 2025)
39. Z. Qi, H. Liu, W. Gao, Variable-rate point cloud compression with channel-wise entropy modeling, in *2025 Data Compression Conference (DCC)* (IEEE, 2025)
40. H. Zheng, W. Gao, Z. Yu, T. Zhao, G. Li, Viewpcgc: view-guided learned point cloud geometry compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 7152–7161
41. L. Xie, W. Gao, H. Zheng, G. Li, Roi-guided point cloud geometry compression towards human and machine vision, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024)
42. Y. Wu, L. Xie, S. Sun, W. Gao, Y. Yan, Adaptive intra period size for deep learning-based screen content video coding, in *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2024)
43. H. Zheng, W. Gao, End-to-end rgb-d image compression via exploiting channel-modality redundancy. *Proc. AAAI Conf. Artif. Intell.* **38**(7), 7562–7570 (2024)
44. L. Tao, W. Gao, G. Li, C. Zhang, Adanic: towards practical neural image compression via dynamic transform routing, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 16 879–16 888
45. Y. Wu, W. Gao, End-to-end lossless compression of high precision depth maps guided by pseudo-residual (2022). [arXiv:2201.03195](https://arxiv.org/abs/2201.03195)
46. Y. Wu, Z. Qi, H. Zheng, L. Tao, W. Gao, Deep image compression with latent optimization and piece-wise quantization approximation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 1926–1930
47. W. Gao, L. Tao, L. Zhou, D. Yang, X. Zhang, Z. Guo, Low-rate image compression with super-resolution learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2020), pp. 154–155
48. W. Gao, S. Sun, H. Zheng, Y. Wu, H. Ye, Y. Zhang, Opendmc: an open-source library and performance evaluation for deep-learning-based multi-frame compression, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9685–9688
49. W. Gao, H. Zheng, C. Zhang, K. Zheng, Z. Yu, Y. Li, H. Ye, Y. Zhang, Opendic: an open-source library and performance evaluation for deep-learning-based image compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 202–11 205
50. Y. Li, W. Gao, G. Li, S. Ma, Saliency segmentation oriented deep image compression with novel bit allocation. *IEEE Trans. Image Process.* **34**, 16–29 (2025)
51. Y. Li, Z. Yu, X. Chen, S. Wang, H. Yuan, W. Gao, Zero-shot quantization for large-kernels via shape-based distribution and diversity self-distillation, in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2025)

52. X. Chen, Y. Wang, Y. Li, X. Ling, M. Li, R. Liu, M. Ouyang, K. Zhao, T. Guan, Y. He, Low bit-width zero-shot quantization with soft feature-infused hints for iot systems, *IEEE Internet of Things Journal* (2024)
53. C. Zhang, W. Gao, Learned rate control for frame-level adaptive neural video compression via dynamic neural network, in *European Conference on Computer Vision* (Springer, 2024)
54. Z. Yang, W. Gao, G. Li, Y. Yan, Sur-driven video coding rate control for jointly optimizing perceptual quality and buffer control. *IEEE Trans. Image Process.* (2023)
55. F. Shen, Z. Cai, W. Gao, An efficient rate control algorithm for intra frame coding in avs3, in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2021), pp. 3164–3169
56. H. Yuan, W. Gao, J. Wang, Dynamic computational resource allocation for fast inter frame coding in video conferencing applications, in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6
57. W. Gao, Q. Jiang, R. Wang, S. Ma, G. Li, S. Kwong, Consistent quality oriented rate control in hevc via balancing intra and inter frame coding. *IEEE Trans. Ind. Inform.* **18**(3), 1594–1604 (2021)
58. H. Yuan, W. Gao, A new coding unit partitioning mode for screen content video coding, in *Proceedings of the 2021 5th International Conference on Digital Signal Processing* (2021), pp. 66–72
59. W. Gao, On the performance evaluation of state-of-the-art rate control algorithms for practical video coding and transmission systems, in *Proceedings of the 2020 4th International Conference on Video and Image Processing* (2020), pp. 179–185
60. W. Gao, S. Kwong, Q. Jiang, C.-K. Fong, P.H. Wong, W.Y. Yuen, Data-driven rate control for rate-distortion optimization in hevc based on simplified effective initial qp learning. *IEEE Trans. Broadcast.* **65**(1), 94–108 (2018)
61. W. Gao, A multi-objective optimization perspective for joint consideration of video coding quality, in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (IEEE, 2019), pp. 986–991
62. W. Gao, S. Kwong, Y. Jia, Joint machine learning and game theory for rate control in high efficiency video coding. *IEEE Trans. Image Process.* **26**(12), 6074–6089 (2017)
63. W. Gao, S. Kwong, Y. Zhou, H. Yuan, Ssim-based game theory approach for rate-distortion optimized intra frame ctu-level bit allocation. *IEEE Trans. Multimedia* **18**(6), 988–999 (2016)
64. W. Gao, S. Kwong, H. Yuan, X. Wang, Dct coefficient distribution modeling and quality dependency analysis based frame-level bit allocation for hevc. *IEEE Trans. Circuits Syst. Video Technol.* **26**(1), 139–153 (2015)
65. W. Gao, S. Kwong, Phase congruency based edge saliency detection and rate control for perceptual image and video coding, in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2016), pp. 000 264–000 269
66. H. Yuan, W. Gao, Openfastvc: an open source library for video coding fast algorithm implementation, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9660–9663
67. H. Yuan, W. Gao, S. Ma, Y. Yan, Divide-and-conquer-based rdo-free cu partitioning for 8k video compression. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(4), 1–20 (2024)
68. L. Tao, W. Gao, A hardware implementation of entropy encoder for 8k video coding, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
69. Y. Guo, W. Gao, S. Ma, G. Li, Accelerating transform algorithm implementation for efficient intra coding of 8k uhd videos. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* **18**(4), 1–20 (2022)
70. Z. Cai, W. Gao, Efficient fast algorithm and parallel hardware architecture for intra prediction of avs3, in *IEEE International Symposium on Circuits and Systems (ISCAS)* (IEEE, 2021), pp. 1–5
71. W. Gao, H. Yuan, Y. Guo, L. Tao, Z. Cai, G. Li, Openhardwarevc: an open source library for 8k uhd video coding hardware implementation, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 7339–7342

72. W. Gao, H. Yuan, G. Liao, Z. Guo, J. Chen, Pp8k: a new dataset for 8k uhd video compression and processing. *IEEE MultiMedia* (2023)
73. W. Liu, W. Gao, G. Li, S. Ma, T. Zhao, H. Yuan, Enlarged motion-aware and frequency-aware network for compressed video artifact reduction. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
74. A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, A.A. Bharath, Generative adversarial networks: an overview. *IEEE Signal Process. Mag.* **35**(1), 53–65 (2018)
75. Z. Li, F. Liu, W. Yang, S. Peng, J. Zhou, A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Trans. Neural Netw. Learn. Syst.* **33**(12), 6999–7019 (2021)
76. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al., An image is worth 16x16 words: Transformers for image recognition at scale (2020). [arXiv:2010.11929](https://arxiv.org/abs/2010.11929)
77. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need (2017). [arXiv:1706.03762](https://arxiv.org/abs/1706.03762)
78. C.-Y. Liou, W.-C. Cheng, J.-W. Liou, D.-R. Liou, Autoencoder for words. *Neurocomputing* **139**, 84–96 (2014)
79. D.P. Kingma, M. Welling et al., An introduction to variational autoencoders. *Found. Trends® Mach. Learn.* **12**(4), 307–392 (2019)
80. J. Ballé, V. Laparra, E.P. Simoncelli, End-to-end optimized image compression (2016). [arXiv:1611.01704](https://arxiv.org/abs/1611.01704)
81. J. Ballé, D. Minnen, S. Singh, S.J. Hwang, N. Johnston, Variational image compression with a scale hyperprior (2018). [arXiv:1802.01436](https://arxiv.org/abs/1802.01436)
82. L. Xie, X. Mu, G. Li, W. Gao et al., Pku-dpcc: a new dataset for dynamic point cloud compression. *APSIPA Trans. Signal Inf. Process.* **13**(6) (2024)
83. H. Yuan, W. Gao, W. Gao, Opensep: an open source subjective experiment platform, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 248–11 251
84. S. Fan, W. Gao, Screen-based 3d subjective experiment software, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9672–9675
85. J. Wang, W. Gao, G. Li, Zoom to perceive better: No-reference point cloud quality assessment via exploring effective multiscale feature. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
86. J. Wang, W. Gao, G. Li, Applying collaborative adversarial learning to blind point cloud quality measurement. *IEEE Trans. Instrum. Meas.* (2023)
87. H. Li, W. Gao, Information exploration of projected views for point cloud quality measurement. *IEEE Trans. Instrum. Meas.* (2024)
88. B. Qu, X. Liang, S. Sun, W. Gao, Exploring aigc video quality: a focus on visual harmony, video-text consistency and domain distribution gap, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2024)
89. B. Qu, H. Li, W. Gao, Bringing textual prompt to ai-generated image quality assessment, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024)
90. S. Sun, X. Liang, S. Fan, W. Gao, W. Gao, Ve-bench: Subjective-aligned benchmark suite for text-driven video editing quality assessment, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
91. Z. Guo, W. Gao, H. Wang, J. Wang, S. Fan, No-reference deep quality assessment of compressed light field images, in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6
92. D. Minnen, J. Ballé, G. Toderici, Joint autoregressive and hierarchical priors for learned image compression (2018). [arXiv:1809.02736](https://arxiv.org/abs/1809.02736)
93. E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, L.V. Gool, Generative adversarial networks for extreme learned image compression, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 221–231
94. W. Liu, W. Gao, X. Mu, Fast inter-frame motion prediction for compressed dynamic point cloud attribute enhancement. *Proc. AAAI Conf. Artif. Intell.* **38**(4), 3720–3728 (2024)

95. Z. Yang, W. Gao, X. Lu, Danet: density-adaptive network for geometry-based point cloud compression artifacts removal, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
96. X. Fan, G. Li, D. Li, Y. Ren, W. Gao, T.H. Li, Deep geometry post-processing for decompressed point clouds, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
97. X. Zhang, G. Liao, W. Gao, G. Li, Tdrnet: transformer-based dual-branch restoration network for geometry based point cloud compression artifacts, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
98. Z. Li, G. Li, T. H. Li, S. Liu, W. Gao, Semantic point cloud upsampling. *IEEE Trans. Multimedia* (2022)
99. Z. Li, S. Liu, W. Gao, G. Li, G. Li, S4r: rethinking point cloud sampling via guiding upsampling-aware perception. *IEEE Trans. Multimedia* (2025)
100. R. Zhang, W. Gao, G. Li, T.H. Li, Qinet: decision surface learning and adversarial enhancement for quasi-immune completion of diverse corrupted point clouds. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–14 (2022)
101. R. Bao, Y. Ren, G. Li, W. Gao, S. Liu, Flow-based point cloud completion network with adversarial refinement, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2559–2563
102. J. Chen, G. Li, R. Zhang, T.H. Li, W. Gao, Pointvae: invertible variational autoencoder framework for 3d point cloud generation, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 3216–3220
103. R. Zhang, J. Chen, W. Gao, G. Li, T.H. Li, Pointot: interpretable geometry-inspired point cloud generative model via optimal transport. *IEEE Trans. Circuits Syst. Video Technol.* **32**(10), 6792–6806 (2022)
104. Z. Li, G. Li, T. Li, S. Liu, W. Gao, Information-growth attention network for image super-resolution, in *Proceedings of the 29th ACM International Conference on Multimedia* (2021), pp. 544–552
105. L. Zhou, W. Gao, G. Li, H. Yuan, T. Zhao, G. Yue, Disentangled feature distillation for light field super-resolution with degradations, in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2023), pp. 116–121
106. L. Zhou, W. Gao, G. Li, End-to-end spatial-angular light field super-resolution using parallax structure preservation strategy, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 3396–3400
107. W. Gao, L. Zhou, L. Tao, A fast view synthesis implementation method for light field applications. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* **17**(4), 1–20 (2021)
108. X. Zhang, W. Gao, G. Li, Q. Jiang, R. Cong, Image quality assessment-driven reinforcement learning for mixed distorted image restoration. *ACM Trans. Multimedia Comput. Commun. Appl.* **19**(1s), 1–23 (2023)
109. X. Zhang, W. Gao, H. Yuan, G. Li, Je 2 net: joint exploitation and exploration in reinforcement learning based image restoration, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2090–2094
110. X. Zhang, W. Gao, Hirl: hybrid image restoration based on hierarchical deep reinforcement learning via two-step analysis, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2445–2449
111. S. Wang, Y. Lu, W. Xia, P. Xia, Z. Wang, W. Gao, Light field angular super-resolution by view-specific queries. *Visual Comput.* 1–16 (2024)
112. Y. Sun, Z. Li, S. Wang, W. Gao, Depth-assisted calibration on learning-based factorization for a compressive light field display. *Opt. Express* **31**(4), 5399–5413 (2023)
113. Y. Sun, Z. Li, L. Li, S. Wang, W. Gao, Optimization of compressive light field display in dual-guided learning, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2075–2079
114. W. Gao, S. Fan, G. Li, W. Lin, A thorough benchmark and a new model for light field saliency detection. *IEEE Trans. Pattern Anal. Mach. Intell.* (2023)

115. G. Li, G. Wei, W. Gao, *Point Cloud Compression: Technologies and Standardization* (Springer Nature, 2024)
116. G. Li, W. Gao, W. Gao, Introduction, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 1–28
117. G. Li, W. Gao, W. Gao, Background knowledge, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 29–51
118. G. Li, W. Gao, W. Gao, Predictive coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 53–70
119. G. Li, W. Gao, W. Gao, Transform coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 71–96
120. G. Li, W. Gao, W. Gao, Quantization techniques, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 97–112
121. G. Li, W. Gao, W. Gao, Entropy coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 113–133
122. G. Li, W. Gao, W. Gao, Mpeg geometry-based point cloud compression (g-pcc) standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 135–165
123. G. Li, W. Gao, W. Gao, Avs point cloud compression standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 167–197
124. G. Li, W. Gao, W. Gao, Mpeg video-based point cloud compression (v-pcc) standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 199–218
125. G. Li, W. Gao, W. Gao, Mpeg ai-based 3d graphics coding standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 219–241
126. G. Li, W. Gao, W. Gao, Future work, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 243–250

Chapter 3

Quality Assessment and Perception

Models for Image and Video



3.1 Introduction

Assessing the image and video quality is essential for ensuring high-quality human experiences in various multimedia applications, such as video conferencing and broadcasting [1–83]. The human visual system (HVS) plays a crucial role in perceiving visual distortions, making it necessary to develop models that align with human perception. Quality assessment (QA) methods can be broadly classified into subjective and objective approaches. Subjective QA relies on human evaluators to rate visual quality, providing the most accurate reflection of perceptual experience. However, it is time-consuming, costly, and influenced by external factors such as viewing conditions and individual biases. Objective QA, on the other hand, employs computational models to predict quality automatically. The objective quality assessment modeling methods can be divided into three types, including full-reference, reduced-reference, and no-reference. Recent advancements highlight the effectiveness of deep learning in capturing complex distortions and aligning closely with human visual perception [84–93, 93–126].

Just-Noticeable-Distortion (JND) is a fundamental concept within QA. It is used to distinguish perceptible distortions from imperceptible ones, ensuring that objective QA models focus on distortions that significantly impact visual quality. It serves as a perceptual threshold, enabling QA models to disregard imperceptible distortions and focus on those that significantly impact perceived quality. By integrating JND into QA processes, models become more aligned with human visual perception, enhancing the accuracy and reliability of quality assessments. This alignment allows QA models to predict quality scores that better correlate with subjective human evaluations.

Another critical aspect is saliency detection, which identifies visually prominent regions to prioritize in quality evaluation. Acknowledging the limitations inherent in the human visual nervous system, which render it incapable of simultaneously apprehending or processing all available visual information, the cognitive processes

naturally gravitate towards focusing on salient components. This detection mechanism is thus rooted in capitalizing on this physiological constraint, aiming to intelligently and autonomously isolate visually compelling targets within a scene. Consequently, it facilitates expedited information processing by directing attention to the most pertinent aspects, thereby augmenting the efficiency of the visual cognition. In QA, saliency detection highlights regions that have the most impact on perceptual quality, allowing models to prioritize these areas for evaluation. By aligning JND with saliency, QA methods can refine their accuracy further, ensuring assessments that closely match human perception and improving reliability in multimedia contexts.

This chapter introduces these concepts in detail, covering subjective and objective QA methods, JND modeling, and saliency-based approaches for both images and videos. The discussion highlights state-of-the-art techniques, their advantages, limitations, and applications in modern multimedia systems.

3.2 Quality Assessment of Image and Video

Image Quality Assessment (IQA) [127] is a crucial area in the field of image processing [42–51, 85] and computer vision [10–41, 93, 112–114, 128–132, 132–162], aimed at evaluating the perceptual quality of images. The increasing prevalence of image-based applications, ranging from multimedia to medical imaging, necessitates robust and reliable methods to ensure images meet quality standards. IQA problems can be broadly categorized into two kinds, including the subjective experiments and the objective modeling.

Subjective IQA involves human observers who evaluate image quality based on personal perceptions. This method is considered the gold standard since the ultimate users of images are humans. However, subjective evaluations are often impractical for real-world applications due to their time-consuming and resource-intensive nature. Factors such as viewing distance, display device, and observers' physical and emotional conditions can also affect the results.

Objective IQA aims to develop mathematical models that predict image quality automatically and efficiently, emulating human judgment as closely as possible. Objective methods can be divided into three types depending on whether a reference image is available, and they can be Full-Reference (FR-IQA), Reduced-Reference (RR-IQA), and No-Reference (NR-IQA).

Meanwhile, Video Quality Assessment (VQA) is crucial for ensuring that the visual quality of video content meets user expectations during transmission, processing, and storage. High video quality is essential for applications such as streaming, video conferencing, surveillance, and virtual reality, directly impacting user experience. However, VQA faces several challenges: the subjective nature of human visual perception, diverse distortion types, real-time processing requirements, and the large data volumes of high-definition and ultra-high-definition videos.

Current VQA methods are categorized into subjective and objective approaches. Subjective methods, e.g., the Absolute Category Rating (ACR) and the Double Stim-

ulus Continuous Quality Scale (DSCQS), rely on human viewers to rate video quality, providing the most accurate user experience feedback but at a high cost and limited real-time applicability. Objective methods include Full-Reference (FR), No-Reference (NR), and Reduced-Reference (RR) approaches. FR methods compare the processed videos with the original references using metrics like PSNR, SSIM, and VMAF. NR methods predict video quality without reference videos, often using machine learning models, while RR methods use partial reference information, balancing between FR and NR methods. These VQA methods effectively detect and quantify visual quality, with subjective methods offering authentic feedback and objective methods providing quick, efficient evaluations. Deep learning-based VQA methods are particularly promising for handling complex distortions and delivering precise quality predictions [8, 41].

Next, the details of subjective quality assessment and objective quality assessment are introduced.

3.2.1 Subjective Quality Assessment

Subjective IQA involves evaluating the quality of images based on human judgment. It is considered the gold standard for image quality assessment as it directly reflects the perceptual experiences of the end-users, who are typically human observers. Subjective assessments provide the most accurate representation of perceived image quality, but they are also resource-intensive and can be influenced by various external factors.

Subjective IQA employs several standardized testing methods to ensure reliable results. Some of the common methods are listed in Table 3.1. Mean Opinion Score (MOS) [163] is a numerical measure of multimedia quality based on the average ratings given by a group of human evaluators on a predefined scale. In Single Stimulus Categorical Rating [164] method, test images are shown to observers for a fixed duration, after which they rate the image quality on a categorical scale, i.e., the excellent, good, fair, poor, bad scales. For Double Stimulus Categorical Rating [165], we can show the test image and the reference image, and observers rate the quality of the test image based on the reference image. For Ordering by Forced-Choice Pair-Wise Comparison [166], the pairs of images are shown, and observers choose the image with higher quality. Pair-Wise Similarity Judgments [167] need observers to choose the higher-quality image and to indicate the difference level.

Subjective video quality assessment [169] involves gathering ratings from sufficient participants, who evaluate the quality of the target videos based on the predefined criteria. The Mean Opinion Score (MOS) and the Difference Mean Opinion Score (DMOS) are used to represent the subjective quality. MOS represents the average rating for specific stimuli, while DMOS quantifies the mean difference between reference and distorted stimuli ratings, effectively minimizing the impact of video content by adjusting for reference material scores. The primary assessment

Table 3.1 Subjective image quality assessment methods

Method	Advantages	Disadvantages
Mean Opinion Score (MOS) [163]	Simple and easy to understand, widely used, reflects subjective perception of viewers	Highly influenced by individual subjectivity, time-consuming, and costly
Double Stimulus Continuous Quality Scale (DSCQS) [165]	Reduces random errors in scoring, improves reliability of evaluation results	Complex experiments, long evaluation time, high demands on viewer fatigue
Single Stimulus Continuous Quality Evaluation (SSCQE) [164]	Suitable for evaluating dynamic video, can reflect changes in image quality in real-time	Viewer subjective perception easily influenced by environmental factors, high variability in evaluation results
Double Stimulus Impairment Scale (DSIS) [168]	Accurately distinguish between good and poor image quality, reduces subjective differences in scoring	Long evaluation time, requires training for viewers, costly
Ordering by Forced-Choice Pair-Wise Comparison [166]	Suitable for various types of content and straightforward for participants	Data analysis and experimental design can be complex and require careful planning
Pair-Wise Similarity Judgments [167]	Comprehensive evaluation of multiple aspects of images, including clarity, color, motion, etc.	Complex experiment, long evaluation time, requires some level of viewer expertise

Source Author

methodologies include Absolute Category Rating (ACR), Comparison Category Rating (CCR), and Degradation Category Rating (DCR).

Table 3.2 summarizes six databases commonly employed in visual quality assessment research for general video content. Among these, the LIVE Video Quality Assessment Database (LIVE-VQA) [170] is the widely-used database which includes the 10 pristine and the 160 distorted videos (768×432 resolution, 25/50 fps, 10 s duration) in the formats of YUV and H.264, along with subjective data such as DMOS and standard deviation (σ). The IVP database for subjective video quality evaluation, on the other hand, comprises 10 original and 138 impaired videos, all in 1080p resolution, 25 frames per second, and a 10 s duration in YUV format. It also offers subjective metrics such as DMOS and σ .

Subjective quality assessment serves as the gold standard for evaluating perceptual quality, as it directly mirrors human judgment. However, its inherent limitations—such as uncontrollability and high implementation costs often render it impractical for large-scale or real-world applications. Consequently, objective quality assessment methods have emerged as indispensable alternatives, offering automated, efficient, and scalable solutions. Despite their advantages, objective metrics must be rigorously validated and calibrated against subjective benchmarks to ensure alignment with human perception. Thus, while objective methods dominate practical implementations, subjective assessment remains a critical foundation for their development and refinement.

Table 3.2 Video quality assessment dataset

Dataset	Description
LIVE Video Quality Database [170] ^a	Contains videos with various types and levels of distortion, commonly used for developing and testing video quality assessment algorithms
CSIQ Video Quality Database [171] ^b	Includes videos with multiple distortion types and encoding settings, providing subjective scores for model training and validation
VQEG HD3 Database [172] ^c	Provided by the Video Quality Experts Group (VQEG), includes HD video clips with subjective ratings
Waterloo Video Quality Database [173] ^d	Contains videos with multiple distortion types and scenes, used for research in no-reference video quality assessment
Netflix Public Dataset [174, 175] ^e	Includes public datasets provided by Netflix for research on video quality assessment, with encoded videos and corresponding subjective scores
TID2013 [176] ^f	Contains images and videos with 24 types of distortion, used for testing full-reference and no-reference quality assessment algorithms

Source Author

^ahttps://live.ece.utexas.edu/research/quality/live_video.html

^b<https://s2.smu.edu/~eclarson/csiq.html>

^c<http://www.cdvl.org/>

^d<https://ivc.uwaterloo.ca/database/Video-Database>

^e<https://www.kaggle.com/datasets/shivamb/netflix-shows>

^f<https://ponomarenko.info/tid2013.html>

3.2.2 Objective Quality Assessment

Image Quality Assessment

Objective IQA [177] aims to create computational models that can accurately and automatically predict perceived image quality. In contrast to subjective IQA methods, which rely on human judgments, objective IQA employs algorithms to provide a quantitative evaluation of image quality. Objective IQA techniques generally have three types, including Full-Reference (FR-IQA), Reduced-Reference (RR-IQA), and No-Reference (NR-IQA). Table 3.4 presents representative objective IQA methods.

FR-IQA methods assess image quality by the means of comparing the distorted image against the undistorted reference image, assuming that the reference is fully accessible and serves as a benchmark for quality assessment. The standard FR-IQA framework involves several stages, including preprocessing, channel decomposition, feature extraction, divergence calculation, and pooling to yield a final quality score. Metrics like Peak Signal-to-Noise Ratio (PSNR) [178] and Structural Similarity Index (SSIM) [179] are commonly employed in FR-IQA. The PSNR quantifies the ratio of the maximum possible image signal power and the power of noise that distorts the image. A higher PSNR value indicates lower distortion and, therefore,

Table 3.3 Objective image quality assessment methods

Method	Type	Advantages	Disadvantages
Peak Signal-to-Noise Ratio (PSNR) [178]	FR	Simple to calculate, widely used, easy to understand and implement	Does not reflect human visual perception well, ignores structural information in images
Structural Similarity Index (SSIM) [179]	FR	Considers luminance, contrast, and structural information, closer to human visual perception	Sensitive to noise and occlusions, relatively high computational complexity
Reduced-Reference Entropic Differences (RRED) [180]	RR	Utilizes statistical characteristics of images, provides accurate quality assessment	Requires partial reference image information, high computational complexity
Blind Image Spatial Quality Evaluator (BRISQUE) [181]	NR	No reference image needed, utilizes natural scene statistics for quality assessment, and efficient computation	Sensitivity to different types of distortions varies, dependent on training data
Natural Image Quality Evaluator (NIQE) [181]	NR	No reference image needed, based on natural scene statistics and deep learning models, captures complex image features	Requires large datasets for training, high computational complexity, and high demand on computational resources
Learning-based Image Quality Assessment methods [182]	NR	Captures complex image features and are highly correlated with the human visual system	Requires a large amount of training data, and the training process and computational complexity are high

Source Author

better perceived image quality. While PSNR is a simple and computationally efficient metric for assessing image quality, its limitations make it inadequate for applications requiring high perceptual accuracy. For a more reliable assessment of image quality, especially in contexts where human perception is critical, metrics that incorporate aspects of the human visual system, such as SSIM, are preferred. PSNR can still be useful for preliminary assessments and comparisons, particularly in technical contexts where perceptual factors are less critical.

RR-IQA approaches rely on limited information from the reference image, making them effective in situations where the reference image is partially unavailable. In image transmission systems, characteristics are derived from the source image at the transmitting terminal and transmitted to the receiving end. On the receiving side, we can obtain features of the degraded image, and then contrasted with those from the original image. This comparison facilitates image quality evaluation. RR-IQA metrics typically employ features that maintain robustness and exhibit minimal sensitivity to transmission distortions (Tables 3.3 and 3.4).

Table 3.4 Learning-based image quality assessment methods

Method	Advantages	Disadvantages
DeepRN [183]	Simple to calculate, widely used, easy to understand and implement	Does not reflect human visual perception well, ignores structural information in images
Two-stream convolutional network [184]	Considers luminance, contrast, and structural information, closer to human visual perception	Sensitive to noise and occlusions, relatively high computational complexity
Hallucinated-IQA [185]	Simple to calculate, easy to implement	Insensitive to small differences in images, does not reflect human visual quality perception
Transformers in image quality assessment (TRIQ) [186]	Utilizes statistical characteristics of images, provides accurate quality assessment	Requires partial reference image information, high computational complexity
TReS [187]	No reference image needed, utilizes natural scene statistics for quality assessment, and efficient computation	Sensitivity to different types of distortions varies, dependent on training data

Source Author

NR-IQA, also referred to as blind IQA, assesses image quality independently of a reference image. This task is inherently complex, as it relies solely on the image’s inherent attributes to estimate quality. NR-IQA models are typically trained on extensive datasets with annotated quality scores, enabling them to predict the quality of unseen images. Features such as Natural Scene Statistics (NSS) are pivotal in these models. The ultimate objective is to create a generalizable NR-IQA metric capable of handling a wide range of distortions. A prominent method in this domain is the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [181], which operates by extracting NSS in the spatial domain and using a Support Vector Machine to predict the perceived quality without the reference information. Expanding on the NSS-based framework, the Natural Image Quality Evaluator (NIQE) [181] introduces an entirely blind, unsupervised approach, comparing the NSS of the target image to a statistical model derived from pristine images. Further enhancing NR-IQA capabilities, deep learning-based techniques [182] leverage convolutional neural networks and other advanced algorithms to automatically extract complex features and predict image quality, offering increased accuracy and robustness across diverse distortion types.

Recent advancements in IQA have focused on deep learning incorporating human visual system characteristics to improve the accuracy of these models. A part of the deep learning-based IQA methods are illustrated on Table 3.5. For example, Deep

RN [183] and Two-stream model [184] consider deep neural network-based NR-IQA methods with aspects of human visual perception to provide more reliable quality assessments. Varga et al. [183] proposed DeepRN, a blind image quality assessment method using ResNet-101 with adaptive spatial pyramidal pooling. Based on popular datasets, DeepRN achieved leading performance in quality assessment. In the same year, Yan et al. [184] developed the two-stream convolutional network using an image and a gradient map as inputs to extract different information levels. The final image score is averaged from block scores. With the development of Generative Adversarial Networks, Lin et al. [185] introduced Hallucinated-IQA that predicts difference between illusionary reference image and distorted image. The Generative Adversarial Network synthesizes numerous realistic images that closely resemble those found in the IQA database. Firstly introduced in 2017 as an advancement over Recurrent Neural Networks for natural language processing, Transformer has since gained considerable attention in the field of image quality assessment [186–189]. For example, You et al. [186] investigated the application of Transformers to image quality evaluation, presenting the TRIQ framework. Their method employs the Vision Transformer (ViT), which incorporates a Transformer encoder with adaptive positional embeddings, allowing it to process images at different resolutions. Golestaneh et al. [187] introduced an innovative strategy named TReS, which combines Transformers, Relative Ranking, and Self-Consistency. This technique enhances robustness and accuracy by integrating both local and global feature extraction, relative image quality ranking, and a self-consistency mechanism.

Video Quality Assessment

The increasing demand for high-quality video content across diverse applications, including streaming services and video conferencing, has driven substantial interest in VQA [190, 191]. VQA approaches are generally classified into three main categories: full-reference (FR VQA), reduced-reference (RR VQA), and no-reference (NR VQA) methods.

FR VQA methods [192] rely on the availability of the pristine undistorted video, against which the quality of the distorted video is compared. These methods are predominantly knowledge-driven, utilizing domain-specific knowledge of human visual systems (HVS) or low-level features such as structural similarity (SSIM) [193]. Additionally, feature fusion techniques enhance performance by combining multiple feature sets to better approximate human perception. These techniques are designed based on principles such as SSIM, which measures structural degradation, and HVS-based models that take into account human sensitivity to visual distortions.

RR VQA approaches [194] operate with partial information about the reference video. These approaches usually derive a collection of features from the pristine video and utilize them to evaluate the quality of the target video. Similar to FR VQA, RR VQA approaches are often knowledge-driven, relying on handcrafted features to evaluate quality. However, unlike FR methods, they use fewer reference points, making them more practical in real-world scenarios where full-reference videos may not be available.

NR VQA methods [195, 196] are applied when no reference video is available, making them particularly useful for real-time and large-scale applications. These methods are predominantly data-driven and often employ machine learning techniques, including pre-trained and end-to-end neural networks. Recent advances also incorporate self-supervised learning, which can further enhance their robustness and generalization capabilities. Traditional NR-VQA methods have typically depended on handcrafted features, such as natural scene statistics (NSS) and discrete cosine transform (DCT) coefficients, to capture and model spatial and temporal artifacts in videos. Approaches like TLVQM [197] and RAPIQUE [198] exemplify efficient handcrafted feature-based models that balance predictive accuracy with computational efficiency. While these techniques utilize large-scale datasets to train deep learning models for reference-free video quality prediction, the end-to-end design of such models enables them to extract complex, high-level features, enhancing their accuracy in predicting perceived video quality.

While general-purpose methods provide a broad evaluation framework, specific-purpose methods target particular video characteristics or scenarios [199]. Compressed VQA methods [200, 201] assess the quality of videos that have undergone compression. This is a highly relevant area due to the widespread use of video compression algorithms in streaming services. These methods must account for artifacts such as blocking, blurring, and ringing that typically arise from compression. Streaming VQA methods [202, 203] focus on assessing videos streamed in real-time over the internet. Streaming introduces unique challenges, such as varying bandwidth and network latency, which can lead to quality degradations like stalling, jitter, and frame freezing. As streaming becomes more prevalent with services like YouTube, Netflix, and gaming platforms, these methods are increasingly vital. Stereo Video Quality Assessment methods [204, 205] are specifically designed to assess the quality of stereoscopic (3D) video content, which introduces challenges like depth perception and binocular disparity. These methods are crucial for ensuring accurate and immersive experiences in applications ranging from 3D films to medical imaging. With the growing integration of virtual reality and augmented reality technologies in sectors like entertainment, training, and industry, VR/AR VQA methods [206, 207] have become essential for evaluating the quality of immersive content. These approaches account for unique factors such as field-of-view, head tracking, and rendering latency, all of which significantly impact user experience in VR and AR settings.

Recent advancements in deep learning have significantly transformed NR-VQA by utilizing convolutional neural networks and recurrent architectures to automatically capture spatiotemporal features. While these models achieve strong performance, their high computational demands and substantial data requirements restrict their practical applications [208, 209]. Future research should aim to develop more comprehensive models capable of addressing both general and specific video quality issues by harnessing advancements in deep learning, self-supervised learning, and multi-modal learning.

In conclusion, objective IQA methods are essential for numerous applications, including image compression, transmission, and enhancement. They offer a practical and efficient way to ensure image quality in various real-world scenarios, complementing subjective evaluations and providing valuable tools for the development and optimization of image processing systems.

3.3 Just-Noticeable-Distortion of Image and Video

Just-Noticeable-Distortion is an important vision concept in the human visual system, which represents the minimum distortion threshold that humans can perceive. Due to the limited discrimination ability of the human eye, it is impossible to detect subtle visual impairments, which indicates that HVS has a certain tolerance to changes in visual signals. Thus, Just-Noticeable-Distortion can be used as an indicator of perceptual redundancy in the visual signal. The corresponding Just-Noticeable-Distortion model is an effective way to estimate visual redundancy and has been widely used in various fields. First of all, in the image processing [210, 211], Just-Noticeable-Distortion can examine the distortion model can measure the sensitivity of the human eye to different regions of image distortion. Considering that humans are the receivers of images/videos, the visual redundancy in images can be effectively eliminated by exploiting the characteristics of HVS. Second, in the field of video coding [212–215], Just-Noticeable-Distortion can examine the distortion model can be used to further improve the video coding efficiency. By simulating the perceptual characteristics inherent in the human eye perception system, accurate mode decision as well as the elimination of spatial and temporal redundancy can be achieved more efficiently while considering the compression objective. In addition, Just-Noticeable-Distortion also has important potential applications in quality assessment [216, 217] and digital watermarking [218, 219] field. It plays a key role in evaluating perceptual redundancy and optimizing coding efficiency [8, 9, 41].

The rest of this section are illustrated as follows. The datasets for Just-Noticeable-Distortion are firstly introduced. Next, representative pixel-domain and frequency-domain methods is introduced. Finally, the methods of image and video coding with just-noticeable-distortion are presented.

3.3.1 *Just-Noticeable-Distortion Datasets*

In the process of constructing Just-Noticeable-Distortion dataset, many rounds of comparison are needed to arrive at the final judgment of Just-Noticeable-Distortion, so the subjective labeling experiment of Just-Noticeable-Distortion is very time-consuming. In addition, the observation of Just-Noticeable-Distortion requires the subjects to make a detailed observation comparison between the two visual content models at the same time, which is relatively subjective and volatile. Due to

a variety of factors, the existing Just-Noticeable-Distortion datasets are still very scarce, which largely hinders the research of Just-Noticeable-Distortion modeling. Currently, the public Just-Noticeable-Distortion datasets are mainly for compressed degraded images/videos.

The first Just-Noticeable-Distortion dataset MCL-JCI¹ [220] consists of 50 source images and corresponding JPEG encoded compressed images, each with 100 quality levels. The study observed that people can only distinguish between a limited number of quality levels and perceived quality is described as a step function. As the first image Just-Noticeable-Distortion dataset in the industry, the MCL-JCI dataset helps to understand the perception change patterns of ordinary people at different quality levels, provides important resources and valuable research insights for statistical analysis of perceived image quality, and provides reliable data support for further research.

Subsequently, Wang et al. [221] proposed a compressed video quality assessment dataset MCL-JCV² based on Just-Noticeable-Distortion model to evaluate video quality in H.264/AVC compression. MCL-JCV includes 30 video clips with different contents as sources, and lossy videos are obtained by applying 50 different quality levels of H.264/AVC compression. This dataset addresses the limitations of existing video quality assessment datasets by focusing on the discreteness of perceived video quality based on JND points. They proposed to approximate the Just-Noticeable-Distortion as a Gaussian distribution, and used the Expectation Maximization algorithm for joint Just-Noticeable-Distortion modeling and processing to better fit the collected Just-Noticeable-Distortion data. In addition, the JND distribution is modeled by combining multiple Just-Noticeable-Distortion points with Gaussian Mixture Model.

Considering that the previous MCL-JCI and MCL-JCV are small-scale Just-Noticeable-Distortion datasets, Wang et al. [222] published a large-scale coded video quality dataset based on Just-Noticeable-Distortion, named VideoSet.³ 220 video clips with four resolutions are included. It uses the H.264/AVC codec to provide 51 compression levels for each sequence. Research shows that the algorithm and strategy of Just-Noticeable-Distortion search process are very important for video quality assessment, and the distribution modeling of Just-Noticeable-Distortion points is helpful to understand the characteristics of video content. By analyzing the experimental results, the post-processed Just-Noticeable-Distortion data show the characteristics of Gaussian distribution after normality test, which indicates that it can be approximated as Gaussian random variable, and further reveals the internal mechanism of Just-Noticeable-Distortion.

In addition, Liu et al. [223]⁴ constructed the first image Just-Noticeable-Distortion dataset covering various distortion types. The dataset includes 106 source images and 25 distortion types, totaling 1,642 distorted images. This study selects distorted

¹ <https://mcl.usc.edu/mcl-jci-dataset/>.

² <https://mcl.usc.edu/mcl-jci-dataset/>.

³ <https://iee-dataport.org/documents/videoset>.

⁴ <https://github.com/jianjin008>.

Table 3.5 A list of just-noticeable-distortion datasets for images/videos

Dataset	Pub year	Source no.	Distortion no.	Distortion type
MCL-JCI [220] ^a	2016	50	100	JPEG compression
MCL-JCV [221] ^b	2016	30	50	H.264/AVC compression
VideoSet [222] ^c	2017	880	51	H.264/AVC compression
Liu et al. [223] ^d	2023	105	1642	Additive Gaussian noise, Color shift, and so on

Source Author

^a<https://mcl.usc.edu/mcl-jci-dataset/>

^b<https://mcl.usc.edu/mcl-jcv-dataset/>

^c<https://ieee-dataport.org/documents/videoset>

^d<https://github.com/jianjin008>

images from existing IQA datasets based on MOS and subjective evaluation, and highlights the importance of considering different distortion types in the modeling of Just-Noticeable-Distortion. A coarse-to-fine Just-Noticeable-Distortion selection process is also proposed.

3.3.2 Pixel-Domain Just-Noticeable-Distortion Model

Pixel-domain Just-Noticeable-Distortion models provide an intuitive threshold of Just-Noticeable-Distortion in the pixel domain, and usually combine Luminance Adaptation (LA) and spatial Contrast Masking (CM) to estimate pixel-level Just-Noticeable-Distortion maps. Chou et al. [224] proposed the first pixel-domain model for Just-Noticeable-Distortion estimation, which takes into account background luminance adaptation and texture masking effects within a perceptual framework. The authors emphasize the importance of removing perceptually insignificant components from image signals, apply Just-Noticeable-Distortion configurations to quantify perceptual redundancy, and provide distorted visibility thresholds for each signal to achieve image coding with improved visual quality at low bitrates. Since then, Wu et al. [225] developed a method to estimate the Just-Noticeable-Distortion threshold of an image by combining brightness differences and structural regularity in spatial masking. The authors measure the structural regularity in the image by adaptive non-local self-similarity analysis, and then use the similarity coefficients to weight the brightness difference, thereby deriving a new spatial masking function to accurately estimate the Just-Noticeable-Distortion.

Prior pixel-based Just-Noticeable-Distortion models tend to underestimate the JND threshold for disordered regions. Therefore, Wu et al. [226] developed an image Just-Noticeable-Distortion estimation model based on the free energy principle and an autoregressive model that simulates the human brain and considers both ordered and disordered regions. By deriving an autoregressive model to simulate the

active prediction process of the brain, the proposed Just-Noticeable-Distortion model considers both ordered and disordered regions. To better predict Just-Noticeable-Distortion thresholds, Wu et al. [227] introduced an innovative pattern masking function that integrates the luminance contrast and the structural uncertainty. They employed an autoregressive model to isolate unpredictable uncertainty from the input image, and applied an adjusted local binary pattern technique to determine structural uncertainty. The proposed pattern masking function shows better consistency with human vision system. Wu et al. [228] introduced an enhanced Just-Noticeable-Distortion model by incorporating pattern complexity as a key factor for the overall masking effect, while accounting for the interplay between visual elements. They highlighted that the interaction within regular patterns is relatively straightforward, resulting in a weaker visual masking effect, whereas irregular patterns exhibit complex interactions and a stronger visual masking effect. By integrating pattern complexity and luminance contrast, they developed a new spatial masking estimation function, thereby refining the Just-Noticeable-Distortion estimation model. Shen et al. [229] designed Just-Noticeable-Distortion model based on block decomposition strategy and simulated structural visibility masking by neural network. The proposed model decomposes an image patch into three independent components for just noticeable estimation and learns structural degradation using a deep learning-based model. Wang et al. [230] proposed a novel approach for Just-Noticeable-Distortion estimation in image compression based on hierarchical predictive coding theory. They modeled the human visual system as a three-level communication system, which divided visual perception into three stages. The proposed just noticeable estimation scheme takes into account both positive and negative visual incentive perception effects and incorporates this Just-Noticeable-Distortion model into the VVC codec.

3.3.3 *Frequency-Domain Just-Noticeable-Distortion Model*

The Just-Noticeable-Distortion model in the frequency domain employs a transformation to shift the image into the frequency domain, subsequently integrating the Contrast Sensitivity Function (CSF), luminance adaptation, and contrast masking to characterize the JND threshold for each subband.

Jia et al. [231] initially introduced a spatio-temporal model for estimating Just-Noticeable-Distortion based on the Discrete Cosine Transform (DCT) domain, which considers both the temporal and spatial characteristics of the human visual system. The model combines the contrast sensitivity function, eye movement, luminance adaptation, and contrast masking to align with human visual perception. Additionally, they proposed spatio-temporal Just-Noticeable-Distortion configurations for grayscale images and videos in the DCT domain to determine the just noticeable thresholds within the DCT subband. Following this, Wei et al. [232] introduced a comprehensive spatio-temporal Just-Noticeable-Distortion model for monochrome images and videos in the DCT domain, incorporating both spatial and temporal sensitivity functions. This model integrates a spatial contrast sensitivity function,

luminance adaptation, contrast masking, and a temporal modulation factor to effectively capture perceptual redundancy. Parameterized through physical experiments, the model demonstrates alignment with the characteristics of the human visual system. In contrast, Bae et al. [233] proposed a structural contrast metric to evaluate texture complexity, taking into account both the intensity of contrast and the structural composition of the image. Leveraging this metric, they developed a Just-Noticeable-Distortion model for monochrome images utilizing DCT analysis. This model incorporates contrast masking, introduces a novel masking function for variable block size transforms, and integrates the contrast sensitivity function, luminance adaptation, and contrast masking to formulate a comprehensive Just-Noticeable-Distortion framework based on DCT.

Recently, foveated masking has also been incorporated into Just-Noticeable-Distortion models. In order to solve the limitations of the existing DCT-based perceptual distortion, Bae et al. [234] designed a Temporal-Foveated Masking (TFM) perceptual distortion model by incorporating the foveated masking effect for the first time. It is able to estimate Just-Noticeable-Distortion thresholds for motion/still image regions and provides a Just-Noticeable-Distortion adjustment method to prevent it from being overestimated. In order to overcome the limitations of the previous Just-Noticeable-Distortion model in contrast masking, Wan et al. [235] used DCT coefficients to analyze the directional characteristics of the image, and classified the DCT blocks according to the direction regularity and frequency energy to build an improved contrast masking model and its corresponding Just-Noticeable-Distortion model. The proposed model classifies the DCT blocks into five types according to the direction regularity and the frequency texture energy. By combining these block types with the perceptual sensitivity, a more accurate contrast masking model is proposed, and a novel DCT-based Just-Noticeable-Distortion model is established in the DCT domain by combining luminance adaptation and contrast sensitivity function. Ki et al. [236] proposed an energy-efficient Just-Noticeable-Distortion model based on DCT transform for video coding and further extended to two learning-based models for preprocessing in video coding. The model considers quantization effect, adjusts distortion level automatically according to quantization step, and achieves high coding efficiency by using perceptual video coding guided by Just-Noticeable-Distortion to eliminate perceptual redundancy.

3.3.4 Image and Video Coding with Just-Noticeable-Distortion

Recent advances in machine vision have introduced perceptually-driven distortion metrics, such as Just Recognizable Distortion and Just-Noticeable-Distortion, to optimize coding efficiency and adversarial robustness while preserving task performance. These frameworks leverage deep learning, ensemble methods, and perceptual mod-

eling to quantify distortion thresholds, guide compression strategies, and balance visual quality with computational efficiency across diverse vision tasks.

Zhang et al. [237] gave the concept of Just Recognizable Distortion (JRD) in the field of machine vision. To optimize coding efficiency, they developed an ensemble learning-based framework to predict the JRD for various vision tasks under different conditions, ensuring that data compression does not significantly compromise the performance of machine vision models. Jin et al. [238, 239] investigated the concept of just perceptible distortion in Deep Machine Vision (DMV) and proposed a model named DMV-JND. The authors employed unsupervised learning to generate Just-Noticeable-Distortion, enabling them to determine the threshold at which DMV can tolerate such distortions. Furthermore, they developed a semantic-guided redundancy evaluation strategy to regulate both the magnitude and spatial distribution of just perceptible distortion. Chen et al. [240] proposed a framework that utilizes neural networks to extract the high-level instance segmentation maps and the low-level signal features. This framework employs a combination of lossless and lossy codecs for encoding semantic information and signal features, while also incorporating an image predictor to achieve high-quality image reconstruction. Furthermore, the residual map is compressed to enhance the quality of image reconstruction. Tian et al. [215] developed an efficient perceptual image compression framework comprising an adaptive model for predicting Just-Noticeable-Distortion at the image level and a perceptual coding scheme. The framework employs a convolutional neural network to predict the label of Just-Noticeable-Distortion in images, utilizes support vector regression to determine the number of distinct levels of Just-Noticeable-Distortion, and incorporates a generation algorithm for calculating the quality factor associated with each level. Akan et al. [241] investigated the utilization of the concept of Just-Noticeable-Distortion for generating adversarial images with minimal perturbations. The proposed algorithm employs additive noise to distort the image until misclassifications occurs, while simultaneously minimizing the introduced perturbation. Additionally, they introduced a novel cost function aimed at preserving the perceptual similarity between adversarial and input images. Zhang et al. [213] proposed a method for optimizing JPEG compression by utilizing a quantization table based on Just-Noticeable-Distortion, with the aim of simulating perceptual distortion in the DCT domain and achieving optimal rate-distortion cost across all frequency bands to obtain an improved perceptual quality.

3.4 Visual Attention Modeling of Image and Video

Within the realm of visual saliency research, the examination of both salient object detection and fixation prediction falls under the umbrella of saliency detection, albeit with distinct focuses and applications. While united by their common objective to elucidate areas of visual interest, these two approaches diverge in their methodologies and the nature of information they provide. This research problem

belong to both visual perception modeling [84–93] and machine perception modeling [128–155, 157–162].

Salient object detection [242] is dedicated to the meticulous segmentation and isolation of visually prominent objects from intricate backgrounds. This process underscores the importance of precise boundary definition and comprehensive object recovery, furnishing unambiguous target information for further high-level visual processing tasks. Leveraging advanced algorithms and technologies, including deep learning architectures and graphical segmentation techniques, salient object detection achieves pixel-level precision, thereby forming a robust foundation for advancements in computer vision applications [9, 34]. In contrast, fixation prediction [243] is centered on anticipating regions that naturally captivate human or machine visual attention, approximating the focal points of visual exploration. Unlike salient object detection, the resultant fixation maps are not uniform. Instead, they emphasize the identification of key areas attention, reflecting the uneven and selective nature of human attention allocation. This approach, grounded in psychological principles and often augmented by machine learning methodologies, is instrumental in comprehending visual perception mechanisms, enhancing human-computer interface design, and informing the development of attention models in intelligent systems.

In summary, while salient object detection and fixation prediction are integral components of saliency detection, they diverge in their emphasis: the former on exact localization and segmentation of objects, and the latter on predicting regions guiding attention. Collectively, they advance the understanding of visual attention, enriching both theoretical frameworks and practical applications within the field. Next, the representative methods of Fixation Prediction and Salient Object Detection are introduced in details.

3.4.1 *Fixation Prediction Methods*

Numerous deep learning-based models for visual fixation prediction have emerged in recent years. Among these, the “Ensembles of Deep Networks” [244] stands as one of the pioneering models to employ shallow Convolutional Neural Networks (CNNs) for detecting visual saliency in natural images. This model extracts saliency features using CNNs and combines them with a linear classifier to produce saliency maps. Since then, advancements in deep neural networks and the development of large-scale saliency datasets have propelled further significant achievements in saliency prediction through deep learning methodologies.

DeepGaze II [245], leveraging VGGNet [246], underscores the efficacy of harnessing pre-trained networks as feature extractors for training deeper networks dedicated to saliency prediction. Huang et al. [247] revealed VGGNet’s superior performance in saliency detection, thereby stimulating the proposition of multiple VGGNet-based visual saliency prediction models, as seen in studies [248, 249]. Recently, the research focus has shifted towards exploiting multi-scale information. CNN-based models [247, 250] is introduced to extract multi-scale features, inte-

grating these to capture salient semantic objects across different granularity levels, thereby refining saliency prediction. The Deep Visual Attention [248], for instance, employs three decoders operating at distinct scales to generate multi-resolution saliency estimates for enhanced prediction. EML-NET [251] utilizes multi-scale feature maps from an encoder network to encapsulate global scene characteristics for saliency inference. Meanwhile, MSI-Net [249] adopts convolutional layers with varied dilation rates to augment multi-scale information utilization in saliency prediction. GazeGAN [252], a generative adversarial network tailored for saliency prediction, incorporates skip connections within a modified U-Net [253] architecture to handle multi-scale information. Similarly, UNISAL [254] benefits from skip connections to furnish its decoder with multi-scale features. These investigations collectively affirm the advantageous role of multi-scale information in enhancing visual saliency prediction capabilities.

3.4.2 Datasets for Salient Object Detection

Several prominent datasets employed in the detection of salient objects in images and videos are introduced as follows. In the realm of image saliency detection, historical precedents were marked by a scarcity of datasets, exemplified by MSRA-A and MSRA-B [281]. These datasets comprised images with simplistic compositions and featured clear contrasts between backgrounds and foregrounds. The maturation of deep learning methodologies has since catalyzed the proposal of numerous additional datasets, many tailored for model training purposes. Contemporary datasets introduce increased complexity in backgrounds, thereby augmenting the challenge of detection and often incorporating multiple objects of interest. Table 3.6 enumerates existing image datasets, highlighting various attributes including quantity, resolution, and the year of introduction, among others.

Turning to salient object detection in videos, the discussion encompasses several esteemed datasets, summarized in Table 3.7. The DAVIS [286] dataset stands as a widely utilized and inherently challenging resource, embodying 50 video sequences complete with pixel-level annotations for each frame. More recently, the DAVSOD [287] dataset has been introduced, which stratifies its 226 videos (90 for training, 46 for validation, and 90 for testing) in a manner that underscores assess-

Table 3.6 Representative RGB salient object detection datasets

Dataset	Num.	Res.	Obj.	Backg.
PASCAL-S [282]	850	500 × 500	Multiple, moderate	Complex
HKU-IS [283]	4447	400 × 400	Multiple, moderate	Complex
DUTS [284]	10553(5019)	400 × 400	Single, small	Complex
SOC [285]	6000	600 × 600	Single, small	Complex

Source Author

Table 3.7 Representative video saliency detection datasets

DS.	Fra. Num	Vid. Num.	Res.	Obj.	Backg.
SegTrackV2 [288]	1065	14	640×360	Single	Diverse
FBMS [275]	13,860	59	960×540	Single	Diverse
ViSal [274]	963	17	512×228	Single	Diverse
DAVIS [286]	3455	50	1920×1080	Multiple	Complex
VOS [289]	116,103	200	800×800	Single	Complex
DAVSOD [287]	23,938	226	640×360	Multiple	Complex

Source Author

ment rigor. Notably, the test subset is further divided into subsets of 35 simple, 30 moderate, and 25 challenging videos, facilitating the evaluation of models across a spectrum of difficulties. This nuanced categorization not only serves as a barometer for the diverse capabilities of detection algorithms but also encourages advancements that can handle a broad range of real-world complexities.

3.4.3 Image Salient Object Detection

In the task of salient object detection, supervision labels play a key role. Different from previous methods that use region and boundary labels [255, 256], Wu et al. [257] proposed a new Label Decoupling Framework (LDF), in which new supervision labels are introduced.

Label Decoupling. The prediction complexity associated with a pixel is inherently linked to its positional context within an image. Edges are fraught with a chaotic background, rendering bordering pixels susceptible to erroneous predictions, whereas centrally located pixels benefit from the inherent coherence of the salient object, thereby exhibiting heightened prediction accuracy. To address this disparity, a more nuanced approach involves stratified pixel treatment based on their distinct attributes. This work introduces a framework that bifurcates the initial label into two components: the body label and the detail label. The cornerstone of this methodology lies in adopting Distance Transformation (DT), a classical image processing technique, for label disentanglement.

In technical terms, the Distance Transformation process commences with a binary input image I , comprising two disjoint sets: the foreground I_{fg} and the background I_{bg} . Pixel p within I assumes a value of 1 if it falls within I_{fg} , otherwise 0 for membership in I_{bg} . DT employs a Euclidean distance metric $f(p, q)$ to gauge proximity between pixels p and q . Specifically, for every foreground pixel p , DT identifies the closest background pixel q and assigns p the computed distance $f(p, q)$; background pixels are assigned a distance of zero. This transformation yields a new image I' , mathematically encapsulated as [257]:

$$I'(p) = \begin{cases} \min_{q \in I_{bg}} f(p, q), & p \in I_{fg}, \\ 0, & p \in I_{bg}. \end{cases} \quad (3.1)$$

After transformation, pixel intensities in I' undergo normalization to span the interval $[0, 1]$, achieved through a linear mapping $I' = \frac{I' - \min(I')}{\max(I') - \min(I')}$. Unlike I , where pixels are classified, I' imparts additional spatial information, with central object pixels assuming maximal values and periphery or background pixels minimal ones. Thus, I' embodies the body component, emphasizing easier-to-predict central pixels, and serving as the body label (BL). Conversely, subtracting I' from I isolates the detail image, focusing on peripheral pixels, designated as the detail label (DL). To suppress background influence, these labels are further refined via multiplication with I [257]:

$$Label \Rightarrow \begin{cases} BL = I * I', \\ DL = I * (1 - I'). \end{cases} \quad (3.2)$$

Through this mechanism, the initial label is effectively decomposed into dual supervisory signals, enabling the learning of body and detail features with their unique characteristics.

Feature Interaction Network. The proposed detection framework is systematically structured around a quartet of core components: a backbone encoder network, an interaction encoder network, accompanied by a dedicated body decoder network, and a detail decoder network. The well-established ResNet-50 architecture [258] serves as the backbone, tasked with extracting a hierarchy of multi-level features represented by sets $B = \{B_i | i = 2, 3, 4, 5\}$ and $D = \{D_i | i = 2, 3, 4, 5\}$. These extracted features are then channeled through distinct pathways. The body decoder network processes features B to produce body maps, capturing the central essence of salient objects. In parallel, the detail decoder network handles features D , generating detail maps that emphasize peripheral and fine-grained structures.

A naive approach to merge these parallel streams might involve direct concatenation followed by a convolutional operation to yield the final saliency maps. However, such an approach overlooks the potential synergy between the two branches. To bridge this gap and facilitate a more dynamic exchange of information, the framework introduces an interaction encoder network. The interaction encoder accepts concatenated outputs from both the body and detail decoders as its input. Through a series of stacked convolutional layers, it distills these inputs into multi-level features, which are then tailored via 3×3 convolutional layers to align with the specific requirements of the body and detail decoders. Subsequently, these interaction features are directly added to corresponding features from the backbone encoder, enhancing the precision of the resultant saliency maps.

This design fosters an iterative feature interaction process. Initially, in the first iteration, the two branches operate independently, without cross-communication.

Commencing from the second iteration onwards, a reciprocal exchange of information takes place between the branches, fostering a cycle of refined feature learning and ultimately leading to saliency maps that are more comprehensive and nuanced in their representation of salient objects amidst complex scenes.

Loss Function. Loss function is streamlined to aggregate outcomes from successive iterations, succinctly formulated as [257]:

$$\mathcal{L} = \sum_{k=1}^K \alpha_k \ell^{(k)}, \quad (3.3)$$

where $\ell^{(k)}$ represents the loss at the k -th iteration, with K denoting the total iterations and α_k set uniformly to 1 for simplicity, treating each iteration equally. Each iteration yields three outputs: body, detail, and segmentation maps, each associated with a specific loss. Thus, $\ell^{(k)}$ consolidates these into [257]:

$$\ell^{(k)} = \ell_{body}^{(k)} + \ell_{detail}^{(k)} + \ell_{segm}^{(k)}, \quad (3.4)$$

where $\ell_{body}^{(k)}$, $\ell_{detail}^{(k)}$ and $\ell_{segm}^{(k)}$ representing the respective losses for body, detail, and segmentation. Both body and detail losses adopt the Binary Cross Entropy (BCE), a prevalent choice for binary classification and segmentation tasks [257]:

$$\ell_{bce} = - \sum_{(x,y)} [g(x,y) \log(p(x,y)) + (1-g(x,y)) \log(1-p(x,y))], \quad (3.5)$$

where $g(x, y)$ and $p(x, y)$ denote the ground truth label and the predicted saliency probability for pixel (x, y) , respectively.

Considering BCE's limitation in capturing global structure, we follow the suggestion of [259] and employ Intersection over Union (IoU) loss for $\ell_{segm}^{(k)}$ to assess overall image similarity [257]:

$$\ell_{iou} = 1 - \frac{\sum_{(x,y)} [g(x, y) * p(x, y)]}{\sum_{(x,y)} [g(x, y) + p(x, y) - g(x, y) * p(x, y)]}. \quad (3.6)$$

Notably, IoU loss is not applied to body or detail losses due to their non-binary nature, which would conflict with IoU's requirement for binary ground truths. Furthermore, reference to Table 3.8 is recommended for a comprehensive view of alternative salient object detection methodologies in imagery. This table enumerates not only a selection of conventional techniques but also highlights prominent deep learning-based approaches that have emerged in recent years. Comparisons drawn between the primary technologies underscore the prevailing trends in the field, thereby facilitating a nuanced understanding of the underlying detection mechanisms and orienting insights towards the trajectory of technological progress. Consequently, this anal-

Table 3.8 A brief summary of SOD model

Model	Year	Publication	Contributions
Traditional			
GMR [260]	2013	CVPR	Based on background-prior
UFO [261]	2013	ICCV	Based on global-contrast, focusness-prior and objectness prior
RBD [262]	2014	CVPR	Based on background connectivity
PDE [263]	2014	CVPR	Based on center-prior, background-prior and color prior
LPS [264]	2015	TIP	Based on global-contrast and spatial distribution
Deep learning			
BASNet [259]	2019	CVPR	Based on hybrid loss (BCE+IoU+SSIM)
CPD [265]	2019	CVPR	Based on cascaded partial decoder
MIN [266]	2020	CVPR	Based on integration of the features from adjacent levels
VST [267]	2021	ICCV	Based on Transformer as an encoder
EDN [268]	2022	TIP	Based on extreme downsampling technique

Source Author

ysis paves the way for the conception of innovative strategies aimed at enhancing detection performance, thereby contributing to the advancement of the discipline.

3.4.4 Video Salient Object Detection

The recognition and isolation of salient objects in video sequences, known as Video Salient Object Detection (VSOD), necessitates a keen understanding of spatiotemporal cues, particularly given humans’ innate attraction to moving objects. Conventional methods in this field have relied on techniques such as the Long Short-Term Memory (LSTM) networks and the 3D Convolutional Neural Networks (C3D), which progressively handle temporal data via sequential propagation. While effective to an extent, these methods are limited in their ability to instantaneously and directly decipher complex motion dynamics.

The advent of non-local operations marked a significant stride towards capturing long-range dependencies, yet integrating this mechanism seamlessly into VSOD frameworks poses dual challenges. It inadvertently overlooks motion-specific cues, leaning more towards learning a holistic, motion-agnostic context, and its resource-intensive nature—both computationally and in terms of memory—renders it impractical for high-resolution, temporally dense prediction tasks like VSOD.

To navigate these hurdles, Gu et al. [269] introduced a novel concept: Constrained Self-Attention (CSA). Rooted in the understanding that objects typically traverse in

consistent trajectories, CSA is meticulously designed to pinpoint and analyze motion patterns effectively. By further arranging CSA modules into hierarchical, Pyramid Constrained Self-Attention (PCSA) structures, the authors ensure the capture of object movements across diverse scales and velocities.

Constrained Self-Attention Mechanism. To effectively harness motion information within a video sequence composed of T frames, the authors initially extract static salient features of dimensions $T \times H \times W$ using a pre-trained backbone network. These features are then projected into three distinct spaces—query (Q), key (K), and value (V)—utilizing linear transformations. A critical insight leveraged by CSA is the temporal consistency in object positions across adjacent frames. Exploiting this, when querying a feature element $x_q = (t, h, w)$ from the query space Q , this work focuses on the local neighborhood of x_q in the key space K to gauge similarity. This is realized by defining a responsive window with a radius r and dilation factor d , yielding the surrounding region $S(x_q, K)$ as follows [269]:

$$S(x_q, K) = \{K_{(t', h', w')}\}_{t'=1, h'=h-dr, w'=w-dr}^{T, h+dr, w+dr}, \quad (3.7)$$

where $S(x_q, K)$ is shaped in $\mathbb{R}^{TR^2 \times C}$, reflecting a compact representation of the contextual neighborhood around the queried point across time and space.

The affinity computation is thus revised to [269]:

$$W_{\text{att}} = f(Q, K) = QS(Q, K)^T. \quad (3.8)$$

W_{att} is normalized by softmax function, and static salient features are dynamically enhanced by aggregating information across frames guided by these attention weights, encapsulated in [269]:

$$Y = W_{\text{att}}S(Q, V). \quad (3.9)$$

By confining attention to a constrained response region, CSA diverges from non-local operations that capture global contexts, focusing instead on motion-driven affinities. Unlike ConvLSTM’s sequential propagation, CSA instantaneously accesses multi-frame information. Integrated seamlessly into a backbone via residual connection ($Z = X + Y$), CSA augments feature representations with temporal dynamics.

Pyramid Combination. While the previously discussed Constrained Self-Attention (CSA) mechanism effectively harnesses local motion patterns by leveraging a motion-centric prior, the complexity of dynamic scenes demands a more nuanced approach. Objects of varying sizes often move at different velocities, necessitating adaptable mechanisms that go beyond a singular, fixed window size inherent to standard CSA. Drawing inspiration from the multi-head attention strategy employed in machine translation, the authors proposed an enhancement where features are projected into distinct subspaces via individually learned linear functions. Departing from the conventional multi-head scheme, this design integrates multi-

scale learning, thereby enhancing adaptability to a broader range of motion characteristics.

The procedure commences with the division of the input feature tensor X into g distinct groups $\{X_i, i = 1, 2, \dots, g\}$. Subsequently, a suite of parallel CSA modules, each designated as $(CSA_i, i = 1, 2, \dots, g)$, is employed. These modules are configured with varying window dimensions ($r' = (r_1, r_2, \dots, r_g)$) and dilation rates ($d = (d_1, d_2, \dots, d_g)$) to accommodate motion cues extraction across the diversified X_i groups. The resultant features, collected as $\{Y_i, i = 1, 2, \dots, g\}$, are concatenated and fused via a $1 \times 1 \times 1$ convolutional layer to generate a spatiotemporal feature Y , which is ultimately combined with the input X . The conventional CSA with a uniform window size R may inadequately represent fast or large-scale moving targets due to its rigidity. In contrast, PCSA innovates by embracing a multi-head methodology, where each head is calibrated with unique window sizes and dilation parameters to cater to a broader spectrum of motion scenarios.

Network Design. The architectural foundation of the proposed network lies in MobileNetV3 [270]. To ensure the retention of spatial information within feature maps, adjustments are made to the encoder part. Specifically, the convolution strides for both the third and final layers are altered from 2 to 1. Enhancing the backbone's initial stages, a modified version of the RFB block [271] is incorporated, wherein traditional convolutions are substituted with separable convolutions for added efficiency.

In integrating the PCSA module, the configuration entails setting $g = 4$, with window sizes $r = \{3, 4\}$, and dilation factors $d = \{1, 2\}$ to tackle multi-scale motion dynamics effectively. As for the decoder component, the second-stage output from the encoder serves as the low-level feature source. The PCSA module's output, embodying spatiotemporal features, undergoes processing via a dilated convolution to diminish its dimensionality from 128 to 32, facilitating subsequent computations. To align the spatiotemporal feature map with the lower-resolution counterpart, bilinear interpolation is employed for upsampling. Following this, the upsampled spatiotemporal features are concatenated with the preserved low-level features. A subsequent convolutional layer, defined by a kernel size of 3, processes this concatenated information to yield the final prediction output.

Loss Function Configuration. For model training, the employment of a binary cross entropy loss function is central. Let P symbolize the predicted saliency map output, with G representing the corresponding Ground Truth map. The formulation for the binary cross entropy loss, denoted as L_{bce} [269], is given by:

$$L_{bce}(P, G) = -\frac{1}{N} \sum_{i=1}^N [g_i \log(p_i) + (1 - p_i) \log(1 - p_i)]. \quad (3.10)$$

where N signifies the total number of pixels under consideration, g_i denotes the ground truth value (either 0 or 1) for the i -th pixel, and p_i represents the predicted probability of the i -th pixel of the salient class, also confined within the range

[0, 1]. This loss quantifies the dissimilarity between the predicted and true distributions, guiding the model optimization toward accurate salient object detection.

In the context of exploring salient target detection techniques in videos, Table 3.9 presents an exhaustive comparative analysis of a variety of existing methods, highlighting the differences and advantages and disadvantages between the traditional methods and the deep learning-based methods. Generally speaking, deep learning-based methods, with their powerful automatic feature extraction capabilities, tend to demonstrate superior performance over traditional methods in terms of detection accuracy and efficiency. This observation not only validates the effectiveness of deep learning techniques in complex visual tasks, but also points out the direction for the development of the field.

By scrutinizing the technical characteristics and differences of these methods, researchers and developers are able to gain a deeper understanding of the unique advantages of each method and its potential limitations. This in-depth understanding provides valuable insights for subsequent model design, and encourages researchers to incorporate the advantages and avoid the shortcomings of various techniques in the conceptualization of new solutions, thus advancing the salient target detection technology towards higher accuracy, stronger robustness, and wider applicability. In short, the comprehensive comparison and deep understanding of current methods lay a solid foundation for the innovation and optimization of model design strategies.

Table 3.9 A brief summary of video SOD model

Model	Year	Publication	Contributions
Traditional			
SSDSR [272]	2012	ICME	Based on sparse reconstruction
SBSSD [273]	2014	TCSVT	Based on Superpixel-level motion features
GFGF [274]	2015	TIP	Based on gradient flow field
SGVS [275]	2015	CVPR	Based on geodesic distance
STBP [276]	2017	TIP	Based on spatiotemporal background prior and SIFT flow
Deep learning			
DSVS [277]	2019	TIP	Based on 3D stereoscopic video saliency
TASED-Net [278]	2019	ICCV	Based on an encoder-Decoder-based approach
COSNet [279]	2022	WACV	Based on contrastive features and attention modules
CoSTFormer [280]	2023	TNNLS	Based on complementary spatial and temporal contexts

Source Author

3.5 Summary

This chapter proposes a comprehensive framework for quality assessment and perception modeling in images and videos, covering three key research directions: quality assessment methodologies, Just-Noticeable-Distortion modeling and Salient Object Detection. Quality assessment is systematically categorized into subjective and objective methods, with the latter further divided into Full-Reference, Reduced-Reference, and No-Reference approaches, highlighting the role of deep learning in improving accuracy. JND modeling is explored in both pixel and frequency domains, emphasizing its applications in optimizing coding efficiency and perceptual quality. For Salient Object Detection, advanced techniques such as the Label Decoupling Framework for images and Constrained Self-Attention mechanisms for videos are introduced, alongside benchmark datasets. Collectively, these contributions advance the alignment of computational models with human visual perception, offering practical solutions for multimedia applications.

Exercises

1. Please explain the aim of image and video quality assessment.
2. Please explain the capability of transformer-based model in IQA tasks compare to traditional convolutional neural networks.
3. Please explain how multimodal information can be integrated with Transformer models to enhance IQA performance.
4. Please describe the advantages and limitations of existing transformer-based VQA methods in handling the integration of spatial and temporal information.
5. Please describe the key to handling different types of distortions and noise in VQA.
6. Please describe the difference between fixation prediction and salient object detection.
7. Please describe what types of labels can be used to train models for salient object detection in images.
8. Please describe what should be the output of a detection model when there are no salient objects in the scene.
9. Please explain the advantages of using point clouds for salient object detection compared to image data.
10. Is it possible to incorporate optical flow information in video salient object detection, and what benefits can it bring?

References

1. W. Gao, G. Li, H. Yuan, R. Hamzaoui, Z. Li, S. Liu, Apccpa'22: 1st international workshop on advances in point cloud compression, processing and analysis, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 7392–7393
2. W. Gao, G. Li, Point cloud compression, enhancement and applications: from 3d perception to large models, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 292–11 293
3. H. Sun, G. Li, S. Chen, L. Li, W. Gao, Special topic on 3d point cloud processing and applications. *ZTE Commun.* **21**(4), 1 (2023)
4. W. Gao, W. Gao, X. Mu, C. Peng, G. Li, Overview and comparison of avs point cloud compression standard. *APSIPA Trans. Signal Inf. Process.* (2024)
5. W. Gao, S. Wang, X. Zhang, S. Kwong, Editorial for special issue on deep learning-based data compression. *APSIPA Trans. Signal Inf. Process.* **13**(6) (2024)
6. W. Gao, H. Ye, G. Li, H. Zheng, Y. Wu, L. Xie, Openpointcloud: An open-source algorithm library of deep learning based point cloud compression, in *Proceedings of the 30th ACM international conference on multimedia* (2022), pp. 7347–7350
7. Y. Zhang, W. Gao, G. Li, Openpointcloud-v2: a deep learning based open-source algorithm library of point cloud processing, in *Proceedings of the 1st International Workshop on Advances in Point Cloud Compression, Processing and Analysis* (2022), pp. 51–55
8. L. Xie, W. Gao, Learningpcc: a pytorch library for learning-based point cloud compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 234–11 238
9. L. Xie, W. Gao, Pchmvision: an open-source library of point cloud compression for human and machine vision, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 239–11 243
10. T. Qin, G. Li, W. Gao, S. Liu, Multi-grained point cloud geometry compression via dual-model prediction with extended octree. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(9), 1–30 (2024)
11. Y. Shao, W. Gao, S. Liu, G. Li, Advanced patch-based affine motion estimation for dynamic point cloud geometry compression. *Sensors* **24**(10), 3142 (2024)
12. Y. Shao, F. Song, W. Gao, S. Liu, G. Li, Texture-guided graph transform optimization for point cloud attribute compression. *Appl. Sci.* **14**(10), 4094 (2024)
13. Y. Shao, X. Yang, W. Gao, S. Liu, G. Li, 3d point cloud attribute compression using diffusion-based texture-aware intra prediction. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
14. J. Zhang, Y. Chen, G. Liu, W. Gao, G. Li, Efficient point cloud attribute compression framework using attribute-guided graph fourier transform, in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2024), pp. 8426–8430
15. W. Gao, H. Yuan, G. Li, Z. Li, H. Yuan, Low complexity coding unit decision for video-based point cloud compression. *IEEE Trans. Image Process.* **33**, 149–162 (2023)
16. Y. Shao, G. Li, Q. Zhang, W. Gao, S. Liu, Nonrigid registration-based progressive motion compensation for point cloud geometry compression. *IEEE Trans. Geosci. Remote Sens.* (2023)
17. F. Song, G. Li, X. Yang, W. Gao, S. Liu, Block-adaptive point cloud attribute coding with region-aware optimized transform. *IEEE Trans. Circuits Syst. Video Technol.* (2023)
18. Y. An, Y. Shao, G. Li, W. Gao, S. Liu, A fast motion estimation method with hamming distance for lidar point cloud compression, in *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2022), pp. 1–5
19. H. Yuan, W. Gao, G. Li, Z. Li, Rate-distortion-guided learning approach with cross-projection information for v-pcc fast cu decision, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 3085–3093
20. F. Song, G. Li, W. Gao, T.H. Li, Rate-distortion optimized graph for point cloud attribute coding. *IEEE Signal Process. Lett.* **29**, 922–926 (2022)

21. F. Song, G. Li, X. Yang, W. Gao, T.H. Li, Fine-grained correlation representation for graph-based point cloud attribute compression, in *2022 IEEE International Conference on Multi-media and Expo (ICME)* (IEEE, 2022), pp. 1–6
22. F. Shen, W. Gao, A rate control algorithm for video-based point cloud compression, in *2021 International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2021), pp. 1–5
23. F. Song, Y. Shao, W. Gao, H. Wang, T. Li, Layer-wise geometry aggregation framework for lossless lidar point cloud compression. *IEEE Trans. Circuits Syst. Video Technol.* **31**(12), 4603–4616 (2021)
24. Z. Cai, W. Gao, G. Li, W. Gao, Distortion propagation model-based v-pcc rate control for 3d point cloud broadcasting. *IEEE Trans. Broadcast.* 1–13 (2024)
25. X. Mu, W. Gao, H. Yuan, S. Wang, G. Li, Adaptive lpu decision for dynamic point cloud compression. *IEEE Signal Process. Lett.* (2024)
26. C. Peng, W. Gao, Laplacian matrix learning for point cloud attribute compression with ternary search-based adaptive block partition, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 10 412–10 420
27. Y. Wang, W. Gao, X. Mu, H. Yuan, Rate control optimization for joint geometry and attribute coding of lidar point clouds, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
28. T. Guo, H. Yuan, T. Wang, W. Gao, Graph filter-based fast motion matching for inter frame coding of mpeg g-pcc, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 1151–1155
29. L. Xie, W. Gao, H. Zheng, G. Li, Spcgc: Scalable point cloud geometry compression for machine vision, in *2024 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2024), pp. 17 272–17 278
30. L. Xie, W. Gao, H. Zheng, H. Ye, Semantic-aware visual decomposition for point cloud geometry compression, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 595–595
31. Z. Qi, W. Gao, Variable-rate point cloud geometry compression based on feature adjustment and interpolation, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 63–72
32. Z. Yu, W. Gao, When dynamic neural network meets point cloud compression: computation-aware variable rate and checkerboard context, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 600–600
33. L. Xie, W. Gao, S. Fan, Z. Yao, Pdnet: parallel dual-branch network for point cloud geometry compression and analysis, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 596–596
34. L. Xie, W. Gao, H. Zheng, End-to-end point cloud geometry compression and analysis with sparse tensor, in *International Workshop on Advances in Point Cloud Compression, Processing and Analysis* (2022), pp. 27–32
35. C. Fu, G. Li, R. Song, W. Gao, S. Liu, Octattention: octree-based large-scale contexts model for point cloud compression. *Proc. AAAI Conf. Artif. Intell.* **36**(1), 625–633 (2022)
36. C. Zhang, W. Gao, Adapcc: adaptive rate control and rate-distortion-complexity optimization for dynamic point cloud compression, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
37. K. Wang, W. Gao, Unipcgc: towards practical point cloud geometry compression via an efficient unified approach, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
38. K. Zheng, W. Gao, H. Zheng, Octree-based learned point cloud geometry compression: a lossy perspective, in *2025 Data Compression Conference (DCC)* (IEEE, 2025)
39. Z. Qi, H. Liu, W. Gao, Variable-rate point cloud compression with channel-wise entropy modeling, in *2025 Data Compression Conference (DCC)* (IEEE, 2025)
40. H. Zheng, W. Gao, Z. Yu, T. Zhao, G. Li, Viewpcgc: view-guided learned point cloud geometry compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 7152–7161

41. L. Xie, W. Gao, H. Zheng, G. Li, Roi-guided point cloud geometry compression towards human and machine vision, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 3741–3750
42. Y. Wu, L. Xie, S. Sun, W. Gao, Y. Yan, Adaptive intra period size for deep learning-based screen content video coding, in *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2024)
43. H. Zheng, W. Gao, End-to-end rgb-d image compression via exploiting channel-modality redundancy, *Proc. AAAI Conf. Artif. Intell.* **38**(7), 7562–7570 (2024)
44. L. Tao, W. Gao, G. Li, C. Zhang, Adanic: towards practical neural image compression via dynamic transform routing, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 16 879–16 888
45. Y. Wu, W. Gao, End-to-end lossless compression of high precision depth maps guided by pseudo-residual (2022). [arXiv:2201.03195](https://arxiv.org/abs/2201.03195)
46. Y. Wu, Z. Qi, H. Zheng, L. Tao, W. Gao, Deep image compression with latent optimization and piece-wise quantization approximation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 1926–1930
47. W. Gao, L. Tao, L. Zhou, D. Yang, X. Zhang, Z. Guo, Low-rate image compression with super-resolution learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2020), pp. 154–155
48. W. Gao, S. Sun, H. Zheng, Y. Wu, H. Ye, Y. Zhang, Opendmc: an open-source library and performance evaluation for deep-learning-based multi-frame compression, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9685–9688
49. W. Gao, H. Zheng, C. Zhang, K. Zheng, Z. Yu, Y. Li, H. Ye, Y. Zhang, Opendic: an open-source library and performance evaluation for deep-learning-based image compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 202–11 205
50. Y. Li, W. Gao, G. Li, S. Ma, Saliency segmentation oriented deep image compression with novel bit allocation. *IEEE Trans. Image Process.* **34**, 16–29 (2025)
51. C. Zhang, W. Gao, Learned rate control for frame-level adaptive neural video compression via dynamic neural network, in *European Conference on Computer Vision* (Springer, 2024)
52. Z. Yang, W. Gao, G. Li, Y. Yan, Sur-driven video coding rate control for jointly optimizing perceptual quality and buffer control. *IEEE Trans. Image Process.* (2023)
53. F. Shen, Z. Cai, W. Gao, An efficient rate control algorithm for intra frame coding in avs3, in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2021), pp. 3164–3169
54. H. Yuan, W. Gao, J. Wang, Dynamic computational resource allocation for fast inter frame coding in video conferencing applications, in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6
55. W. Gao, Q. Jiang, R. Wang, S. Ma, G. Li, S. Kwong, Consistent quality oriented rate control in hevc via balancing intra and inter frame coding. *IEEE Trans. Ind. Inform.* **18**(3), 1594–1604 (2021)
56. H. Yuan, W. Gao, A new coding unit partitioning mode for screen content video coding, in *Proceedings of the 2021 5th International Conference on Digital Signal Processing* (2021), pp. 66–72
57. W. Gao, On the performance evaluation of state-of-the-art rate control algorithms for practical video coding and transmission systems, in *Proceedings of the 2020 4th International Conference on Video and Image Processing* (2020), pp. 179–185
58. W. Gao, S. Kwong, Q. Jiang, C.-K. Fong, P.H. Wong, W.Y. Yuen, Data-driven rate control for rate-distortion optimization in hevc based on simplified effective initial qp learning. *IEEE Trans. Broadcast.* **65**(1), 94–108 (2018)
59. W. Gao, A multi-objective optimization perspective for joint consideration of video coding quality, in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (IEEE, 2019), pp. 986–991

60. W. Gao, S. Kwong, Y. Jia, Joint machine learning and game theory for rate control in high efficiency video coding. *IEEE Trans. Image Process.* **26**(12), 6074–6089 (2017)
61. W. Gao, S. Kwong, Y. Zhou, H. Yuan, Ssim-based game theory approach for rate-distortion optimized intra frame ctu-level bit allocation. *IEEE Trans. Multimedia* **18**(6), 988–999 (2016)
62. W. Gao, S. Kwong, H. Yuan, X. Wang, Dct coefficient distribution modeling and quality dependency analysis based frame-level bit allocation for hevcc. *IEEE Trans. Circuits Syst. Video Technol.* **26**(1), 139–153 (2015)
63. W. Gao, S. Kwong, Phase congruency based edge saliency detection and rate control for perceptual image and video coding, in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2016), pp. 000 264–000 269
64. H. Yuan, W. Gao, Openfastvc: An open source library for video coding fast algorithm implementation, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9660–9663
65. H. Yuan, W. Gao, S. Ma, Y. Yan, Divide-and-conquer-based rdo-free cu partitioning for 8k video compression. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(4), 1–20 (2024)
66. L. Tao, W. Gao, A hardware implementation of entropy encoder for 8k video coding, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
67. Y. Guo, W. Gao, S. Ma, G. Li, Accelerating transform algorithm implementation for efficient intra coding of 8k uhd videos. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* **18**(4), 1–20 (2022)
68. Z. Cai, W. Gao, Efficient fast algorithm and parallel hardware architecture for intra prediction of avs3, in *IEEE International Symposium on Circuits and Systems (ISCAS)* (IEEE, 2021), pp. 1–5
69. W. Gao, H. Yuan, Y. Guo, L. Tao, Z. Cai, G. Li, Openhardwarevc: an open source library for 8k uhd video coding hardware implementation, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 7339–7342
70. W. Gao, H. Yuan, G. Liao, Z. Guo, J. Chen, Pp8k: a new dataset for 8k uhd video compression and processing. *IEEE MultiMedia* (2023)
71. W. Liu, W. Gao, G. Li, S. Ma, T. Zhao, H. Yuan, Enlarged motion-aware and frequency-aware network for compressed video artifact reduction. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
72. G. Li, G. Wei, W. Gao, *Point Cloud Compression: Technologies and Standardization* (Springer Nature, 2024)
73. G. Li, W. Gao, W. Gao, Introduction, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 1–28
74. G. Li, W. Gao, W. Gao, Background knowledge, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 29–51
75. G. Li, W. Gao, W. Gao, Predictive coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 53–70
76. G. Li, W. Gao, W. Gao, Transform coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 71–96
77. G. Li, W. Gao, W. Gao, Quantization techniques, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 97–112
78. G. Li, W. Gao, W. Gao, Entropy coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 113–133
79. G. Li, W. Gao, W. Gao, Mpeg geometry-based point cloud compression (g-pcc) standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 135–165
80. G. Li, W. Gao, W. Gao, Avs point cloud compression standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 167–197
81. G. Li, W. Gao, W. Gao, Mpeg video-based point cloud compression (v-pcc) standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 199–218
82. G. Li, W. Gao, W. Gao, Mpeg ai-based 3d graphics coding standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 219–241

83. G. Li, W. Gao, W. Gao, Future work, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 243–250
84. L. Xie, X. Mu, G. Li, W. Gao et al., Pku-dpcc: a new dataset for dynamic point cloud compression. *APSIPA Trans. Signal Inf. Process.* **13**(6) (2024)
85. H. Yuan, W. Gao, W. Gao, Opensep: an open source subjective experiment platform, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 248–11 251
86. S. Fan, W. Gao, Screen-based 3d subjective experiment software, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9672–9675
87. J. Wang, W. Gao, G. Li, Zoom to perceive better: No-reference point cloud quality assessment via exploring effective multiscale feature. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
88. J. Wang, W. Gao, G. Li, Applying collaborative adversarial learning to blind point cloud quality measurement. *IEEE Trans. Instrum. Meas.* (2023)
89. H. Li, W. Gao, Information exploration of projected views for point cloud quality measurement. *IEEE Trans. Instrum. Meas.* (2024)
90. B. Qu, X. Liang, S. Sun, W. Gao, Exploring aigc video quality: A focus on visual harmony, video-text consistency and domain distribution gap, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2024)
91. B. Qu, H. Li, W. Gao, Bringing textual prompt to ai-generated image quality assessment, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024)
92. S. Sun, X. Liang, S. Fan, W. Gao, W. Gao, Ve-bench: subjective-aligned benchmark suite for text-driven video editing quality assessment, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
93. Z. Guo, W. Gao, H. Wang, J. Wang, S. Fan, No-reference deep quality assessment of compressed light field image, in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6
94. W. Liu, W. Gao, X. Mu, Fast inter-frame motion prediction for compressed dynamic point cloud attribute enhancement. *Proc. AAAI Conf. Artif. Intell.* **38**(4), 3720–3728 (2024)
95. Z. Yang, W. Gao, X. Lu, Danet: density-adaptive network for geometry-based point cloud compression artifacts removal, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
96. X. Fan, G. Li, D. Li, Y. Ren, W. Gao, T.H. Li, Deep geometry post-processing for decompressed point clouds, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
97. X. Zhang, G. Liao, W. Gao, G. Li, Tdrnet: transformer-based dual-branch restoration network for geometry based point cloud compression artifacts, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
98. Z. Li, G. Li, T. H. Li, S. Liu, W. Gao, Semantic point cloud upsampling. *IEEE Trans. Multimedia* (2022)
99. Z. Li, S. Liu, W. Gao, G. Li, G. Li, S4r: rethinking point cloud sampling via guiding upsampling-aware perception. *IEEE Trans. Multimedia* (2025)
100. R. Zhang, W. Gao, G. Li, T.H. Li, Qinet: decision surface learning and adversarial enhancement for quasi-immune completion of diverse corrupted point clouds. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–14 (2022)
101. R. Bao, Y. Ren, G. Li, W. Gao, S. Liu, Flow-based point cloud completion network with adversarial refinement, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2559–2563
102. J. Chen, G. Li, R. Zhang, T. H. Li, W. Gao, Pointvae: invertible variational autoencoder framework for 3d point cloud generation, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 3216–3220
103. R. Zhang, J. Chen, W. Gao, G. Li, T.H. Li, Pointot: interpretable geometry-inspired point cloud generative model via optimal transport. *IEEE Trans. Circuits Syst. Video Technol.* **32**(10), 6792–6806 (2022)

104. Z. Li, G. Li, T. Li, S. Liu, W. Gao, Information-growth attention network for image super-resolution, in *Proceedings of the 29th ACM International Conference on Multimedia* (2021), pp. 544–552
105. L. Zhou, W. Gao, G. Li, H. Yuan, T. Zhao, G. Yue, Disentangled feature distillation for light field super-resolution with degradations, in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2023), pp. 116–121
106. L. Zhou, W. Gao, G. Li, End-to-end spatial-angular light field super-resolution using parallax structure preservation strategy, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 3396–3400
107. W. Gao, L. Zhou, L. Tao, A fast view synthesis implementation method for light field applications. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* 17(4), 1–20 (2021)
108. X. Zhang, W. Gao, G. Li, Q. Jiang, R. Cong, Image quality assessment-driven reinforcement learning for mixed distorted image restoration. *ACM Trans. Multimedia Computing, Commun. Appl.* 19(1s), 1–23 (2023)
109. X. Zhang, W. Gao, H. Yuan, G. Li, Je 2 net: joint exploitation and exploration in reinforcement learning based image restoration, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2090–2094
110. X. Zhang, W. Gao, Hirl: hybrid image restoration based on hierarchical deep reinforcement learning via two-step analysis, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2445–2449
111. S. Wang, Y. Lu, W. Xia, P. Xia, Z. Wang, W. Gao, Light field angular super-resolution by view-specific queries. *Visual Comput.* pp. 1–16 (2024)
112. Y. Sun, Z. Li, S. Wang, W. Gao, Depth-assisted calibration on learning-based factorization for a compressive light field display. *Opt. Express* 31(4), 5399–5413 (2023)
113. Y. Sun, Z. Li, L. Li, S. Wang, W. Gao, Optimization of compressive light field display in dual-guided learning, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2075–2079
114. W. Gao, S. Fan, G. Li, W. Lin, A thorough benchmark and a new model for light field saliency detection. *IEEE Trans. Pattern Anal. Mach. Intell.* (2023)
115. W. Gao, G. Li, *Deep Learning for 3D Point Clouds* (Springer Nature, 2024)
116. W. Gao, G. Li, Introduction to 3d point clouds: Datasets and perception, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 1–27
117. W. Gao, G. Li, Learning basics for 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 29–70
118. W. Gao, G. Li, Deep-learning-based point cloud enhancement i, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 71–97
119. W. Gao, G. Li, Deep-learning-based point cloud enhancement ii, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 99–130
120. W. Gao, G. Li, Deep-learning-based point cloud analysis i, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 131–162
121. W. Gao, G. Li, Deep-learning-based point cloud analysis ii, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 163–193
122. W. Gao, G. Li, Point cloud pre-trained models and large models, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 195–225
123. W. Gao, G. Li, Point cloud-language multi-modal learning, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 227–254
124. W. Gao, G. Li, Open-source projects for 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 255–272
125. W. Gao, G. Li, Typical engineering applications of 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2025), pp. 273–299
126. W. Gao, G. Li, Future work on deep learning-based point cloud technologies, in *Deep Learning for 3D Point Clouds* (Springer, 2025), pp. 301–315
127. Z. Wang, A.C. Bovik, A comprehensive review on image quality assessment. *IEEE Signal Process. Mag.* 28(6), 98–117 (2011)

128. Y. Guo, W. Gao, G. Li, Interpretable task-inspired adaptive filter pruning for neural networks under multiple constraints. *Int. J. Comput. Vision* 1–17 (2024)
129. W. Gao, Y. Guo, S. Ma, G. Li, S. Kwong, Efficient neural network compression inspired by compressive sensing. *IEEE Trans. Neural Netw. Learn. Syst.* (2022)
130. Y. Guo, W. Gao, Semantic-driven automatic filter pruning for neural networks, in *IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
131. L. Tao, W. Gao, Efficient channel pruning based on architecture alignment and probability model bypassing, in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2021), pp. 3232–3237
132. Y. Li, Z. Yu, X. Chen, S. Wang, H. Yuan, W. Gao, Zero-shot quantization for large-kernels via shape-based distribution and diversity self-distillation, in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2025)
133. X. Zang, W. Gao, G. Li, H. Fang, C. Ban, Z. He, H. Sun, A baseline investigation: Transformer-based cross-view baseline for text-based person search, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 7737–7746
134. G. Liao, W. Gao, G. Li, J. Wang, S. Kwong, Cross-collaborative fusion-encoder network for robust rgb-thermal salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **32**(11), 7646–7661 (2022)
135. W. Gao, G. Liao, S. Ma, G. Li, Y. Liang, W. Lin, Unified information fusion network for multimodal rgb-d and rgb-t salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **32**(4), 2091–2106 (2021)
136. Y. Chen, S. Sun, G. Li, W. Gao, T.H. Li, Closing the gap between theory and practice during alternating optimization for gans. *IEEE Trans. Neural Netw. Learn. Syst.* (2023)
137. Y. Chen, C. Jin, G. Li, T. H. Li, W. Gao, Mitigating label noise in gans via enhanced spectral normalization. *IEEE Trans. Circuits Syst. Video Technol.* (2023)
138. X. Zang, G. Li, W. Gao, Multidirection and multiscale pyramid in transformer for video-based pedestrian retrieval. *IEEE Trans. Ind. Inform.* **18**(12), 8776–8785 (2022)
139. X. Zang, G. Li, W. Gao, X. Shu, Learning to disentangle scenes for person re-identification. *Image Vision Comput.* **116**, 104330 (2021)
140. X. Zang, G. Li, W. Gao, X. Shu, Exploiting robust unsupervised video person re-identification. *IET Image Process.* **16**(3), 729–741 (2022)
141. Z. Yue, G. Li, W. Gao, Cross-level guided attention for human-object interaction detection, in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2023), pp. 284–289
142. S. Sun, J. Liu, H. Li, G. Liu, T.H. Li, W. Gao, Streamflow: Streamlined multi-frame optical flow estimation for video sequences, in *Advances in Neural Information Processing Systems* (2024)
143. Z. Yao, W. Gao, Iterative saliency aggregation and assignment network for efficient salient object detection in optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* (2024)
144. G. Liao, W. Gao, Rethinking feature mining for light field salient object detection. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(10), 1–24 (2024)
145. B. Yang, W. Gao, G. Li, Focus and adjust: Progressive refinement network for human object interaction detection, in *2022 26th International Conference on Pattern Recognition (ICPR)* (IEEE, 2022), pp. 2546–2552
146. G. Liao, W. Gao, Q. Jiang, R. Wang, G. Li, Mmnet: multi-stage and multi-scale fusion network for rgb-d salient object detection, in *Proceedings of the 28th ACM International Conference on Multimedia* (2020), pp. 2436–2444
147. W. Gao, S. Kwong, Y. Zhou, Y. Jia, J. Zhang, W. Wu, Multiscale phase congruency analysis for image edge visual saliency detection, in *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 1 (IEEE, 2016), pp. 75–80
148. W. Gao, S. Kwong, Y. Zhou, X. Wang, Phase congruency analysis of down-sampled and blurring images for foreground extraction, in *2015 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)* (IEEE, 2015), pp. 152–157

149. S. Fan, W. Gao, G. Li, Salient object detection for point clouds, in *European Conference on Computer Vision* (Springer, 2022), pp. 1–19
150. S. Luo, W. Gao, A general framework for rotation invariant point cloud analysis, in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2024), pp. 3665–3669
151. X. Lu, W. Gao, Attentivenet: Detecting small objects for lidar point clouds by attending to important points, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
152. Z. Pan, N. Zhang, W. Gao, S. Liu, G. Li, Less is more: label recommendation for weakly supervised point cloud semantic segmentation. *Proc. AAAI Conf. Artif. Intell.* **38**(5), 4397–4405 (2024)
153. Z. Pan, G. Liu, W. Gao, T.H. Li, Epcontrast: Effective point-level contrastive learning for large-scale point cloud understanding, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024), pp. 1–6
154. N. Zhang, Z. Pan, T.H. Li, W. Gao, G. Li, Improving graph representation for point cloud segmentation via attentive filtering, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 1244–1254
155. K. Wen, N. Zhang, G. Li, W. Gao, Mpvnn: multi-resolution point-voxel non-parametric network for 3d point cloud processing, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024), pp. 1–6
156. X. Chen, Y. Wang, Y. Li, X. Ling, M. Li, R. Liu, M. Ouyang, K. Zhao, T. Guan, Y. He, Low bit-width zero-shot quantization with soft feature-infused hints for iot systems. *IEEE Internet Things J.* (2024)
157. Z. Pan, N. Zhang, W. Gao, S. Liu, G. Li, Point cloud semantic segmentation with sparse and inhomogeneous annotations, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
158. R. Zhang, G. Li, W. Gao, S. Liu, A quantum-inspired framework in master-servant mode for large-scale multi-modal place recognition. *IEEE Trans. Intell. Transp. Syst.* (2024)
159. D. Yang, W. Gao, Pointchd: a point cloud benchmark for congenital heart disease classification and segmentation. *IEEE J. Biomed. Health Inform.* (2024)
160. S. Fan, W. Gao, G. Li, Point-mpp: Point cloud self-supervised learning from masked position prediction. *IEEE Trans. Neural Netw. Learn. Syst.* (2024)
161. Z. Pan, W. Gao, G. Li, Distribution guidance network for weakly supervised point cloud semantic segmentation. *Adv. Neural Inf. Process. Syst.* (2024)
162. R. Zhang, G. Li, W. Gao, T.H. Li, Compoint: can complex-valued representation benefit point cloud place recognition? *IEEE Trans. Intell. Transp. Syst.* (2024)
163. ITU-T, Methods for subjective determination of transmission quality. *ITU-T Recommendation P 800* (1996)
164. ITU-T, Subjective video quality assessment methods for multimedia applications. *ITU-T Recommendation P 910* (2008)
165. ITU-R, Methodology for the subjective assessment of the quality of television pictures. *ITU-R Recommendation BT 500(13)* (2012)
166. ITU-T, Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment. *ITU-T Recommendation P 913* (2016)
167. H. David, *The Method of Paired Comparisons* (Oxford University Press, 1988)
168. I.T. Union, Methodology for the subjective assessment of the quality of television pictures. *ITU-R Recommendation BT 500(13)* (2012)
169. Z. Shang, J.P. Ebenezzer, A.C. Bovik, Y. Wu, H. Wei, S. Sethuraman, Subjective assessment of high dynamic range videos under different ambient conditions, in *IEEE International Conference on Image Processing* (2022), pp. 786–790
170. K. Seshadrinathan, R. Soundararajan, A.C. Bovik, L.K. Cormack, Study of subjective and objective quality assessment of video. *IEEE Trans. Image Process.* **19**(6), 1427–1441 (2010)

171. P.V. Vu, D.M. Chandler, Vis 3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. *J. Electron. Imaging* **23**(1), 013 016–013 016 (2014)
172. N. Barman, Y. Reznik, M.G. Martini, A subjective dataset for multi-screen video streaming applications, in *International Conference on Quality of Multimedia Experience* (2023), pp. 270–275
173. J. Wang, S. Wang, Z. Wang, Asymmetrically compressed stereoscopic 3d videos: quality assessment and rate-distortion performance evaluation. *IEEE Trans. Image Process.* **26**(3), 1330–1343 (2017)
174. C.G. Bampis, Z. Li, A.K. Moorthy, I. Katsavounidis, A. Aaron, A.C. Bovik, Study of temporal effects on subjective video quality of experience. *IEEE Trans. Image Process.* **26**(11), 5217–5231 (2017)
175. C.G. Bampis, Z. Li, I. Katsavounidis, T.-Y. Huang, C. Ekanadham, A.C. Bovik, Towards perceptually optimized adaptive video streaming-a realistic quality of experience database. *IEEE Trans. Image Process.* **30**, 5182–5197 (2021)
176. N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti et al., Image database tid2013: peculiarities, results and perspectives. *Signal Process. Image Commun.* **30**, 57–77 (2015)
177. H.-J. Park, D.-H. Har, Subjective image quality assessment based on objective image quality measurement factors. *IEEE Trans. Consum. Electron.* **57**(3), 1176–1184 (2011)
178. Q. Huynh-Thu, M. Ghanbari, Scope of validity of psnr in image/video quality assessment. *Electron. Lett.* **44**(13), 800–801 (2008)
179. Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**(4), 600–612 (2004)
180. R. Soundararajan, A.C. Bovik, Rred indices: reduced reference entropic differencing for image quality assessment. *IEEE Trans. Image Process.* **21**(2), 517–526 (2012)
181. A. Mittal, A.K. Moorthy, A.C. Bovik, No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.* **21**(12), 4695–4708 (2012)
182. L. Kang, P. Ye, Y. Li, D. Doermann, Convolutional neural networks for no-reference image quality assessment, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 1733–1740
183. D. Varga, D. Saupe, T. Szirányi, Deepn: a content preserving deep architecture for blind image quality assessment, in *IEEE International Conference on Multimedia and Expo* (2018), pp. 1–6
184. S. Bianco, L. Celona, G. Ciocca, P. Napolitano, Two-stream convolutional networks for blind image quality assessment, in *IEEE International Conference on Image Processing* (2018), pp. 3773–3777
185. W. Lin, J. Wang, Peiran Ren, X. Chu, M.-H. Yang, Hallucinated-iqa: No-reference image quality assessment via adversarial learning. *IEEE Trans. Multimedia* **20**(8), 2004–2015 (2018)
186. J. You, J. Korhonen, Transformer for image quality assessment, in *IEEE International Conference on Image Processing* (2021), pp. 1389–1393
187. S.A. Golestaneh, S. Dadsetan, K.M. Kitani, No-reference image quality assessment via transformers, relative ranking, and self-consistency, in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision* (2022), pp. 1220–1227
188. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in *European Conference on Computer Vision* (2020), pp. 213–229
189. N. Le, H. Zhang, F. Cricri, R. Ghaznavi-Youvalari, E. Rahtu, Image coding for machines: an end-to-end learned approach, in *IEEE International Conference on Acoustics, Speech and Signal Processing* (2021), pp. 1590–1594
190. A. Saha, S.K. Pentapati, Z. Shang, R. Pahwa, B. Chen, H.E. Gedik, S. Mishra, A.C. Bovik, Perceptual video quality assessment: the journey continues! *Front. Signal Process.* **3**, 1193523 (2023)
191. H. Duan, G. Zhai, X. Yang, D. Li, W. Zhu, Ivqad, an immersive video quality assessment database, in *International Conference on Systems, Signals and Image Processing* (2017), pp. 1–5

192. W. Lin, C.-C.J. Kuo, Perceptual visual quality metrics: a survey. *J. Vis. Commun. Image Represent.* **22**(4), 297–312 (2011)
193. Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**(4), 600–612 (2004)
194. Z. Wang, A.C. Bovik, Reduced-and no-reference image quality assessment. *IEEE Signal Process. Mag.* **28**(6), 29–40 (2011)
195. P. Chen, L. Li, L. Ma, J. Wu, G. Shi, Rirnet: recurrent-in-recurrent network for video quality assessment, in *Proceedings of the 28th ACM International Conference on Multimedia* (2020), pp. 834–842
196. B. Chen, L. Zhu, G. Li, F. Lu, H. Fan, S. Wang, Learning generalized spatial-temporal deep feature representation for no-reference video quality assessment. *IEEE Trans. Circuits Syst. Video Technol.* **32**(4), 1903–1916 (2021)
197. J. Korhonen, Two-level approach for no-reference consumer video quality assessment. *IEEE Trans. Image Process.* **28**(12), 5923–5938 (2019)
198. Z. Tu, X. Yu, Y. Wang, N. Birkbeck, B. Adsumilli, A.C. Bovik, Rapique: rapid and accurate video quality prediction of user generated content. *IEEE Open J. Signal Process.* **2**, 425–440 (2021)
199. X. Min, H. Duan, W. Sun, Y. Zhu, G. Zhai, Perceptual video quality assessment: a survey (2024). [arXiv:2402.03413](https://arxiv.org/abs/2402.03413)
200. M. Smirnov, A. Gushchin, A. Antsiferova, D. Vatolin, R. Timofte, Z. Jia, Z. Zhang, W. Sun, J. Qian, Y. Cao et al., Aim 2024 challenge on compressed video quality assessment: methods and results (2024). [arXiv:2408.11982](https://arxiv.org/abs/2408.11982)
201. W. Liu, Z. Duanmu, Z. Wang, End-to-end blind quality assessment of compressed videos using deep neural networks, in *ACM Multimedia* (2018), pp. 546–554
202. W. Zhou, X. Min, H. Li, Q. Jiang, A brief survey on adaptive video streaming quality assessment. *J. Vis. Commun. Image Represent.* **86**, 103526 (2022)
203. L. Li, P. Chen, W. Lin, M. Xu, G. Shi, From whole video to frames: weakly-supervised domain adaptive continuous-time qoe evaluation. *IEEE Trans. Image Process.* **31**, 4937–4951 (2022)
204. C. Galkandage, J. Calic, S. Dogan, J.-Y. Guillemaut, Full-reference stereoscopic video quality assessment using a motion sensitive hvs model. *IEEE Trans. Circuits Syst. Video Technol.* **31**(2), 452–466 (2020)
205. B. Appina, S.V.R. Dendi, K. Manasa, S.S. Channappayya, A.C. Bovik, Study of subjective quality and objective blind quality prediction of stereoscopic videos. *IEEE Trans. Image Process.* **28**(10), 5027–5040 (2019)
206. Y. Fang, L. Huang, J. Yan, X. Liu, Y. Liu, Perceptual quality assessment of omnidirectional images. *Proc. AAAI Conf. Artif. Intell.* **36**(1), 580–588 (2022)
207. W. Sun, K. Gu, S. Ma, W. Zhu, N. Liu, G. Zhai, A large-scale compressed 360 degree spherical image database: from subjective quality evaluation to objective model comparison, in *IEEE International Workshop on Multimedia Signal Processing* (2018), pp. 1–6
208. L. Xie, M. Hu, X. Bai, towards hardware-friendly and robust facial landmark detection method, in *International Conference on Neural Information Processing* (Springer, 2022), pp. 432–444
209. L. Xie, M. Hu, X. Bai, online improved vehicle tracking accuracy via unsupervised route generation, in *IEEE 34th International Conference on Tools with Artificial Intelligence* (IEEE, 2022), pp. 788–792
210. H. Su, C. Jung, Perceptual enhancement of low light images based on two-step noise suppression. *IEEE Access* **6**, 7005–7018 (2018)
211. H. Su, L. Yu, C. Jung, Joint contrast enhancement and noise reduction of low light images via jnd transform. *IEEE Trans. Multimedia* **24**, 17–32 (2020)
212. S.-H. Bae, J. Kim, M. Kim, Hvc-based perceptually adaptive video coding using a dct-based local distortion detection probability model. *IEEE Trans. Image Process.* **25**(7), 3343–3357 (2016)
213. X. Zhang, S. Wang, K. Gu, W. Lin, S. Ma, W. Gao, Just-noticeable difference-based perceptual optimization for jpeg compression. *IEEE Signal Process. Lett.* **24**(1), 96–100 (2016)

214. M. Zhou, X. Wei, S. Kwong, W. Jia, B. Fang, Just noticeable distortion-based perceptual rate control in hevcc. *IEEE Trans. Image Process.* **29**, 7603–7614 (2020)
215. T. Tian, H. Wang, L. Zuo, C.-C.J. Kuo, S. Kwong, Just noticeable difference level prediction for perceptual image compression. *IEEE Trans. Broadcast.* **66**(3), 690–700 (2020)
216. T. Zhu, L. Karam, A no-reference objective image quality metric based on perceptually weighted local noise. *EURASIP J. Image Video Process.* **2014**, 1–8 (2014)
217. Y. Zhang, H. Liu, Y. Yang, X. Fan, S. Kwong, C.J. Kuo, Deep learning based just noticeable difference and perceptual quality prediction models for compressed video. *IEEE Trans. Circuits Syst. Video Technol.* **32**(3), 1197–1212 (2021)
218. W. Wan, K. Zhou, K. Zhang, Y. Zhan, J. Li, Jnd-guided perceptually color image watermarking in spatial domain. *IEEE Access* **8**, 164 504–164 520 (2020)
219. K. Zhou, Y. Zhang, J. Li, Y. Zhan, W. Wan, Spatial-perceptual embedding with robust just noticeable difference model for color image watermarking. *Mathematics* **8**(9), 1506 (2020)
220. L. Jin, J.Y. Lin, S. Hu, H. Wang, P. Wang, I. Katsavounidis, A. Aaron, C.-C.J. Kuo, Statistical study on perceived jpeg image quality via mcl-jci dataset construction and analysis. *Electron. Imaging* **2016**(13), 1–9 (2016)
221. H. Wang, W. Gan, S. Hu, J.Y. Lin, L. Jin, L. Song, P. Wang, I. Katsavounidis, A. Aaron, C.-C.J. Kuo, Mcl-jcv: a jnd-based h. 264/avc video quality assessment dataset, in *IEEE International Conference on Image Processing* (IEEE, 2016), pp. 1509–1513
222. H. Wang, I. Katsavounidis, J. Zhou, J. Park, S. Lei, X. Zhou, M.-O. Pun, X. Jin, R. Wang, X. Wang et al., Videoset: a large-scale compressed video quality dataset based on jnd measurement. *J. Vis. Commun. Image Represent.* **46**, 292–302 (2017)
223. Y. Liu, J. Jin, Y. Xue, W. Lin, The first comprehensive dataset with multiple distortion types for visual just-noticeable differences, in *IEEE International Conference on Image Processing* (2023), pp. 2820–2824
224. C.-H. Chou, Y.-C. Li, A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile. *IEEE Trans. Circuits Syst. Video Technol.* **5**(6), 467–476 (1995)
225. J. Wu, F. Qi, G. Shi, Self-similarity based structural regularity for just noticeable difference estimation. *J. Vis. Commun. Image Represent.* **23**(6), 845–852 (2012)
226. J. Wu, G. Shi, W. Lin, A. Liu, F. Qi, Just noticeable difference estimation for images with free-energy principle. *IEEE Trans. Multimedia* **15**(7), 1705–1710 (2013)
227. J. Wu, W. Lin, G. Shi, X. Wang, F. Li, Pattern masking estimation in image with structural uncertainty. *IEEE Trans. Image Process.* **22**(12), 4892–4904 (2013)
228. J. Wu, L. Li, W. Dong, G. Shi, W. Lin, C.-C.J. Kuo, Enhanced just noticeable difference model for images with pattern complexity. *IEEE Trans. Image Process.* **26**(6), 2682–2693 (2017)
229. X. Shen, Z. Ni, W. Yang, X. Zhang, S. Wang, S. Kwong, Just noticeable distortion profile inference: a patch-level structural visibility learning approach. *IEEE Trans. Image Process.* **30**, 26–38 (2020)
230. H. Wang, L. Yu, J. Liang, H. Yin, T. Li, S. Wang, Hierarchical predictive coding-based jnd estimation for image compression. *IEEE Trans. Image Process.* **30**, 487–500 (2020)
231. Y. Jia, W. Lin, A.A. Kassim, Estimating just-noticeable distortion for video. *IEEE Trans. Circuits Syst. Video Technol.* **16**(7), 820–829 (2006)
232. Z. Wei, K.N. Ngan, Spatio-temporal just noticeable distortion profile for grey scale image/video in dct domain. *IEEE Trans. Circuits Syst. Video Technol.* **19**(3), 337–346 (2009)
233. S.-H. Bae, M. Kim, A novel generalized dct-based jnd profile based on an elaborate cm-jnd model for variable block-sized transforms in monochrome images. *IEEE Trans. Image process.* **23**(8), 3227–3240 (2014)
234. S.-H. Bae, M. Kim, A dct-based total jnd profile for spatiotemporal and foveated masking effects. *IEEE Trans. Circuits Syst. Video Technol.* **27**(6), 1196–1207 (2017)
235. W. Wan, J. Wu, X. Xie, G. Shi, A novel just noticeable difference model via orientation regularity in dct domain. *IEEE Access* **5**, 22 953–22 964 (2017)
236. S. Ki, S.-H. Bae, M. Kim, H. Ko, Learning-based just-noticeable-quantization-distortion modeling for perceptual video coding. *IEEE Trans. Image Process.* **27**(7), 3178–3193 (2018)

237. Q. Zhang, S. Wang, X. Zhang, S. Ma, W. Gao, Just recognizable distortion for machine vision oriented image and video coding. *Int. J. Comput. Vision* **129**(10), 2889–2906 (2021)
238. J. Jin, X. Zhang, X. Fu, H. Zhang, W. Lin, J. Lou, Y. Zhao, Does deep machine vision have just noticeable difference (jnd) (2021). [arXiv:2102.08168](#)
239. J. Jin, X. Zhang, X. Fu, H. Zhang, W. Lin, J. Lou, Y. Zhao, Just noticeable difference for deep machine vision. *IEEE Trans. Circuits Syst. Video Technol.* **32**(6), 3452–3461 (2021)
240. S. Chen, J. Jin, L. Meng, W. Lin, Z. Chen, T.-S. Chang, Z. Li, H. Zhang, A new image codec paradigm for human and machine uses (2021). [arXiv:2112.10071](#)
241. A.K. Akan, E. Akbas, F.T.Y. Vural, Just noticeable difference for machine perception and generation of regularized adversarial images with minimal perturbation. *Signal, Image Video Process.* **16**(6), 1595–1606 (2022)
242. X. Deng, P. Zhang, W. Liu, H. Lu, Recurrent multi-scale transformer for high-resolution salient object detection, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 7413–7423
243. Y.A.D. Djilali, K. McGuinness, N. O'Connor, Learning saliency from fixations, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2024), pp. 383–393
244. E. Vig, M. Dorr, D. Cox, Large-scale optimization of hierarchical features for saliency prediction in natural images, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 2798–2805
245. M. Kummerer, T.S. Wallis, L.A. Gatys, M. Bethge, Understanding low-and high-level contributions to fixation prediction, in *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 4789–4798
246. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition (2014). [arXiv:1409.1556](#)
247. X. Huang, C. Shen, X. Boix, Q. Zhao, Salicon: reducing the semantic gap in saliency prediction by adapting deep neural networks, in *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 262–270
248. W. Wang, J. Shen, Deep visual attention prediction. *IEEE Trans. Image Process.* **27**(5), 2368–2378 (2017)
249. A. Kroner, M. Senden, K. Driessens, R. Goebel, Contextual encoder-decoder network for visual saliency prediction. *Neural Netw.* **129**, 261–270 (2020)
250. S. Fan, Z. Shen, M. Jiang, B.L. Koenig, J. Xu, M.S. Kankanhalli, Q. Zhao, Emotional attention: a study of image sentiment and visual attention, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 7521–7531
251. S. Jia, N.D. Bruce, Eml-net: an expandable multi-layer network for saliency prediction. *Image Vision Comput.* **95**, 103887 (2020)
252. Z. Che, A. Borji, G. Zhai, X. Min, G. Guo, P. Le Callet, How is gaze influenced by image transformations? dataset and model. *IEEE Trans. Image Process.* **29**, 2287–2300 (2019)
253. O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2015), pp. 234–241
254. R. Droste, J. Jiao, J.A. Noble, Unified image and video saliency modeling, in *European Conference on Computer Vision* (2020), pp. 419–435
255. Z. Wu, L. Su, Q. Huang, Stacked cross refinement network for edge-aware salient object detection, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 7264–7273
256. J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, M.-M. Cheng, Egnet: edge guidance network for salient object detection, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 8779–8788
257. J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, Q. Tian, Label decoupling framework for salient object detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 13 025–13 034

258. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 770–778
259. X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, M. Jagersand, Basnet: boundary-aware salient object detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 7479–7489
260. C. Yang, L. Zhang, H. Lu, X. Ruan, M.-H. Yang, Saliency detection via graph-based manifold ranking, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 3166–3173
261. P. Jiang, H. Ling, J. Yu, J. Peng, Salient region detection by ufo: uniqueness, focusness and objectness, in *Proceedings of the IEEE International Conference on Computer Vision* (2013), pp. 1976–1983
262. W. Zhu, S. Liang, Y. Wei, J. Sun, Saliency optimization from robust background detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 2814–2821
263. R. Liu, J. Cao, Z. Lin, S. Shan, Adaptive partial differential equation learning for visual saliency detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 3866–3873
264. H. Li, H. Lu, Z. Lin, X. Shen, B. Price, Inner and inter label propagation: salient object detection in the wild. *IEEE Trans Image Process.* **24**(10), 3176–3186 (2015)
265. Z. Wu, L. Su, Q. Huang, Cascaded partial decoder for fast and accurate salient object detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 3907–3916
266. Y. Pang, X. Zhao, L. Zhang, H. Lu, Multi-scale interactive network for salient object detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 9413–9422
267. N. Liu, N. Zhang, K. Wan, L. Shao, J. Han, Visual saliency transformer, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 4722–4732
268. Y.-H. Wu, Y. Liu, L. Zhang, M.-M. Cheng, B. Ren, Edn: Salient object detection via extremely-downsampled network. *IEEE Trans. Image Process.* **31**, 3125–3136 (2022)
269. Y. Gu, L. Wang, Z. Wang, Y. Liu, M.-M. Cheng, S.-P. Lu, Pyramid constrained self-attention network for fast video salient object detection. *Proc. AAAI Conf. Artif. Intell.* **34**(07), 10 869–10 876 (2020)
270. A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan et al., Searching for mobilenetv3, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 1314–1324
271. S. Liu, D. Huang et al., Receptive field block net for accurate and fast object detection, in *European Conference on Computer Vision* (2018), pp. 385–400
272. Z. Ren, S. Gao, D. Rajan, L.-T. Chia, Y. Huang, Spatiotemporal saliency detection via sparse representation, in *IEEE International Conference on Multimedia and Expo* (2012), pp. 158–163
273. Z. Liu, X. Zhang, S. Luo, O. Le Meur, Superpixel-based spatiotemporal saliency detection. *IEEE Trans. Circuits Syst. Video Technol.* **24**(9), 1522–1540 (2014)
274. W. Wang, J. Shen, L. Shao, Consistent video saliency using local gradient flow optimization and global refinement. *IEEE Trans. Image Process.* **24**(11), 4185–4196 (2015)
275. W. Wang, J. Shen, F. Porikli, Saliency-aware geodesic video object segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 3395–3402
276. T. Xi, W. Zhao, H. Wang, W. Lin, Salient object detection with spatiotemporal background priors for video. *IEEE Trans. Image Process.* **26**(7), 3425–3436 (2016)
277. Y. Fang, G. Ding, J. Li, Z. Fang, Deep3dsaliency: deep stereoscopic video saliency detection model by 3d convolutional networks. *IEEE Trans. Image Process.* **28**(5), 2305–2318 (2018)
278. K. Min, J.J. Corso, Tased-net: temporally-aggregating spatial encoder-decoder network for video saliency detection, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 2394–2403

- 279. Y.-W. Chen, X. Jin, X. Shen, M.-H. Yang, Video salient object detection via contrastive features and attention modules, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2022), pp. 1320–1329
- 280. N. Liu, K. Nan, W. Zhao, X. Yao, J. Han, Learning complementary spatial–temporal transformer for video salient object detection. *IEEE Trans. Neural Netw. Learn. Syst.* 1–11 (2023)
- 281. T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, H.-Y. Shum, Learning to detect a salient object. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(2), 353–367 (2010)
- 282. Y. Li, X. Hou, C. Koch, J.M. Rehg, A.L. Yuille, The secrets of salient object segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 280–287
- 283. G. Li, Y. Yu, ‘Visual saliency based on multiscale deep features, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 5455–5463
- 284. L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, X. Ruan, Learning to detect salient objects with image-level supervision, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 136–145
- 285. D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, A. Borji, Salient objects in clutter: bringing salient object detection to the foreground, in *European Conference on Computer Vision* (2018), pp. 186–202
- 286. F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, A. Sorkine-Hornung, A benchmark dataset and evaluation methodology for video object segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 724–732
- 287. D.-P. Fan, W. Wang, M.-M. Cheng, J. Shen, Shifting more attention to video salient object detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 8554–8564
- 288. F. Li, T. Kim, A. Humayun, D. Tsai, J.M. Rehg, Video segmentation by tracking many figure-ground segments, in *Proceedings of the IEEE International Conference on Computer Vision* (2013), pp. 2192–2199
- 289. J. Li, C. Xia, X. Chen, A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection. *IEEE Trans. Image Process.* **27**(1), 349–364 (2017)

Chapter 4

Deep Learning-based Image Coding



4.1 Introduction

Learned image compression (LIC) has been widely developed and applied in recent years [1–23]. It is widely applied in fields like multimedia [11–64], and computer vision [65–97, 97–101, 101–142], offering higher compression rates and better visual quality compared to traditional methods. The core idea of LIC is to learn a compact representation directly from the input image via a deep neural network, rather than relying on a traditional hand-craft encoder. LIC has many advantages over traditional codecs. First, traditional codecs design more complicated tools to optimize each component in the compression loop independently. Due to module interactions, the optimization of individual modules cannot guarantee optimal overall performance. While LIC methods jointly optimize all modules through a unified RD loss function, ensuring that the trained model achieves the global optimal. Second, benefiting from powerful nonlinear transformations and efficient entropy models, LIC learns more compact latent representations and accurate entropy models. The rate-distortion (R-D) performance of deep image compression has gradually outperformed that of conventional codecs, such as JPEG [143], JPEG2000 [144], and even the best conventional coding standard VVC [145]. Third, compared to traditional coding, LIC offers greater flexibility and adjustability. It has a more significant human visual perception advantage, and existing methods have already greatly surpassed traditional methods in terms of MS-SSIM [146], a human-oriented image quality assessment.

Next, we will present a comprehensive review of LIC methods. Lossless image compression, lossy image compression, and generative image compression will be introduced sequentially.

4.2 Typical Methods of Lossless Image Coding

The purpose of lossless compression is to reduce data storage space or transmission size without losing any information. Probability-based compression algorithms, such as Huffman coding and arithmetic coding, have become the most commonly used methods. According to Shannon's information theory [147], the bit rate when using estimated probability distribution $Q(X)$ to perform entropy coding on a discrete symbol set X with an actual marginal probability distribution $P(X)$ is as follows:

$$\begin{aligned} R(X) &= H(P(X)) + D_{\text{KL}}(P(X) \parallel Q(X)) \\ &= - \sum_i P(x_i) \log_2 P(x_i) + \sum_i P(x_i) \log_2 \left(\frac{P(x_i)}{Q(x_i)} \right), \end{aligned} \quad (4.1)$$

where $H(p(x))$ represents the information entropy of the original symbol source, which is the lower bound of the coding rate. The KL divergence represents the rate increase caused by errors in probability estimation. Therefore, accurate probability estimation for image distributions is the key to lossless entropy coding. Recent discrete generative models have made great progress in probability distribution modeling. They are intricately connected with lossless compression since entropy encoding needs the guide of probability distribution estimation. However, some state-of-the-art generative models predict the probability in an autoregressive manner, such as PixelRNN [148], PixelCNN++ [149], and Multiscale-PixelCNN [150], which are time-consuming and renders them hard to be applied in practice. Mentzer et al. [151] proposed the first lossless compression work to address aforementioned problems. They developed a fully parallel learned framework outperforming traditional methods PNG, WebP and JPEG2000 as shown in Fig. 4.1. Hierarchical feature extractors are introduced, and three auxiliary hierarchical feature representations are learned to help with the prediction. The three parallel autoencoders extract features at different scales [152, 153]. The predictor transforms pixel probability estimation from an autoregressive approach to joint distribution probability prediction between the image and latent features as follows [151]:

$$\begin{aligned} p(x, y^{(1)}, \dots, y^{(S)}) &= p(x|y^{(1)}, \dots, y^{(S)}) \prod_{s=1}^S p(y^{(s)}|y^{(s+1)}, \dots, y^{(S)}) \\ &= p(x|y^{(1)}, \dots, y^{(S)}) \prod_{s=1}^S p(y^{(s)}|g^{(s+1)}). \end{aligned} \quad (4.2)$$

The parallel structure of their model allows for significantly faster decoding compared to autoregressive models, thus making it possible to entropy coding in practice. The system is trained to minimize the cross entropy of the different representations where conditional probabilities were modeled with K mixture discretized logistic distributions proposed in [149].

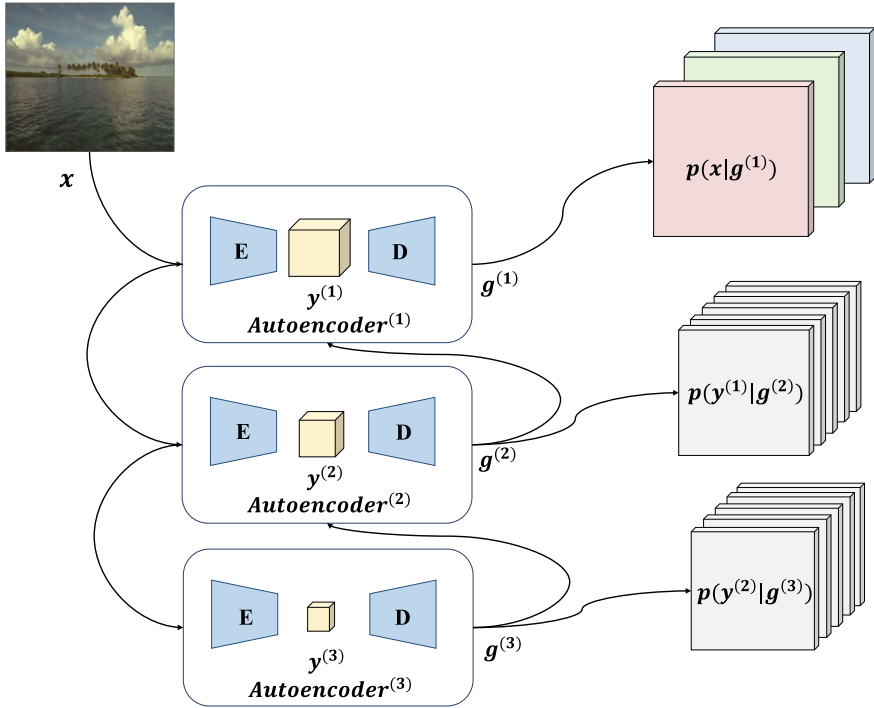


Fig. 4.1 Overall structure of L3C. ©2019 IEEE Reprinted, with permission, from [151]

L3C [151] can be viewed as an autoencoder (AE). Because the overall image can be predicted in one inference pass, methods assisted with latent representations have significant advantages in terms of decoding speed. Townsend et al. [154] addressed the issue that bits back coding is incompatible with the arithmetic coding scheme by outlining a scheme that utilizes asymmetric numeral systems termed Bits Back with ANS (BB-ANS) to implement bits back instead. However, BB-ANS is only feasible for latent models with one layer. Following that, Kingma et al. [155] proposed the Bit-Swap scheme that allows for efficient compression using hierarchical latent variable models structured as a Markov chain by bypassing the large number of initial bits required by BB-ANS for compression with the same models. These hierarchical latent variable models serve as effective density estimators, and when paired with Bit-Swap, they create a low-overhead, efficient lossless compression algorithm capable of handling complex, high-dimensional datasets with ease. Hoogetboom et al. [156] first applied continuous flow-based generative models to lossless compression, where digital media is stored discretely. They proposed Integer Discrete Flows (IDF) [156] followed by IDF++ [157] to overcome the (de)quantization problems. Both approaches employ the reversible transformation connecting the discrete data and the latent variables. Similarly, iVPF [158] achieved a bijection in discrete space using volume-preserving flows. However, these models must enforce certain

constraints on the flow layers to maintain this bijection, which restricts the flexibility of the transformations. Local Bits-Back Coding (LBB) [159] is the first to effectively apply lossless coding with a more versatile set of continuous flows, solving the issue of information loss by addressing numerical inaccuracies. Zhang et al. [160] introduced iFlow, a flow-based neural compression codec with numerical reversibility that unifies high-precision density estimators and efficient coding speeds [16, 29, 31, 32].

Some recent studies combined autoregressive models (ARMs) and variational autoencoders (VAEs) to further improve lossless image compression. Zhang et al. [10] introduced Bit Plane Slicing (BPS), which allows for effective utilization of latent variables and improves autoregression while reducing the risk of posterior collapse by organizing sub-images in order of decreasing importance for the latent representations. Kang et al. [161] developed a model that integrates an AR model with VQ-VAE, along with a semi-dynamic coder which is computationally efficient (Table 4.1).

Table 4.1 Summary of lossless learned image compression methods

Title	Authors	Publication	Method
Pixel recurrent neural networks [148]	Van et al.	ICML 2016	ARM
Pixelcnn++: a pixelcnn implementation with discretized logistic mixture [149]	Salimans et al.	ICLR 2017	ARM
Practical full resolution learned lossless image compression [151]	Mentzer et al.	CVPR 2019	AE
Practical lossless compression with latent variables using bits back coding [154]	Townsend et al.	ICLR 2019	VAE
Bit-swap: recursive bits-back coding for lossless compression with hierarchical latent variables [155]	Kingma et al.	ICML 2019	VAE
IDF++: analyzing and improving integer discrete flows for lossless compression [157]	Van et al.	ICLR 2021	Flow
iFlow: numerically invertible flows for efficient lossless compression via a uniform coder [160]	Zhang et al.	NeurIPS 2021	Flow
PILC: practical image lossless compression with an end-to-end GPU oriented neural framework [161]	Kang et al.	CVPR 2022	Combination
Learned lossless image compression based on bit plane slicing [10]	Zhang et al.	CVPR 2024	Combination

Source Author

4.3 Typical Methods of Lossy Image Coding

Learned lossy image compression typically employs a transform coding paradigm based on autoencoder frameworks as shown in Fig. 4.2. The entire system can be roughly divided into three modules, i.e., the analysis/synthesis transformation, quantization, and entropy model. The analysis transformation denoted by g_a acts as an encoder and transforms the original input image x through nonlinear transformations to extract compact latent representations y . Following that latent representations are quantized to discrete values $\hat{y}$, in order to be transmitted at limited bit rates R . The process of lossless entropy encoding quantized representations into bitstreams, such as arithmetic encoding, typically involves estimating the probability distribution. The entropy model plays a crucial role in determining the actual bit rates because for a given set of latent representations $\hat{y}$, the actual bit rates depend on the degree of alignment for the estimated probability distribution and the marginal distribution. Therefore amounts of works have been dedicated to improving entropy models to provide more accurate probability estimations, which will be elaborated on in detail later. On the decoding side, the inverse process is performed, where the bitstream is decoded into latent representations, and the image $\hat{x}$ is subsequently reconstructed. In this process, the nonlinear synthesis transformation g_s and the entropy model are important.

Corresponding to Fig. 4.2, the overall LIC algorithm can be summarized as follows [162]:

$$\begin{aligned} y &= g_a(x; \theta_{g_a}), \hat{y} = Q(y), \hat{x} = g_s(\hat{y}; \theta_{g_s}), \\ R &= \mathbb{E}_{\hat{y} \sim p(\hat{y})}[-\log_2(p_{em}(\hat{y}; \theta_{em}))], \end{aligned} \quad (4.3)$$

where θ_{g_a} , θ_{g_s} , and θ_{em} denote learnable network parameters of the encoder, decoder, and entropy model respectively. $p_{em}(\hat{y}; \theta_{em})$ denotes the probability distribution estimated by the entropy model and $p(\hat{y})$ is the marginal distribution of latent represen-

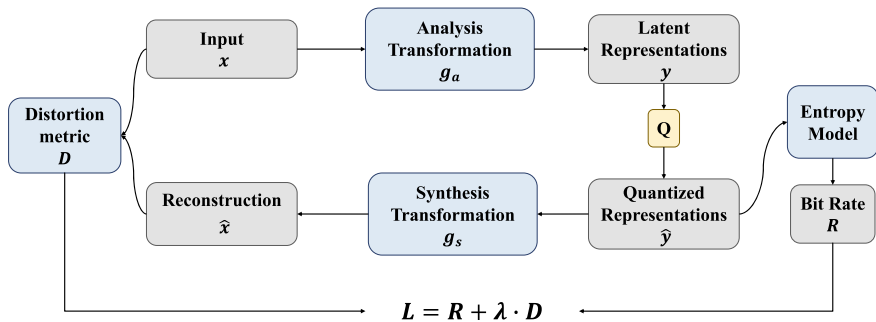


Fig. 4.2 The schematic diagram of learned lossy image compression. Public domain open access image [162]

tations. As mentioned in Sect. 4.1, LIC models can be optimized through a unified rate-distortion loss function as follows:

$$L = R + \lambda D(x, \hat{x}), \quad (4.4)$$

where the distortion metric D can be changed based on different purposes, such as mean squared error (MSE) and the Laplace Multiplicative Factor λ controls different bit rates. Based on the latent distribution modeling approach of the entropy model, the framework can be divided into three categories, which will be discussed later.

4.3.1 Framework

This section will provide a macro classification of LIC frameworks from the perspective of entropy models to give readers a clear and comprehensive understanding. The introduction will follow the chronological order of LIC technological development, and influential cornerstone works for each framework will be discussed. According to the overall scheme of the entropy model, lossy compression can be categorized into three main types, i.e., autoencoders (AEs) [162], variational autoencoders (VAEs) [163], and joint autoregressive and hyperprior (Joints) [164] as shown in Figs. 4.3, 4.4, and 4.5. Recent state-of-the-art methods adopt Joints architectures.

AE-based framework. AE-based methods directly entropy encode representations with fully factorized entropy models as follows [165]:

$$p_{\hat{y}|\psi}(\hat{y}|\psi) = \prod_i p_{\hat{y}|\psi^i}(\psi^i)(\hat{y}_i), \quad (4.5)$$

where ψ are learnable entropy parameters. Ballé et al. [165] utilized a piecewise linear function to fit the marginal probability distribution. Theis et al. [166] adopted the Gaussian scale mixtures model for the distribution. The natural application of autoencoders in the field of compression is simple, but its performance is insufficient. On the one hand, Ballé et al. [163] indicated that latent representations still exhibit significant spatial redundancy; for instance, adjacent values tend to be similar and may share the same distribution. However, the fully factorized entropy model cannot leverage this characteristic. On the other hand, the fully factorized entropy model only fits the overall distribution of the training dataset and cannot adaptively construct a sample-level probability distribution.

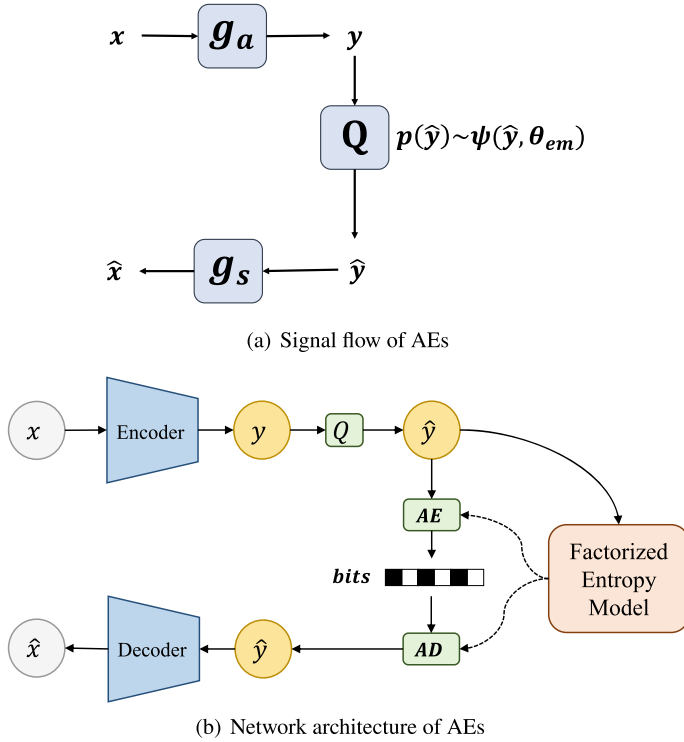


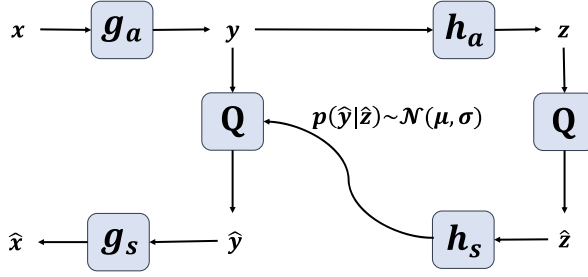
Fig. 4.3 The architecture illustration of AEs. Public domain open access image [162]

VAE-based framework. To overcome above challenge, variational autoencoders (VAEs) framework [163] expressed the representations as a parametric Gaussian distribution (see Fig. 4.4). Specifically, VAE-based methods introduce an auxiliary autoencoder named hyperprior autoencoder to capture the redundancy in latent variables, and the latent representations are modeled by the joint gaussian distribution with hyperprior variables as follows:

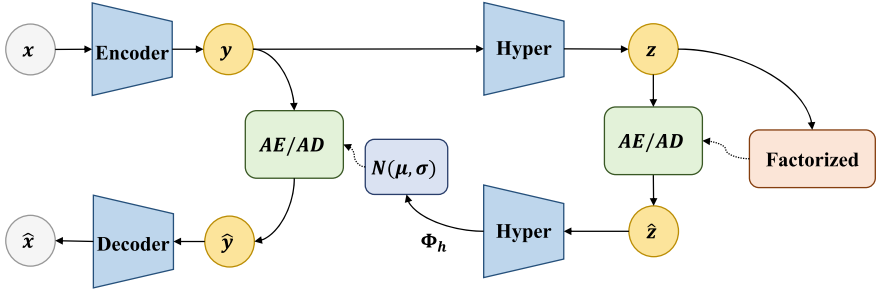
$$p_{\hat{y}|\hat{z}}(\hat{y}|\hat{z}) = \prod_i \mathcal{N}(\mu_i, \sigma_i^2)(\hat{y}_i) \quad (4.6)$$

with $\mu, \sigma = h_s(\hat{z}; \theta_h)$,

where $\mathcal{N}$ is the Gaussian distribution and θ_h is the learnable parameters of hyper decoder. The hyperprior variables model the spatial redundancy within the latent representations and can be viewed as a customized distribution which is associated with the input image. Hu et al. [167] extended multiple hyperprior architectures to capture spatial correlation in a wider range.



(a) Signal flow of VAEs



(b) Network architecture of VAEs

Fig. 4.4 The architecture illustration of VAEs. Public domain open access image [163]

Joint-based framework. Additionally, Minnen et al. [164] combined the successful autoregressive model [148] with hyperprior to further improve the entropy model as shown in Fig. 4.5. Joint autoregressive and hyperprior methods leverage the context information of previously decoded latent variables to better estimate the distribution parameters as follows:

$$p_{\hat{y}|\hat{z}, C_m}(\hat{y}|\hat{z}, C_m) = \prod_i \mathcal{N}(\mu_i, \sigma_i^2)(\hat{y}_i) \quad (4.7)$$

with $\mu, \sigma = h_e(\Phi_h; \Phi_c)$,

where h_e is the entropy parameter estimation network and Φ_h is hyperprior information. Φ_c is the context information provided by all previous decoded latent representations $\hat{y}^{<i}$. Minnen et al. [164] predicted the scale parameters of the distribution at the current coding position through autoregressive context modeling while Lee et al. [168] predicted the mean along with scale parameters in the same way.

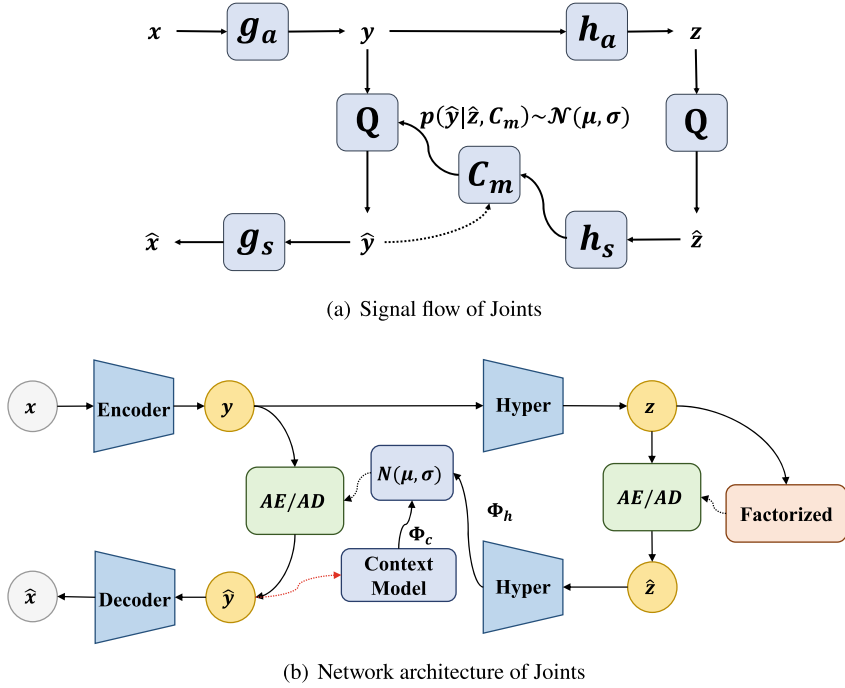


Fig. 4.5 The architecture illustration of Joints. Public domain open access image [164]

Aforementioned early works establish the direction for the development of LIC methods, and the Table 4.2 briefly summarizes their contributions. The foundation of LIC had been established since the joint architecture was proposed and subsequent research improved RD performance within this framework. To clarify the joint autoregressive and hyperprior framework, given a LIC model, the overall optimization objective is shown as:

$$\begin{aligned} \mathcal{L} &= R + \lambda D \\ &= \mathbb{E}_{x \sim p_x} [-\log_2 p_{\hat{y}|\hat{z}, C_m}(\hat{y}|\hat{z}, C_m)] + \mathbb{E}_{x \sim p_x} [-\log_2 p_{\hat{z}}(\hat{z})] + \lambda \cdot \mathbb{E}_{x \sim p_x} [D(x, \hat{x})], \end{aligned} \quad (4.8)$$

where the probability of $\hat{z}$ is generally estimated by a fully factorized model as follows:

$$p_{\hat{z}|\psi}(\hat{z}|\psi) = \prod_i p_{\hat{z}|\psi^i}(\psi^i)(\hat{z}_i). \quad (4.9)$$

The bitstream for $\hat{z}$ constitutes a very small proportion, which can be nearly ignored. The probability of $\hat{y}$ is modeled by the Gaussian distribution, whose parameters μ, σ is predicted by the entropy parameter model leveraging context and hyperpriors:

Table 4.2 Summary of baseline lossy learned image compression methods

Paper	Publication	Framework	Contributions
Ballé et al. [165]	ICLR 2017	AE	End-to-end feasible framework with the factorized entropy model leveraging a piecewise linear function
Theis et al. [166]	ICLR 2017	AE	AE-based architecture with the Gaussian scale mixtures probability distribution
Ballé et al. [163]	ICLR 2018	VAE	Introducing an auxiliary hyperprior autoencoder to reduce spatial redundancy in latent representations and the parameterized Gaussian entropy model
Hu et al. [167]	AAAI 2020	VAE	Extending the entropy model to multi-layer hyperprior architecture to capture spatial correlation in a wider range
Minnen et al. [164]	NeurIPS 2018	Joint	Proposing the integration of a spatial context model with an autoregressive computation method
Lee et al. [168]	ICLR 2019	Joint	Presenting a context-adaptive entropy model framework that integrates two distinct types of contexts for the conditional probability distribution

Source Author

$$p_{\hat{y}|\hat{z}, C_m}(\hat{y}|\hat{z}, C_m) = \prod_i \mathcal{N}(\mu_i, \sigma_i^2)(\hat{y}_i) \quad (4.10)$$

$$\text{with } \mu, \sigma = h_e(\Phi_h; \Phi_c),$$

where C_m denotes the autoregressive context model and $\hat{z}$ denotes hyperpriors. The entropy parameter model h_e predicts distribution parameter utilizing information Φ_h, Φ_c provided by C_m and $\hat{z}$. This portion dominates the final compressed bit rates. Therefore, LIC can coarsely be viewed as a balance between the bit rate term $\mathbb{E}_{x \sim p_x} [-\log_2 p_{\hat{y}|\hat{z}, C_m}(\hat{y}|\hat{z}, C_m)]$ and distortion term $\mathbb{E}_{x \sim p_x} [D(x, \hat{x})]$. At a certain level of distortion, efficient nonlinear transformations extract compact latent representations and accurate entropy models help actual bit rates approximate the lower bound, the information entropy of the marginal distribution. Nonlinear transformations and entropy models are the most important components, which are enhanced by following research. In addition, the non-differentiability issues introduced by quantization brings challenges for optimization, which should be effectively addressed. In the next section, improvements at the module level will be elaborated in detail.

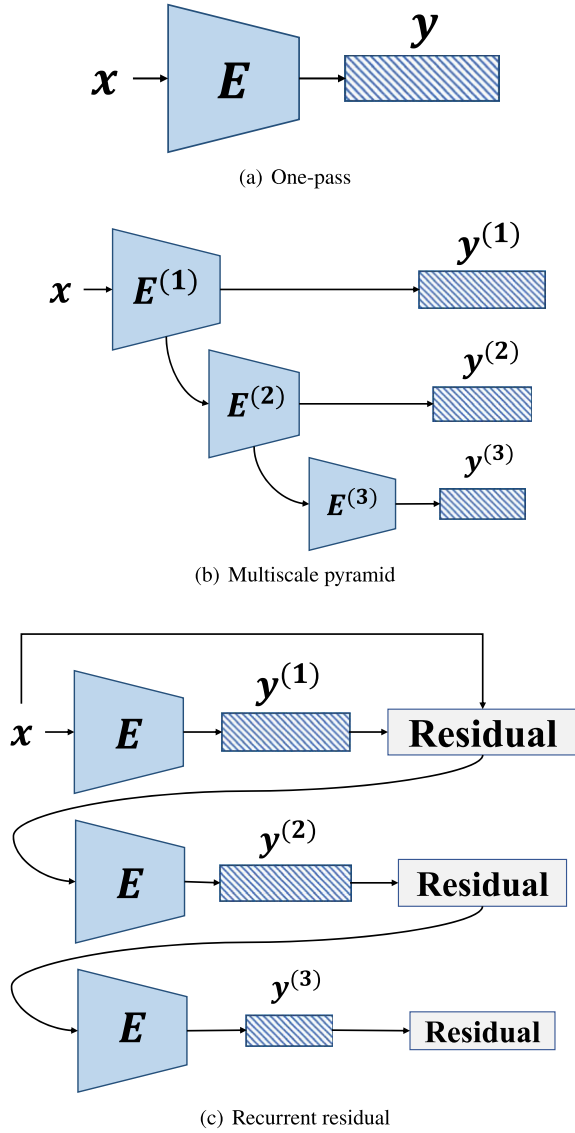
4.3.2 Core Components

Transformation, quantization, and entropy model are three main components in lossy image compression. In this part, some efforts on these three aspects are reviewed.

Transformation. Transformation converts image pixels to latent representations, and a good encoder is desired to extract more compact representations because the information entropy of latent representations is the upper bound of bit rates, which determines the upper limit of encoding performance. Many transformation optimization efforts endeavor to obtain the compact latent representations. The process of extracting latent representations can be roughly categorized into three types, as shown in the Fig. 4.6. One-pass methods [162–164] extract a unique bottleneck representation in one time, followed by entropy encoding. This structure is simple and efficient, and the majority of approaches utilize this structure, which will be elaborated on in detail later. Multiscale pyramid models [169, 170] extract representations at different scales using multiple networks. The difference lies in that Rippel et al. [169] integrated pyramidal representations of different scales into a single comprehensive layer through a fusion module, producing a single bitstream. In contrast, Nakanishi et al. [170] entropy encoded representations at different scales into separate bitstreams for transmission. This allows information to be transmitted between encoders and decoders across multiple scales, rather than solely transmitting the information from the bottleneck layer. The recurrent residual models [171–174] iteratively extract representations, typically employing an RNN network structure. They learn to encode residual information at each step, progressively improving the image reconstruction as more bits are used. The early work [171] introduced a framework for image compression using LSTMs with a focus on variable-rate compression. It was designed to address the need for flexibility in compression, allowing the same model to adapt to different compression levels without needing separate models for each target bit rate. Following that Toderici et al. [172] extended the method to handle full-resolution images and improved the model’s performance. Both methods presented by Toderici et al. emphasized the ability of RNNs to perform variable-rate compression through a progressive approach. Furthermore, [173, 174] integrated CNNs and RNNs to improve the RD performance. Johnston et al. [173] also used a spatial bit allocation algorithm. The summary of RNN-based methods can be viewed in the Table 4.3.

One-pass structure is simple and efficient to extract representations into a single bottleneck layer to be entropy encoded later. This facilitates the optimization of entropy models and is leveraged by subsequent works. Typically, the one-pass extraction employs a symmetrical autoencoder architecture where encoders and decoders both consist of multistage upsampling and downsampling transformations. We categorized these methods based on the network architecture, as shown in the Tables 4.4, and 4.5. Early methods [163–165] adopted a single CNN layer followed by the GDN normalization layer [175] within a transformation stage. Four stages of downsampling transformations are established in the analysis transformation to extract representations whose size is $\frac{1}{16} \times \frac{1}{16}$ of the original image. To further improve the

Fig. 4.6 Categories of transformation architectures.
Source Author



RD performance, subsequent works [176–179] leveraged residual blocks instead of plain CNN layers. In addition, the technique proposed by Cheng et al. [176] incorporated local attention modules into the framework, focusing more on regions with complex context. As a result, these areas receive a higher allocation of bits, leading to an overall improvement in rate-distortion performance. They extended the hyperprior method, originally used for lossy compression, to lossless image compression by predicting a probability model for raw images using a hyperprior and

Table 4.3 Summary of RNN-based transformation image compression methods

Title	Authors	Publication	Contributions
Towards conceptual compression [174]	Gregor et al.	NeurIPS 2016	Hierarchical compression with the first layer transmitting conceptual information and the second layer transmitting detailed information
Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks [173]	Johnston et al.	CVPR 2018	Combining CNNs and RNNs with a spatially adaptive bit allocation algorithm
Variable rate image compression with recurrent neural networks [171]	Toderici et al.	ICLR 2016	The first progressive RNNs-based compression model, reconstructing higher quality images with more bits consumed
Full resolution image compression with recurrent neural networks [172]	Toderici et al.	CVPR 2017	Exploring more efficient RNNs networks to improve the RD performance
Target-quality image compression with recurrent, convolutional neural networks [181]	Covell et al.	arxiv 2017	The RNN is trained using Stop Code Tolerant (SCT) and a two-pass methodology to achieve high-quality reconstruction, particularly in regions requiring fine detail

Source Author

the discretized Gaussian mixture likelihoods. Figure 4.7 illustrates the overall structure. Furthermore, Chen et al. [180] applied non-local attention operations to capture long-range correlation in latent representations, as shown in Fig. 4.8.

Recently, some works focus on the adaptivity of LIC models to enhance the diversity of transformations based on the complexity of the image content. Wang et al. [182] divided the bitstream into a neural syntax layer and a content layer. The decoder applied different transformation kernels for different images based on the transmitted statements. The introduced neural syntax captured abstract contextual information of the image/representation, aiding in projecting the encoded representation into a more compact subspace of transformation coefficients. Liu et al. [183] proposed a model with region-adaptive transformations guided by segmentation masks that generate convolutional kernels based on prototype-related masks.

With the development of Vision Transformers (ViTs), recent works build transformer-based transformation modules to achieve more larger receptive field in encodes and decoders. Lu et al. [196] introduced a transformer-based image compression approach, which is developed reusing the VAE architecture. The proposed

Table 4.4 Summary of CNN-based transformation image compression methods

Title	Authors	Publication	Contributions
Learned image compression with discretized Gaussian mixture likelihoods and attention modules [176]	Cheng et al.	CVPR 2020	Applying attention module into transformations for the first time
End-to-end optimized versatile image compression with wavelet-like transform [184]	Ma et al.	TPAMI 2020	Wavelet like transform, entropy coding and dequantization, supporting lossy and lossless coding.
Deep image compression via end-to-end learning [185]	Liu et al.	CVPRW 2018	Training the CNN model with a combination of perceptual and adversarial losses to enhance detail sharpness
Deep image compression using decoder side information [186]	Ayzik et al.	ECCV 2020	Side information consisting of synthetic images is utilized with DSC to achieve high-quality image results
End-to-end optimized ROI image compression [187]	Cai et al.	TIP 2020	A tiered reconstruction loss function that ensures precise pixel-wise consistency in the region of interest (ROI) while ensuring structural similarity across the entire image
Learning convolutional networks for content-weighted image compression [188]	Li et al.	CVPR 2018	Importance maps to enhance reconstruction quality of saliency image regions and training with MSE and SSIM balanced
Variable rate deep image compression with a conditional autoencoder [189]	Choi et al.	ICCV 2019	Employing Lagrange multipliers and the size of quantization bins to control parameter functions
Context-adaptive neural network-based prediction for image compression [190]	Dumas et al.	TIP 2019	CNN-based neural networks trained on undistorted contexts exhibit strong generalization capabilities in distorted contexts, even under significant distortions

(continued)

Table 4.4 (continued)

Title	Authors	Publication	Contributions
A flexible deep CNN framework for image restoration [191]	Jin et al.	TMM 2019	Exceeded JPEG and other denoising and restoration techniques
Wavelet-based deep auto encoder-decoder (wdaed)-based image compression [192]	Mishra et al.	TCSVT 2020	Using a single-level Haar wavelet to extract high-frequency components and improve artifact reduction techniques for extreme compression
Convolutional neural network-based block up-sampling for intra-frame coding [193]	Li et al.	TCSVT 2017	A dense neural network performs intra-prediction by analyzing adjacent image blocks
An end-to-end compression framework based on convolutional neural networks [194]	Jiang et al.	TCSVT 2017	A pair of CNNs working alongside JPEG codec to protect structural details while reducing blocking artifacts in images
Fully connected network-based intra prediction for image coding [195]	Li et al.	TIP 2018	Utilizing a block-wise CNN with downsampling or upsampling strategies for optimizing rate-distortion, where luma and chroma downsampling aids in lowering computational complexity

Source Author

method uses a novel approach to efficiently compress images by combining the strengths of both convolutional networks and transformers. At the core of this method is the Neural Transformation Unit (NTU), which is a building block used in both the main and hyper encoders to analyze and aggregate important information from the input image, making it more compact for compression. The Neural Transformation Unit (NTU) in the Transformer-based Image Compression (TIC) method combines Swin Transformer Blocks (STB) for capturing long-range dependencies and convolutional layers for embedding short-range spatial information. This allows the NTUs to effectively balance global and local feature representation in the compressed latent space. Used in both the main and hyper encoders, the NTUs analyze the input image, aggregate essential information, and produce a more compact representation, leading to improved compression efficiency while maintaining high-quality image reconstruction. The overall structure is shown in Fig. 4.9. Later on, Zhu et al. [197] also utilized the Swin Transformer Block in both the main network and the hyperprior network while the channel autoregressive model continues to employ a CNN architecture. Liu et al. [198] build a model framework utilizing both CNNs and

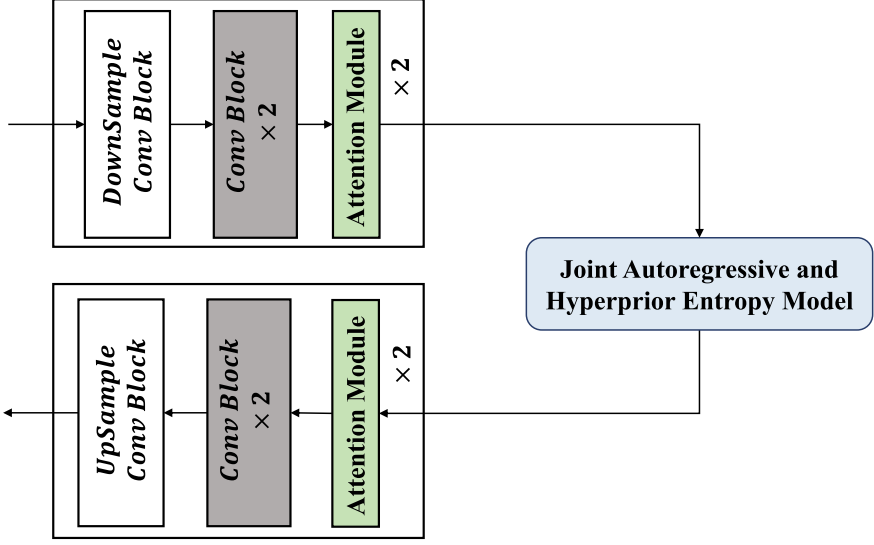


Fig. 4.7 Overall structure of the CNN model. $\times 2$ denotes stack of the two same modules. Public domain open access image [176]

Transformers, combining the ability of CNNs to capture local dependencies with the capability of Swin Transformers to capture global dependencies. Zou et al. [199] divided images into multiple windows, within which correlations are computed. This confirms their hypothesis that local attention is more suited for reconstructing local textures.

Quantization. Lossy compression is fundamentally a non-differentiable problem caused by the quantization component in the compression pipeline, which makes it challenging to train neural networks end-to-end. Most methods utilize uniform additive noise to simulate quantization noise during the training process. It was proposed by Ballé et al. [165] replacing the uniform noise of the quantizer with additive i.i.d. uniform noise, whose width is equal to that of the quantization bin. For example, when the quantization bin equals 1 each discrete probability $p_{\hat{y}_i}$ is equal to the probability mass of the corresponding continuous density p_{y_i} within the quantization bin as follows:

$$\begin{aligned}
 p_{\hat{y}_i}(\hat{y}_i | \hat{\mu}_i, \hat{\sigma}_i) &= \int_{\hat{y}_i - \frac{1}{2}}^{\hat{y}_i + \frac{1}{2}} p_{y_i}(y | \hat{\mu}_i, \hat{\sigma}_i) dy \\
 &= \int_{\hat{y}_i - \frac{1}{2}}^{\hat{y}_i + \frac{1}{2}} \mathcal{N}(y | \hat{\mu}_i, \hat{\sigma}_i) dy.
 \end{aligned} \tag{4.11}$$

Note that the integral term is equivalent to the result of convolving a normal distribution with a uniform distribution function, which corresponds to the probability

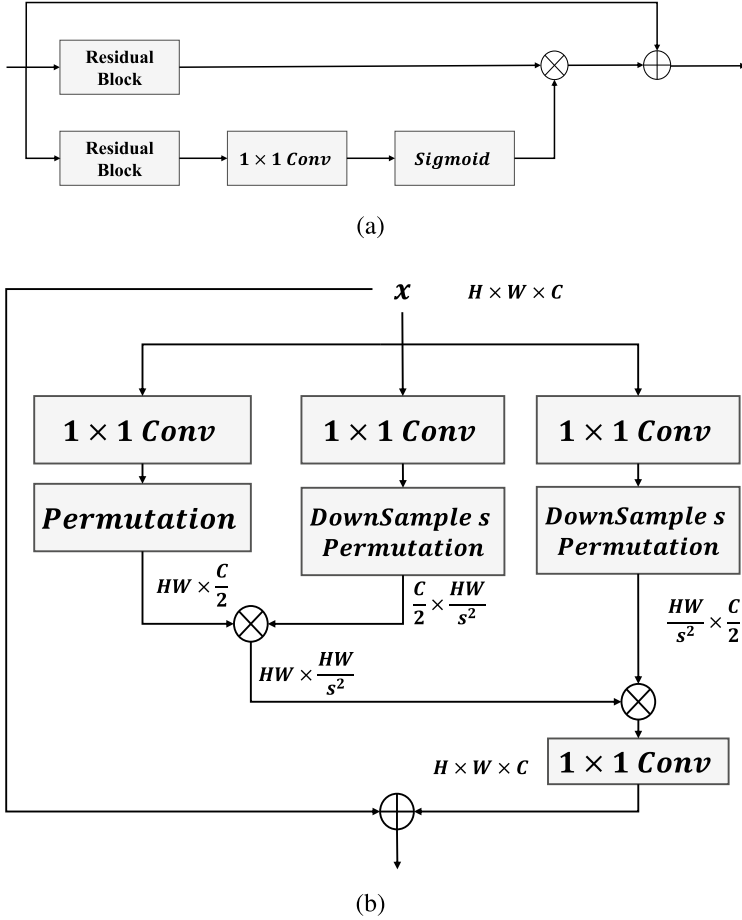


Fig. 4.8 The architecture of attention module in LIC. **a** Local attention module. Public domain open access [176]. **b** Non-local attention module. Public domain open access image [180]

distribution of $\tilde{y} = y + \Delta y$. Therefore the density $p_{\tilde{y}_i}$ provides a continuous function that can interpolate the discrete probability values at integer positions as follows:

$$\begin{aligned}
 p_{\tilde{y}_i}(\hat{y}_i | \hat{\mu}_i, \hat{\sigma}_i) &= \int_{\hat{y}_i - \frac{1}{2}}^{\hat{y}_i + \frac{1}{2}} \mathcal{N}(y | \hat{\mu}_i, \hat{\sigma}_i) dy \\
 &= (\mathcal{N}(\hat{\mu}_i, \hat{\sigma}_i) * \mathcal{U}\left(-\frac{1}{2}, \frac{1}{2}\right))(\hat{y}_i) \\
 &= p_{\tilde{y}_i}(\hat{y}_i | \hat{\mu}_i, \hat{\sigma}_i)
 \end{aligned} \tag{4.12}$$

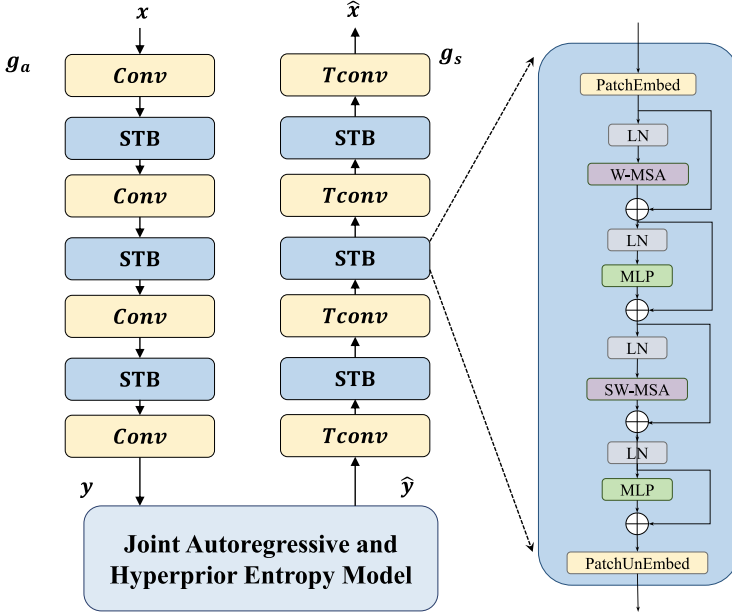


Fig. 4.9 Overall structure of the transformer-based model. Public domain open access image [196]

An alternative solution [171, 172, 188] adopted stochastic binary representation which can effectively address the gradient backpropagation problem as follow:

$$b(x) = x + \varepsilon \in \{-1, 1\}, \quad (4.13)$$

$$\text{with } \varepsilon \sim \begin{cases} 1 - x & \text{with probability } \frac{1+x}{2}, \\ -x - 1 & \text{with probability } \frac{1-x}{2}. \end{cases}$$

Theis et al. [166] further expanded this idea to the integer range as follows:

$$[x] \approx \lfloor x \rfloor + \varepsilon, \quad \varepsilon \in \{0, 1\}, \quad (4.14)$$

$$\text{with } P(\varepsilon = 1) = x - \lfloor x \rfloor,$$

where the derivative of this function is smooth as follows:

$$\frac{d}{dx}[x] := \frac{d}{dx}\mathbb{E}([x]) = \frac{d}{dx}x = 1. \quad (4.15)$$

Some methods [179, 200] combined additive uniform noise and stochastic rounding. Specifically, noisy representations were maintained for the entropy model learning and the rounded representations were fed into decoders. Agustsson et al. [201] pro-

posed a novel approach for acquiring compressible representations. This method depends on a smooth approximation of quantization and entropy modeling, and the discrete counterpart is gradually annealed during the training process. Their method was demonstrated in two demanding tasks, including both the image compression and the neural network model compression. Although these two tasks usually involve different methodologies, the outcomes of their soft-to-hard quantization method are on par with the state-of-the-art methods in both areas.

Entropy Model. The entropy model is the other most important module in learned image compression, which can provide an accurate distribution estimation to decrease bit rates dramatically. Many VAE-based works model the latent variables with single Gaussian distribution and Cheng et al. [176] extend the single Gaussian distribution to discretized Gaussian mixture distribution. The discretized Gaussian mixture likelihoods model can be formulated as:

$$p_{\hat{y}|\hat{z}}(\hat{y} | \hat{z}) = \prod_i p_{\hat{y}|\hat{z}}(\hat{y}_i | \hat{z}_i), \quad (4.16)$$

$$\begin{aligned} p_{\hat{y}|\hat{z}}(\hat{y}_i | \hat{z}_i) &= \left(\sum_{k=1}^K w_i^{(k)} \mathcal{N}(\mu_i^{(k)}, \sigma_i^{2(k)}) * \mathcal{U}\left(-\frac{1}{2}, \frac{1}{2}\right) \right)(\hat{y}_i) \\ &= c\left(\hat{y}_i + \frac{1}{2}\right) - c\left(\hat{y}_i - \frac{1}{2}\right), \end{aligned} \quad (4.17)$$

where i refers to the index in the feature maps, with $\hat{y}_i$ denoting the i -th element of $\hat{y}$ and μ_i the i -th element of μ . k stands for the mixture component index, and each mixture is defined by three parameters: weights $w_i^{(k)}$, means $\mu_i^{(k)}$ and variances $\sigma_i^{2(k)}$ for each element $\hat{y}_i$. The function $c(\cdot)$ is the cumulative distribution function. The range of $\hat{y}$ is learned automatically and is not predetermined. To ensure stability during training, the range of $\hat{y}$ is clipped to $[-255, 256]$, as $\hat{y}$ generally remains within this range. For the boundary cases, when $\hat{y} = -255$, $c(\hat{y}_i - \frac{1}{2})$ is set to zero (i.e., $c(-\infty) = 0$), and when $\hat{y} = 256$, $c(\hat{y}_i + \frac{1}{2})$ is set to one (i.e., $c(+\infty) = 1$). This approach ensures numerical stability during the training process.

Since Minnen et al. [164] proposed the autoregressive module in the entropy model, a large amount of following works improve the autoregressive method to further enhance the context entropy models. All previously decoded latent representations can be viewed as context and provide reference information for entropy models to better estimate the distribution parameters of the current decoding symbol. These methods improve the context models C_m for more accurate parameter estimation reference Φ_c as shown in Fig. 4.5. The objective is to utilize spatial and cross-channel correlations to reduce the distribution redundancy of latent representations. For examples, Ma et al. [202] highlighted the limitations of 2D and 3D convolutions, clarifying that channel dimension references refer to the forward channel features at the same spatial position during encoding. To fully exploit cross-channel correlations, the features are grouped along the channel dimension. Each group can utilize all previously encoded groups to give cross-channel contextual information.

Table 4.5 Summary of transformer-based transformation image compression methods

Title	Authors	Publication	Contributions
Transformer-based image compression [196]	Lu et al.	DCC 2022	Proposing a transformer-based image compression approach, which is built upon the VAE architecture
Transformer-based transform coding [197]	Zhu et al.	CVPR 2020	Utilizing the Swin Transformer Block in both the main network and the hyperprior network while the channel autoregressive model continues to employ a CNN architecture
Learned image compression with mixed transformer-cnn architectures [198]	Liu et al.	CVPR 2023	Building a model framework utilizing both CNNs and Transformers, combining the ability of CNNs to capture local dependencies with the capability of Swin Transformers to capture global dependencies
The devil is in the details: Window-based attention for image compression [199]	Zou et al.	CVPR 2022	Dividing images into multiple windows, within which correlations are computed

Source Author

When encoding the first group, each channel is encoded sequentially, leveraging the previously encoded channels from the first group at each step, with references taken from the top-k most relevant features. Chen et al. [203] used a $5 \times 5 \times 5$ 3D masked convolution to jointly leverage the spatial and cross-channel correlations. Qian et al. [204] computed the similarity map S and the confidence map U to obtain the most relevant encoded image patches, capturing global correlation. He et al. [178] analyzed the context model through a random mask approach, and introduced a new parallel context model featuring checkerboard masked convolutions, serving as an alternative to traditional serial context models. Through experiments, they demonstrated how the distance between reference latent variables and decoded latent variables impacts the bit savings in context modeling for image compression. Closer reference relationships exhibit stronger causality, allowing for greater mutual information during decoding. Additionally, this explains the reason for that modeling context with larger kernel masked convolutions or stacks of 5×5 convolutions (which have a larger receptive field) unexpectedly degrades RD performance. To further leverage correlation between channels, Minnen et al. [200] investigated a channel-conditioned (CC) model that partitions the latent tensor with the channel dimension to get the two equal-sized slices, conditioning the entropy parameters of the second slice based on

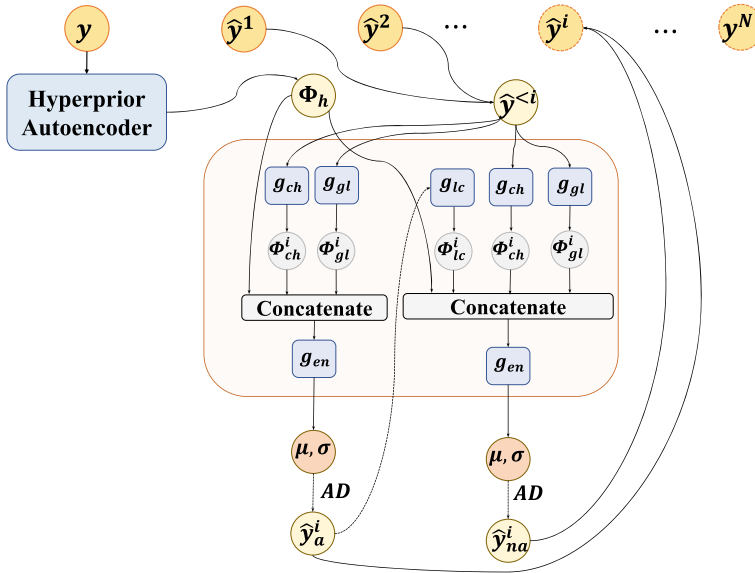


Fig. 4.10 Illustration of the context model with channel-wise and spatial division. Public domain open access image [179]

the first decoded slices. Furthermore, the predicted latent residuals are incorporated into the latent features incrementally for each slice. Furthermore, He et al. [177] implemented an uneven grouping scheme with N divisions as shown in Fig. 4.10, where finer granularity was assigned to the initial slices using fewer channels, and subsequent slices gradually became coarser with the use of more channels. Each group of slices employs a checkerboard pattern to capture spatial correlations. Guo et al. [205] divided channels into two groups. The latent variable $\hat{y}$ at spatial position n is divided into two groups across the channel dimension. The first group utilizes standard masked convolution, while the second group employs masked convolution that can leverage the already encoded latent features from the first group at the current position. By observation, it is found that the correlation distribution of the first half of the channels closely resembles that of the global correlation. This allows for the use of the correlation matrix, calculated from the previously decoded features of the first group, to index the reference positions when encoding the second group of features. Jiang et al. [179] introduced multiple reference information in the entropy model to obtain the state-of-the-art RD performance.

In recent years, many Transformer architectures have been utilized in entropy models to capture global spatial correlations. Qian et al. [206] followed the checkerboard pattern, where the autoregressive context model employed a ViT architecture and incorporates learnable diamond-patterned relative position encoding. The model selects the top- K relevant features as reference. Following that, Koyuncu et al. [207] introduced the context model based on transformer that offers context adaptability

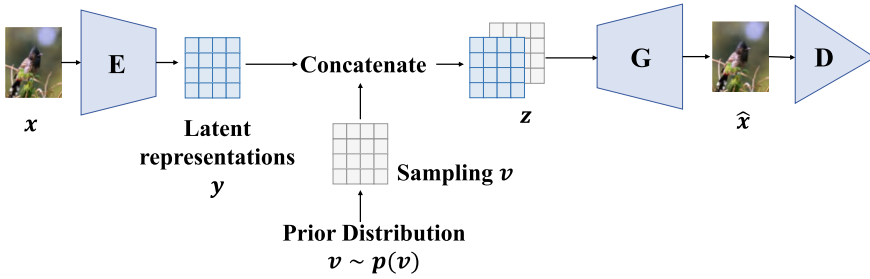


Fig. 4.11 Overall structure of the generative compression model. Public domain open access image [210]

and leverages distant spatial relationships within the latent representations. They not only utilized the attention in the spatial dimensions but also in the channel dimensions for entropy modeling. Building upon the Contextformer, a parallelization strategy Efficient Contextformer was implemented [209], utilizing both windowed and shifted-window attention mechanisms. In [208], the latent features were grouped along both spatial and channel dimensions, employing intra-group token mixers and inter-group token mixers to compute relevant information. These methods are summarized in Table 4.6.

4.4 Typical Methods of Generative Image Coding

Generative compression is an emerging paradigm in the field of data compression that utilizes generative models to facilitate efficient storage and transmission of data, especially for images and videos. Conventional compression techniques, like JPEG and H.264, rely on heuristic approaches to reduce file sizes while attempting to maintain acceptable visual quality. However, these methods often fall short when it comes to achieving high compression rates without sacrificing the perceptual quality of the reconstructed data. In contrast, generative compression aims to address these constraints by modeling the intrinsic distribution of the data being compressed. Instead of directly encoding the raw pixel values, a generative model is trained to understand and generate images from a latent representation. This latent space captures essential features of the data, allowing for a more compact representation that retains critical information. The main advantage of this approach is its ability to model complex distributions, enabling the generation of high-quality reconstructions even at low bit rates.

Generative compression leverages the generative capabilities of generative decoders, introducing randomness at the decoding stage to ensure the production of visually appealing images at extremely low bit rates. Agustsson et al. [210] first applied GAN to introduce stochasticity in LIC. The overall architecture is given in Fig. 4.11. Specifically, the encoder extracts latent representations from original input

Table 4.6 Summary of entropy models

Title	Authors	Publication	Dimension
Channel-wise autoregressive entropy models for learned image compression [200]	Minnen et al.	ICIP 2020	Channel
End-to-end learnt image compression via non-local attention optimization and improved context modeling [180]	Chen et al.	TIP 2021	Spatial + Channel
Checkerboard context model for efficient learned image compression [178]	He et al.	CVPR 2021	Spatial
Elic: efficient learned image compression with unevenly grouped space-channel contextual adaptive coding [177]	He et al.	CVPR 2022	Spatial + Channel
Causal contextual prediction for learned image compression [205]	Guo et al.	TCSVT 2021	Spatial + Channel
Learning accurate entropy model with global reference for image compression [204]	Qian et al.	ICLR 2021	Global spatial
Entroformer: a transformer-based entropy model for learned image compression [206]	Qian et al.	ICLR 2022	Global spatial
Contextformer: a transformer with spatio-channel attention for context modeling in learned image compression [207]	Koyuncu et al.	ECCV 2022	Global spatial + Channel
GroupedMixer: an entropy model with group-wise token-mixers for learned image compression [208]	Li et al.	TCSVT 2024	Global spatial+channel
Learned image compression with mixed transformer-CNN architectures [198]	Liu et al.	CVPR 2023	Global spatial+channel
MLIC: multi-reference entropy model for learned image compression [179]	Jiang et al.	ACMMM 2023	Global spatial+local spatial+channel

Source Author

Table 4.7 Summary of generative compression methods

Title	Authors	Publication	Architecture
Generative adversarial networks for extreme learned image compression [210]	Agustsson et al.	ICCV 2019	GAN/Conditional GAN
High-fidelity generative image compression [211]	Mentzer et al.	NeurIPS 2020	GAN
Deep generative models for distribution-preserving lossy compression [212]	Tschannen et al.	NeurIPS 2018	WGAN
Lossy image compression with conditional diffusion models [213]	Yang et al.	NeurIPS 2024	Conditional Diffusion
Lossy compression with Gaussian diffusion [214]	Theis et al.	arXiv 2022	Gaussian Diffusion

Source Author

images, concatenated with noise v which is sampled from a prior distribution $p(v)$. The generator reconstructs images utilizing latent representations z , followed by a discriminator. And the framework is trained for the objective as follows:

$$\mathcal{L} = \min_{E,G} \max_D \mathbb{E}[f(D(x))] + \mathbb{E}[g(D(G(z)))] + \lambda \mathbb{E}[d(x, G(z))] + \alpha H(\hat{y}), \quad (4.18)$$

where λ and α balance the distortion, GAN, and bit rate terms. However, GAN-based method proposed by [210] favors high-level semantics at extremely low bit rate but ignores image fidelity. Mentzer et al. [211] discarded randomly sampling in latent space to keep high fidelity of reconstructions. Tschannen et al. [212] proposed a generative compression framework that can smoothly interpolate in the latent representation space and utilize the Wasserstein distance to constrain the distribution of original input images and reconstructions. The method generates realistic content sampled from the same distribution of the input image at low bit rates and reconstructs details at high bit rates. All previously mentioned methods build compression frameworks by GAN. Recently diffusion-based generative compression methods have also been proposed [213, 214]. We claim that with the powerful generative capabilities of models, generating compressed transmissions at low bit rates is likely to become a key focus in the future (Table 4.7).

4.5 Summary

This chapter presents a comprehensive categorical review of image compression techniques based on advanced deep learning solutions. Deep image coding techniques are examined in terms of both lossy and lossless coding, with particular focus on lossy

coding methods. Among the lossy coding methods, classical AE and VAE networks serve as the basic framework for most learned image compression, supporting the development of architectures such as CNN, RNN, and Transformer. Regarding the most important modules in lossy image compression, significant research efforts have been devoted to improving the non-linear transformation and entropy models. A comprehensive review of these methods is provided.

Exercises

1. Please describe what is deep learning-based image compression.
2. Please describe the advantages of deep learning-based image compression methods compared with traditional image compression methods.
3. Please explain what is lossless compression and lossy compression.
4. Please explain what is an end-to-end image compression structure.
5. Please explain how can an end-to-end image compression framework be categorized.
6. Please explain what are the differences between AE and VAE based compression methods.
7. Please describe how can Human Visual System (HVS) be used in deep learning-based image compression.
8. Please describe what is Learned Image Compression (LIC), and how does it differ from traditional image compression methods.
9. Please describe the three primary frameworks of lossy image compression, and their characteristics.
10. Please describe the challenges quantization face in learned image compression, and the existing methods to address these challenges.

References

1. Y. Wu, L. Xie, S. Sun, W. Gao, Y. Yan, Adaptive intra period size for deep learning-based screen content video coding, in *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2024)
2. H. Zheng, W. Gao, End-to-end rgb-d image compression via exploiting channel-modality redundancy, *Proc. AAAI Conf. Artif. Intell.* **38**(7), 7562–7570 (2024)
3. L. Tao, W. Gao, G. Li, C. Zhang, Adanic: towards practical neural image compression via dynamic transform routing, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 16 879–16 888
4. Y. Wu, W. Gao, End-to-end lossless compression of high precision depth maps guided by pseudo-residual (2022). [arXiv:2201.03195](https://arxiv.org/abs/2201.03195)
5. Y. Wu, Z. Qi, H. Zheng, L. Tao, W. Gao, Deep image compression with latent optimization and piece-wise quantization approximation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 1926–1930

6. W. Gao, L. Tao, L. Zhou, D. Yang, X. Zhang, Z. Guo, Low-rate image compression with super-resolution learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2020), pp. 154–155
7. W. Gao, S. Sun, H. Zheng, Y. Wu, H. Ye, Y. Zhang, Opendmc: an open-source library and performance evaluation for deep-learning-based multi-frame compression, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9685–9688
8. W. Gao, H. Zheng, C. Zhang, K. Zheng, Z. Yu, Y. Li, H. Ye, Y. Zhang, Opendic: an open-source library and performance evaluation for deep-learning-based image compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 202–11 205
9. Y. Li, W. Gao, G. Li, S. Ma, Saliency segmentation oriented deep image compression with novel bit allocation. *IEEE Trans. Image Process.* **34**, 16–29 (2025)
10. Z. Zhang, H. Wang, Z. Chen, S. Liu, Learned lossless image compression based on bit plane slicing, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 27 579–27 588
11. L. Xie, W. Gao, H. Zheng, G. Li, Spcgc: scalable point cloud geometry compression for machine vision, in *2024 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2024), pp. 17 272–17 278
12. L. Xie, W. Gao, H. Zheng, H. Ye, Semantic-aware visual decomposition for point cloud geometry compression, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 595–595
13. Z. Qi, W. Gao, Variable-rate point cloud geometry compression based on feature adjustment and interpolation, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 63–72
14. Z. Yu, W. Gao, When dynamic neural network meets point cloud compression: computation-aware variable rate and checkerboard context, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 600–600
15. L. Xie, W. Gao, S. Fan, Z. Yao, Pdnet: parallel dual-branch network for point cloud geometry compression and analysis, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 596–596
16. L. Xie, W. Gao, H. Zheng, End-to-end point cloud geometry compression and analysis with sparse tensor, in *International Workshop on Advances in Point Cloud Compression, Processing and Analysis* (2022), pp. 27–32
17. C. Fu, G. Li, R. Song, W. Gao, S. Liu, Octattention: octree-based large-scale contexts model for point cloud compression. *Proc. AAAI Conf. Artif. Intell.* **36**(1), 625–633 (2022)
18. C. Zhang, W. Gao, Adapcc: adaptive rate control and rate-distortion-complexity optimization for dynamic point cloud compression, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
19. K. Wang, W. Gao, Unipcgc: towards practical point cloud geometry compression via an efficient unified approach, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
20. K. Zheng, W. Gao, H. Zheng, ctree-based learned point cloud geometry compression: a lossy perspective, in *2025 Data Compression Conference (DCC)* (IEEE, 2025)
21. Z. Qi, H. Liu, W. Gao, Variable-rate point cloud compression with channel-wise entropy modeling, in *2025 Data Compression Conference (DCC)* (IEEE, 2025)
22. H. Zheng, W. Gao, Z. Yu, T. Zhao, G. Li, Viewpcgc: view-guided learned point cloud geometry compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 7152–7161
23. L. Xie, W. Gao, H. Zheng, G. Li, Roi-guided point cloud geometry compression towards human and machine vision, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024)
24. W. Gao, G. Li, H. Yuan, R. Hamzaoui, Z. Li, S. Liu, Apcpa’22: 1st international workshop on advances in point cloud compression, processing and analysis, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 7392–7393
25. W. Gao, G. Li, Point cloud compression, enhancement and applications: from 3d perception to large models, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 292–11 293

26. H. Sun, G. Li, S. Chen, L. Li, W. Gao, Special topic on 3d point cloud processing and applications. *ZTE Commun.* **21**(4), 1 (2023)
27. W. Gao, W. Gao, X. Mu, C. Peng, G. Li, Overview and comparison of avs point cloud compression standard. *APSIPA Trans. Signal Inf. Process.* (2024)
28. W. Gao, S. Wang, X. Zhang, S. Kwong, Editorial for special issue on deep learning-based data compression. *APSIPA Trans. Signal Inf. Process.* **13**(6) (2024)
29. W. Gao, H. Ye, G. Li, H. Zheng, Y. Wu, L. Xie, OpenPointCloud: an open-source algorithm library of deep learning based point cloud compression, in *ACM International Conference on Multimedia* (2022), pp. 7347–7350
30. Y. Zhang, W. Gao, G. Li, Openpointcloud-v2: a deep learning based open-source algorithm library of point cloud processing, in *Proceedings of the 1st International Workshop on Advances in Point Cloud Compression, Processing and Analysis* (2022), pp. 51–55
31. L. Xie, W. Gao, Learningpcc: a pytorch library for learning-based point cloud compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024)
32. L. Xie, W. Gao, Pchmvision: an open-source library of point cloud compression for human and machine vision, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024)
33. Z. Yang, W. Gao, G. Li, Y. Yan, Sur-driven video coding rate control for jointly optimizing perceptual quality and buffer control. *IEEE Trans. Image Process.* (2023)
34. F. Shen, Z. Cai, W. Gao, An efficient rate control algorithm for intra frame coding in avs3, in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2021), pp. 3164–3169
35. H. Yuan, W. Gao, J. Wang, Dynamic computational resource allocation for fast inter frame coding in video conferencing applications, in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6
36. W. Gao, Q. Jiang, R. Wang, S. Ma, G. Li, S. Kwong, Consistent quality oriented rate control in hevc via balancing intra and inter frame coding. *IEEE Trans. Ind. Inform.* **18**(3), 1594–1604 (2021)
37. H. Yuan, W. Gao, A new coding unit partitioning mode for screen content video coding, in *Proceedings of the 2021 5th International Conference on Digital Signal Processing* (2021), pp. 66–72
38. W. Gao, On the performance evaluation of state-of-the-art rate control algorithms for practical video coding and transmission systems, in *Proceedings of the 2020 4th International Conference on Video and Image Processing* (2020), pp. 179–185
39. W. Gao, S. Kwong, Q. Jiang, C.-K. Fong, P.H. Wong, W.Y. Yuen, Data-driven rate control for rate-distortion optimization in hevc based on simplified effective initial qp learning. *IEEE Trans. Broadcast.* **65**(1), 94–108 (2018)
40. W. Gao, A multi-objective optimization perspective for joint consideration of video coding quality, in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (IEEE, 2019), pp. 986–991
41. W. Gao, S. Kwong, Y. Jia, Joint machine learning and game theory for rate control in high efficiency video coding. *IEEE Trans. Image Process.* **26**(12), 6074–6089 (2017)
42. W. Gao, S. Kwong, Y. Zhou, H. Yuan, Ssim-based game theory approach for rate-distortion optimized intra frame ctu-level bit allocation. *IEEE Trans. Multimedia* **18**(6), 988–999 (2016)
43. W. Gao, S. Kwong, H. Yuan, X. Wang, Dct coefficient distribution modeling and quality dependency analysis based frame-level bit allocation for hevc. *IEEE Trans. Circuits Syst. Video Technol.* **26**(1), 139–153 (2015)
44. W. Gao, S. Kwong, Phase congruency based edge saliency detection and rate control for perceptual image and video coding, in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2016), pp. 000 264–000 269
45. H. Yuan, W. Gao, Openfastvc: an open source library for video coding fast algorithm implementation, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9660–9663

46. H. Yuan, W. Gao, S. Ma, Y. Yan, Divide-and-conquer-based rdo-free cu partitioning for 8k video compression. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(4), 1–20 (2024)
47. L. Tao, W. Gao, A hardware implementation of entropy encoder for 8k video coding, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
48. Y. Guo, W. Gao, S. Ma, G. Li, Accelerating transform algorithm implementation for efficient intra coding of 8k uhd videos, *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* **18**(4), 1–20 (2022)
49. Z. Cai, W. Gao, Efficient fast algorithm and parallel hardware architecture for intra prediction of avs3, in *IEEE International Symposium on Circuits and Systems (ISCAS)* (IEEE, 2021), pp. 1–5
50. W. Gao, H. Yuan, Y. Guo, L. Tao, Z. Cai, G. Li, Openhardwarevc: an open source library for 8k uhd video coding hardware implementation, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 7339–7342
51. W. Gao, H. Yuan, G. Liao, Z. Guo, J. Chen, Pp8k: a new dataset for 8k uhd video compression and processing. *IEEE MultiMedia* (2023)
52. W. Liu, W. Gao, G. Li, S. Ma, T. Zhao, H. Yuan, Enlarged motion-aware and frequency-aware network for compressed video artifact reduction. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
53. G. Li, G. Wei, W. Gao, *Point Cloud Compression: Technologies and Standardization* (Springer Nature, 2024)
54. G. Li, W. Gao, W. Gao, Introduction, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 1–28
55. G. Li, W. Gao, W. Gao, Background knowledge, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 29–51
56. G. Li, W. Gao, W. Gao, Predictive coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 53–70
57. G. Li, W. Gao, W. Gao, Transform coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 71–96
58. G. Li, W. Gao, W. Gao, Quantization techniques, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 97–112
59. G. Li, W. Gao, W. Gao, Entropy coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 113–133
60. G. Li, W. Gao, W. Gao, Mpeg geometry-based point cloud compression (g-pcc) standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 135–165
61. G. Li, W. Gao, W. Gao, Avs point cloud compression standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 167–197
62. G. Li, W. Gao, W. Gao, Mpeg video-based point cloud compression (v-pcc) standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 199–218
63. G. Li, W. Gao, W. Gao, Mpeg ai-based 3d graphics coding standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 219–241
64. G. Li, W. Gao, W. Gao, Future work, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 243–250
65. W. Liu, W. Gao, X. Mu, Fast inter-frame motion prediction for compressed dynamic point cloud attribute enhancement. *Proc. AAAI Conf. Artif. Intell.* **38**(4), 3720–3728 (2024)
66. Z. Yang, W. Gao, X. Lu, Danet: density-adaptive network for geometry-based point cloud compression artifacts removal, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
67. X. Fan, G. Li, D. Li, Y. Ren, W. Gao, T.H. Li, Deep geometry post-processing for decompressed point clouds, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
68. X. Zhang, G. Liao, W. Gao, G. Li, Tdrnet: transformer-based dual-branch restoration network for geometry based point cloud compression artifacts, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6

69. Z. Li, G. Li, T.H. Li, S. Liu, W. Gao, Semantic point cloud upsampling. *IEEE Trans. Multimedia* (2022)
70. Z. Li, S. Liu, W. Gao, G. Li, G. Li, S4r: rethinking point cloud sampling via guiding upsampling-aware perception. *IEEE Trans. Multimedia* (2025)
71. R. Zhang, W. Gao, G. Li, T.H. Li, Qinet: decision surface learning and adversarial enhancement for quasi-immune completion of diverse corrupted point clouds. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–14 (2022)
72. R. Bao, Y. Ren, G. Li, W. Gao, S. Liu, Flow-based point cloud completion network with adversarial refinement, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2559–2563
73. J. Chen, G. Li, R. Zhang, T.H. Li, W. Gao, Pointvae: invertible variational autoencoder framework for 3d point cloud generation, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 3216–3220
74. R. Zhang, J. Chen, W. Gao, G. Li, T.H. Li, Pointot: interpretable geometry-inspired point cloud generative model via optimal transport. *IEEE Trans. Circuits Syst. Video Technol.* **32**(10), 6792–6806 (2022)
75. S. Fan, W. Gao, G. Li, Salient object detection for point clouds, in *European Conference on Computer Vision* (Springer, 2022), pp. 1–19
76. S. Luo, W. Gao, A general framework for rotation invariant point cloud analysis, in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2024), pp. 3665–3669
77. X. Lu, W. Gao, Attentivenet: Detecting small objects for lidar point clouds by attending to important points, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
78. Z. Pan, N. Zhang, W. Gao, S. Liu, G. Li, Less is more: label recommendation for weakly supervised point cloud semantic segmentation. *Proc. AAAI Conf. Artif. Intell.* **38**(5), 4397–4405 (2024)
79. Z. Pan, G. Liu, W. Gao, T.H. Li, Epcontrast: effective point-level contrastive learning for large-scale point cloud understanding, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024), pp. 1–6
80. N. Zhang, Z. Pan, T.H. Li, W. Gao, G. Li, Improving graph representation for point cloud segmentation via attentive filtering, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 1244–1254
81. K. Wen, N. Zhang, G. Li, W. Gao, Mpvnn: multi-resolution point-voxel non-parametric network for 3d point cloud processing, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024), pp. 1–6
82. Z. Pan, N. Zhang, W. Gao, S. Liu, G. Li, Point cloud semantic segmentation with sparse and inhomogeneous annotations, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
83. R. Zhang, G. Li, W. Gao, S. Liu, A quantum-inspired framework in master-servant mode for large-scale multi-modal place recognition. *IEEE Trans. Intell. Transp. Syst.* (2024)
84. D. Yang, W. Gao, Pointchd: a point cloud benchmark for congenital heart disease classification and segmentation. *IEEE J. Biomed. Health Inform.* (2024)
85. S. Fan, W. Gao, G. Li, Point-mpp: Point cloud self-supervised learning from masked position prediction. *IEEE Trans. Neural Netw. Learn. Syst.* (2024)
86. Z. Pan, W. Gao, G. Li, Distribution guidance network for weakly supervised point cloud semantic segmentation, in *Advances in Neural Information Processing Systems* (2024)
87. R. Zhang, G. Li, W. Gao, T.H. Li, Compoint: can complex-valued representation benefit point cloud place recognition? *IEEE Trans. Intell. Transp. Syst.* (2024)
88. L. Xie, X. Mu, G. Li, W. Gao et al., Pku-dpcc: a new dataset for dynamic point cloud compression. *APSIPA Trans. Signal Inf. Process.* **13**(6) (2024)
89. H. Yuan, W. Gao, W. Gao, Opensep: an open source subjective experiment platform, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 248–11 251

90. S. Fan, W. Gao, Screen-based 3d subjective experiment software, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9672–9675
91. J. Wang, W. Gao, G. Li, Zoom to perceive better: No-reference point cloud quality assessment via exploring effective multiscale feature. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
92. J. Wang, W. Gao, G. Li, Applying collaborative adversarial learning to blind point cloud quality measurement. *IEEE Trans. Instrum. Meas.* (2023)
93. H. Li, W. Gao, Information exploration of projected views for point cloud quality measurement. *IEEE Trans. Instrum. Meas.* (2024)
94. B. Qu, X. Liang, S. Sun, W. Gao, Exploring aigc video quality: a focus on visual harmony, video-text consistency and domain distribution gap, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2024)
95. B. Qu, H. Li, W. Gao, Bringing textual prompt to ai-generated image quality assessment, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024)
96. S. Sun, X. Liang, S. Fan, W. Gao, W. Gao, Ve-bench: subjective-aligned benchmark suite for text-driven video editing quality assessment, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
97. Z. Guo, W. Gao, H. Wang, J. Wang, S. Fan, No-reference deep quality assessment of compressed light field images, in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6
98. X. Zang, W. Gao, G. Li, H. Fang, C. Ban, Z. He, H. Sun, A baseline investigation: Transformer-based cross-view baseline for text-based person search, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 7737–7746
99. G. Liao, W. Gao, G. Li, J. Wang, S. Kwong, Cross-collaborative fusion-encoder network for robust rgb-thermal salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **32**(11), 7646–7661 (2022)
100. W. Gao, G. Liao, S. Ma, G. Li, Y. Liang, W. Lin, Unified information fusion network for multimodal rgb-d and rgb-t salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **32**(4), 2091–2106 (2021)
101. Y. Li, Z. Yu, X. Chen, S. Wang, H. Yuan, W. Gao, Zero-shot quantization for large-kernels via shape-based distribution and diversity self-distillation, in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2025)
102. X. Chen, Y. Wang, Y. Li, X. Ling, M. Li, R. Liu, M. Ouyang, K. Zhao, T. Guan, Y. He, Low bit-width zero-shot quantization with soft feature-infused hints for iot systems. *IEEE Internet Things J.* (2024)
103. Y. Chen, S. Sun, G. Li, W. Gao, T.H. Li, Closing the gap between theory and practice during alternating optimization for gans. *IEEE Trans. Neural Netw. Learn. Syst.* (2023)
104. Y. Chen, C. Jin, G. Li, T.H. Li, W. Gao, Mitigating label noise in gans via enhanced spectral normalization. *IEEE Trans. Circuits Syst. Video Technol.* (2023)
105. X. Zang, G. Li, W. Gao, Multidirection and multiscale pyramid in transformer for video-based pedestrian retrieval. *IEEE Trans. Ind. Inform.* **18**(12), 8776–8785 (2022)
106. X. Zang, G. Li, W. Gao, X. Shu, Learning to disentangle scenes for person re-identification. *Image Vis. Comput.* **116**, 104330 (2021)
107. X. Zang, G. Li, W. Gao, X. Shu, Exploiting robust unsupervised video person re-identification. *IET Image Process.* **16**(3), 729–741 (2022)
108. Z. Yue, G. Li, W. Gao, Cross-level guided attention for human-object interaction detection, in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2023), pp. 284–289
109. S. Sun, J. Liu, H. Li, G. Liu, T.H. Li, W. Gao, Streamflow: Streamlined multi-frame optical flow estimation for video sequences, in *Advances in Neural Information Processing Systems* (2024)
110. Z. Yao, W. Gao, Iterative saliency aggregation and assignment network for efficient salient object detection in optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* (2024)
111. G. Liao, W. Gao, Rethinking feature mining for light field salient object detection. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(10), 1–24 (2024)

112. B. Yang, W. Gao, G. Li, Focus and adjust: Progressive refinement network for human object interaction detection, in *2022 26th International Conference on Pattern Recognition (ICPR)* (IEEE, 2022), pp. 2546–2552
113. G. Liao, W. Gao, Q. Jiang, R. Wang, G. Li, Mmnet: multi-stage and multi-scale fusion network for rgb-d salient object detection, in *Proceedings of the 28th ACM International Conference on Multimedia* (2020), pp. 2436–2444
114. W. Gao, S. Kwong, Y. Zhou, Y. Jia, J. Zhang, W. Wu, Multiscale phase congruency analysis for image edge visual saliency detection, in *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 1 (IEEE, 2016), pp. 75–80
115. W. Gao, S. Kwong, Y. Zhou, X. Wang, Phase congruency analysis of down-sampled and blurring images for foreground extraction, in *2015 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)* (IEEE, 2015), pp. 152–157
116. Y. Guo, W. Gao, G. Li, Interpretable task-inspired adaptive filter pruning for neural networks under multiple constraints. *Int. J. Comput. Vis.* 1–17 (2024)
117. W. Gao, Y. Guo, S. Ma, G. Li, S. Kwong, Efficient neural network compression inspired by compressive sensing. *IEEE Trans. Neural Netw. Learn. Syst.* (2022)
118. Y. Guo, W. Gao, Semantic-driven automatic filter pruning for neural networks, in *IEEE international conference on multimedia and expo (ICME)* (IEEE, 2022), pp. 1–6
119. L. Tao, W. Gao, Efficient channel pruning based on architecture alignment and probability model bypassing, in *IEEE international conference on systems, man, and cybernetics (SMC)* (IEEE, 2021), pp. 3232–3237
120. Z. Li, G. Li, T. Li, S. Liu, W. Gao, Information-growth attention network for image super-resolution, in *Proceedings of the 29th ACM International Conference on Multimedia* (2021), pp. 544–552
121. L. Zhou, W. Gao, G. Li, H. Yuan, T. Zhao, G. Yue, Disentangled feature distillation for light field super-resolution with degradations, in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2023), pp. 116–121
122. L. Zhou, W. Gao, G. Li, End-to-end spatial-angular light field super-resolution using parallax structure preservation strategy, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 3396–3400
123. W. Gao, L. Zhou, L. Tao, A fast view synthesis implementation method for light field applications. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* **17**(4), 1–20 (2021)
124. X. Zhang, W. Gao, G. Li, Q. Jiang, R. Cong, Image quality assessment-driven reinforcement learning for mixed distorted image restoration. *ACM Trans. Multimedia Comput. Commun. Appl.* **19**(1s), 1–23 (2023)
125. X. Zhang, W. Gao, H. Yuan, G. Li, Je 2 net: joint exploitation and exploration in reinforcement learning based image restoration, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2090–2094
126. X. Zhang, W. Gao, Hirl: hybrid image restoration based on hierarchical deep reinforcement learning via two-step analysis, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2445–2449
127. S. Wang, Y. Lu, W. Xia, P. Xia, Z. Wang, W. Gao, Light field angular super-resolution by view-specific queries. *Visual Comput.* 1–16 (2024)
128. Y. Sun, Z. Li, S. Wang, W. Gao, Depth-assisted calibration on learning-based factorization for a compressive light field display. *Opt. Express* **31**(4), 5399–5413 (2023)
129. Y. Sun, Z. Li, L. Li, S. Wang, W. Gao, Optimization of compressive light field display in dual-guided learning, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2075–2079
130. W. Gao, S. Fan, G. Li, W. Lin, A thorough benchmark and a new model for light field saliency detection. *IEEE Trans. Pattern Anal. Mach. Intell.* (2023)
131. W. Gao, G. Li, *Deep Learning for 3D Point Clouds* (Springer Nature, 2024)
132. W. Gao, G. Li, Introduction to 3d point clouds: Datasets and perception, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 1–27

133. W. Gao, G. Li, Learning basics for 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 29–70
134. W. Gao, G. Li, Deep-learning-based point cloud enhancement i, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 71–97
135. W. Gao, G. Li, Deep-learning-based point cloud enhancement ii, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 99–130
136. W. Gao, G. Li, Deep-learning-based point cloud analysis i, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 131–162
137. W. Gao, G. Li, Deep-learning-based point cloud analysis ii, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 163–193
138. W. Gao, G. Li, Point cloud pre-trained models and large models, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 195–225
139. W. Gao, G. Li, Point cloud-language multi-modal learning, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 227–254
140. W. Gao, G. Li, Open-source projects for 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 255–272
141. W. Gao, G. Li, Typical engineering applications of 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2025), pp. 273–299
142. W. Gao, G. Li, Future work on deep learning-based point cloud technologies, in *Deep Learning for 3D Point Clouds* (Springer, 2025), pp. 301–315
143. G.K. Wallace, The jpeg still picture compression standard. *Commun. ACM* **34**(4), 30–44 (1991)
144. A. Skodras, C. Christopoulos, T. Ebrahimi, The jpeg 2000 still image compression standard. *IEEE Signal Process. Mag.* **18**(5), 36–58 (2001)
145. B. Bross, J. Chen, S. Liu, Versatile video coding (draft 5). JVET (2019)
146. Z. Wang, E.P. Simoncelli, A.C. Bovik, Multiscale structural similarity for image quality assessment, in *The Thrity-Seventh Asilomar Conference on Signals, Systems and Computers, 2003*, vol. 2 (IEEE, 2003), pp. 1398–1402
147. C.E. Shannon, A mathematical theory of communication. *Bell Syst. Tech. J.* **27**(3), 379–423 (1948)
148. A. Van Den Oord, N. Kalchbrenner, K. Kavukcuoglu, Pixel recurrent neural networks, in *International Conference on Machine Learning* (2016), pp. 1747–1756
149. T. Salimans, A. Karpathy, X. Chen, D.P. Kingma, Pixelcnn++: a pixelcnn implementation with discretized logistic mixture, in *International Conference on Learning Representations* (2017)
150. S. Reed, A. Oord, N. Kalchbrenner, S.G. Colmenarejo, Z. Wang, Y. Chen, D. Belov, N. Freitas, Parallel multiscale autoregressive density estimation, in *International Conference on Machine Learning* (2017), pp. 2912–2921
151. F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, L.V. Gool, Practical full resolution learned lossless image compression, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 10 629–10 638
152. L. Xie, M. Hu, X. Bai, Towards hardware-friendly and robust facial landmark detection method, in *International Conference on Neural Information Processing* (Springer, 2022), pp. 432–444
153. L. Xie, M. Hu, X. Bai, Online Improved Vehicle Tracking Accuracy via Unsupervised Route Generation, in *IEEE 34th International Conference on Tools with Artificial Intelligence* (IEEE, 2022), pp. 788–792
154. J. Townsend, T. Bird, D. Barber, Practical lossless compression with latent variables using bits back coding, in *International Conference on Learning Representations* (2019)
155. F. Kingma, P. Abbeel, J. Ho, Bit-swap: recursive bits-back coding for lossless compression with hierarchical latent variables, in *International Conference on Machine Learning* (2019), pp. 3408–3417
156. E. Hoogeboom, J. Peters, R. Van Den Berg, M. Welling, Integer discrete flows and lossless compression. *Adv. Neural Inf. Process. Syst.* **32**, 12 134–12 144 (2019)

157. R. van den Berg, A.A. Gritsenko, M. Dehghani, C.K. Sønderby, T. Salimans, Idf++: analyzing and improving integer discrete flows for lossless compression, in *International Conference on Learning Representations* (2020)
158. S. Zhang, C. Zhang, N. Kang, Z. Li, ivpf: numerical invertible volume preserving flow for efficient lossless compression, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 620–629
159. J. Ho, E. Lohm, P. Abbeel, Compression with flows via local bits-back coding. *Adv. Neural Inf. Process. Syst.* **32**, 3874–3883 (2019)
160. S. Zhang, N. Kang, T. Ryder, Z. Li, iflow: numerically invertible flows for efficient lossless compression via a uniform coder. *Adv. Neural Inf. Process. Syst.* **34**, 5822–5833 (2021)
161. N. Kang, S. Qiu, S. Zhang, Z. Li, S.-T. Xia, Pilc: practical image lossless compression with an end-to-end gpu oriented neural framework, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 3739–3748
162. J. Ballé, V. Laparra, E.P. Simoncelli, End-to-end optimized image compression, in *5th International Conference on Learning Representations, ICLR 2017* (2017)
163. J. Ballé, D. Minnen, S. Singh, S.J. Hwang, N. Johnston, Variational image compression with a scale hyperprior (2018). [arXiv:1802.01436](https://arxiv.org/abs/1802.01436)
164. D. Minnen, J. Ballé, G. Toderici, Joint autoregressive and hierarchical priors for learned image compression (2018). [arXiv:1809.02736](https://arxiv.org/abs/1809.02736)
165. J. Ballé, V. Laparra, E.P. Simoncelli, End-to-end optimized image compression (2016). [arXiv:1611.01704](https://arxiv.org/abs/1611.01704)
166. L. Theis, W. Shi, A. Cunningham, F. Huszár, Lossy image compression with compressive autoencoders (2017). [arXiv:1703.00395](https://arxiv.org/abs/1703.00395)
167. Y. Hu, W. Yang, J. Liu, Coarse-to-fine hyper-prior modeling for learned image compression. *Proc. AAAI Conf. Artif. Intell.* **34**(07), 11 013–11 020 (2020)
168. J. Lee, S. Cho, S.-K. Beack, Context-adaptive entropy model for end-to-end optimized image compression (2018). [arXiv:1809.10452](https://arxiv.org/abs/1809.10452)
169. O. Rippel, L. Bourdev, Real-time adaptive image compression, in *International Conference on Machine Learning* (2017), pp. 2922–2930
170. K.M. Nakanishi, S.-I. Maeda, T. Miyato, D. Okanohara, Neural multi-scale image compression, in *Asian Conference on Computer Vision* (2019), pp. 718–732
171. G. Toderici, S.M. O'Malley, S. J. Hwang, D. Vincent, D. Minnen, S. Baluja, M. Covell, R. Sukthankar, Variable rate image compression with recurrent neural networks (2015). [arXiv:1511.06085](https://arxiv.org/abs/1511.06085)
172. G. Toderici, D. Vincent, N. Johnston, S. Jin Hwang, D. Minnen, J. Shor, M. Covell, Full resolution image compression with recurrent neural networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 5306–5314
173. N. Johnston, D. Vincent, D. Minnen, M. Covell, S. Singh, T. Chinen, S.J. Hwang, J. Shor, G. Toderici, Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 4385–4393
174. K. Gregor, F. Besse, D. Jimenez Rezende, I. Danihelka, D. Wierstra, Towards conceptual compression. *Adv. Neural Inf. Process. Syst.* **29**, 3549–3557 (2016)
175. J. Ballé, V. Laparra, E.P. Simoncelli, Density modeling of images using a generalized normalization transformation, in *International Conference on Learning Representations* (2016)
176. Z. Cheng, H. Sun, M. Takeuchi, J. Katto, Learned lossless image compression with a hyperprior and discretized gaussian mixture likelihoods (2020). [arXiv:2002.01657](https://arxiv.org/abs/2002.01657)
177. D. He, Z. Yang, W. Peng, R. Ma, H. Qin, Y. Wang, Elic: efficient learned image compression with unevenly grouped space-channel contextual adaptive coding, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 5718–5727
178. D. He, Y. Zheng, B. Sun, Y. Wang, H. Qin, Checkerboard context model for efficient learned image compression, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 14 771–14 780

179. W. Jiang, J. Yang, Y. Zhai, P. Ning, F. Gao, R. Wang, Mlic: multi-reference entropy model for learned image compression (2022). [arXiv:2211.07273](https://arxiv.org/abs/2211.07273)
180. T. Chen, H. Liu, Z. Ma, Q. Shen, X. Cao, Y. Wang, Neural image compression via non-local attention optimization and improved context modeling (2019). [arXiv:1910.06244](https://arxiv.org/abs/1910.06244)
181. M. Covell, N. Johnston, D. Minnen, S.J. Hwang, J. Shor, S. Singh, D. Vincent, G. Toderici, Target-quality image compression with recurrent, convolutional neural networks (2017). [arXiv:1705.06687](https://arxiv.org/abs/1705.06687)
182. D. Wang, W. Yang, Y. Hu, J. Liu, Neural data-dependent transform for learned image compression, in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition* (2022), pp. 17 379–17 388
183. Y. Liu, W. Yang, H. Bai, Y. Wei, Y. Zhao, Region-adaptive transform with segmentation prior for image compression (2024). [arXiv:2403.00628](https://arxiv.org/abs/2403.00628)
184. H. Ma, D. Liu, N. Yan, H. Li, F. Wu, End-to-end optimized versatile image compression with wavelet-like transform. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(3), 1247–1263 (2020)
185. H. Liu, T. Chen, Q. Shen, T. Yue, Z. Ma, Deep image compression via end-to-end learning, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2018), pp. 2575–2578
186. S. Ayzik, S. Avidan, Deep image compression using decoder side information, in *European Conference on Computer Vision* (2020), pp. 699–714
187. C. Cai, L. Chen, X. Zhang, Z. Gao, End-to-end optimized roi image compression. *IEEE Trans. Image Process.* **29**, 3442–3457 (2020)
188. M. Li, W. Zuo, S. Gu, D. Zhao, D. Zhang, Learning convolutional networks for content-weighted image compression, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 3214–3223
189. Y. Choi, M. El-Khamy, J. Lee, Variable rate deep image compression with a conditional autoencoder, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 3146–3154
190. T. Dumas, A. Roumy, C. Guillemot, Context-adaptive neural network-based prediction for image compression. *IEEE Trans. Image Process.* **29**, 679–693 (2019)
191. Z. Jin, M.Z. Iqbal, D. Bobkov, W. Zou, X. Li, E. Steinbach, A flexible deep cnn framework for image restoration. *IEEE Trans. Multimedia* **22**(4), 1055–1068 (2019)
192. D. Mishra, S.K. Singh, R.K. Singh, Wavelet-based deep auto encoder-decoder (wdaed)-based image compression. *IEEE Trans. Circuits Syst. Video Technol.* **31**(4), 1452–1462 (2020)
193. Y. Li, D. Liu, H. Li, L. Li, F. Wu, H. Zhang, H. Yang, Convolutional neural network-based block up-sampling for intra frame coding. *IEEE Trans. Circuits Syst. Video Technol.* **28**(9), 2316–2330 (2017)
194. F. Jiang, W. Tao, S. Liu, J. Ren, X. Guo, D. Zhao, An end-to-end compression framework based on convolutional neural networks. *IEEE Trans. Circuits Syst. Video Technol.* **28**(10), 3007–3018 (2017)
195. J. Li, B. Li, J. Xu, R. Xiong, W. Gao, Fully connected network-based intra prediction for image coding. *IEEE Trans. Image Process.* **27**(7), 3236–3247 (2018)
196. M. Lu, P. Guo, H. Shi, C. Cao, Z. Ma, Transformer-based image compression (2021). [arXiv:2111.06707](https://arxiv.org/abs/2111.06707)
197. Y. Zhu, Y. Yang, T. Cohen, Transformer-based transform coding, in *International Conference on Learning Representations* (2022)
198. J. Liu, H. Sun, J. Katto, Learned image compression with mixed transformer-cnn architectures, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 14 388–14 397
199. R. Zou, C. Song, Z. Zhang, The devil is in the details: Window-based attention for image compression, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 17 492–17 501
200. D. Minnen, S. Singh, Channel-wise autoregressive entropy models for learned image compression, in *IEEE International Conference on Image Processing* (2020), pp. 3339–3343

201. E. Agustsson, F. Mentzer, M. Tschannen, L. Cavigelli, R. Timofte, L. Benini, L.V. Gool, Soft-to-hard vector quantization for end-to-end learning compressible representations. *Adv. Neural Inf. Process. Syst.* **30**, 1141–1151 (2017)
202. C. Ma, Z. Wang, R. Liao, Y. Ye, A cross channel context model for latents in deep image compression (2021). [arXiv:2103.02884](https://arxiv.org/abs/2103.02884)
203. Z. Chen, T. He, Learning based facial image compression with semantic fidelity metric. *Neurocomputing* **338**, 16–25 (2019)
204. Y. Qian, Z. Tan, X. Sun, M. Lin, D. Li, Z. Sun, L. Hao, R. Jin, Learning accurate entropy model with global reference for image compression, in *International Conference on Learning Representations* (2022)
205. Z. Guo, Z. Zhang, R. Feng, Z. Chen, Causal contextual prediction for learned image compression. *IEEE Trans. Circuits Syst. Video Technol.* **32**(4), 2329–2341 (2021)
206. Y. Qian, M. Lin, X. Sun, Z. Tan, R. Jin, Entroformer: a transformer-based entropy model for learned image compression (2022). [arXiv:2202.05492](https://arxiv.org/abs/2202.05492)
207. A.B. Koyuncu, H. Gao, A. Boev, G. Gaikov, E. Alshina, E. Steinbach, Contextformer: a transformer with spatio-channel attention for context modeling in learned image compression, in *European Conference on Computer Vision* (2022), pp. 447–463
208. D. Li, Y. Bai, K. Wang, J. Jiang, X. Liu, W. Gao, Groupedmixer: an entropy model with group-wise token-mixers for learned image compression. *IEEE Trans. Circuits Syst. Video Technol.* **34**(10), 9606–9619 (2024)
209. A.B. Koyuncu, P. Jia, A. Boev, E. Alshina, E. Steinbach, Efficient contextformer: Spatio-channel window attention for fast context modeling in learned image compression. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
210. E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, L. Van Gool, Generative adversarial networks for extreme learned image compression (2018). [arXiv:1804.02958](https://arxiv.org/abs/1804.02958)
211. F. Mentzer, G.D. Toderici, M. Tschannen, E. Agustsson, High-fidelity generative image compression. *Adv. Neural Inf. Process. Syst.* **33**, 11 913–11 924 (2020)
212. M. Tschannen, E. Agustsson, M. Lucic, Deep generative models for distribution-preserving lossy compression. *Adv. Neural Inf. Process. Syst.* **31**, 5933–5944 (2018)
213. R. Yang, S. Mandt, Lossy image compression with conditional diffusion models. *Adv. Neural Inf. Process. Syst.* **36** (2024)
214. L. Theis, T. Salimans, M.D. Hoffman, F. Mentzer, Lossy compression with gaussian diffusion (2022). [arXiv:2206.08889](https://arxiv.org/abs/2206.08889)

Chapter 5

Deep Learning-based Video Coding



5.1 Introduction

Efficient encoding of video data is critical for reducing storage demands and transmission overhead [1–54]. Over the years, a myriad of video compression techniques have been developed, with the goal of striking a balance between minimizing file sizes and preserving perceptual video quality. The advent of deep learning has revolutionized the field of video coding, providing novel solutions that utilize neural networks to enhance both compression efficiency and quality.

The traditional video compression standards, such as H.264 and H.265 (High-Efficiency Video Coding, HEVC), have been the cornerstone of video coding for decades [55–93]. These standards employ various techniques like motion estimation, discrete cosine transform (DCT), and quantization to reduce video data size [94–105]. While effective, they have limitations in achieving substantial compression gains and maintaining high visual quality [54, 106–114], especially at low bit rates.

The advent of deep learning in video coding has introduced a paradigm shift in the field. Deep learning approaches utilize neural networks to exploit spatial and temporal correlations within video frames, allowing for better compression and higher quality reconstruction. The use of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) in video coding has enabled innovative solutions for video compression. The development of deep learning-based video compression methods can be traced back to the 2010s. With the emergence of powerful GPUs and the development of deep learning technology [115–127, 127–161], researchers began to explore the potential of deep learning for video coding. Over the years, the field has witnessed a proliferation of research papers, with constant improvements in compression efficiency and video quality. Deep learning techniques have evolved in tandem with advances in neural network architectures, including the rise of deep residual networks and attention mechanisms. Deep learning-based video compression methods can be categorized based on their core techniques, such as intra-frame coding or inter-frame coding, architecture choices (e.g., CNN, RNN,

VAE), and the extent of end-to-end compression. These categories help researchers and practitioners navigate the diverse landscape of deep learning video compression techniques.

In conclusion, the adoption of end-to-end deep learning video compression has led to remarkable advancements, enabling more efficient video storage, transmission, and streaming without compromising quality. The subsequent sections of this chapter will delve deeper into specific techniques and innovations within this exciting field.

5.2 Framework and Key Components

5.2.1 Overall Framework

Deep learning-based video compression algorithms utilize a sophisticated framework, primarily focusing on efficient data reduction while preserving video quality [162], as shown in Fig. 5.1. The first stage, feature extraction, involves using deep neural networks to analyze and extract significant features from video frames. This initial stage is where deep learning algorithms shine. Feature extraction involves analyzing video frames to identify and extract significant patterns, textures, and elements. Typically, this is achieved using deep learning models like Convolutional Neural Networks (CNNs). Note that these deep learning models are trained to recognize various features in the video, such as edges, colors, and motion patterns, which are crucial for understanding the content and context of the video. The goal here is to transform the raw video data into a more abstract, yet informative representation that captures the essence of the video content with less data. This abstract representation forms the basis for the compression process, as it reduces the redundancy in the video data, making it easier to compress.

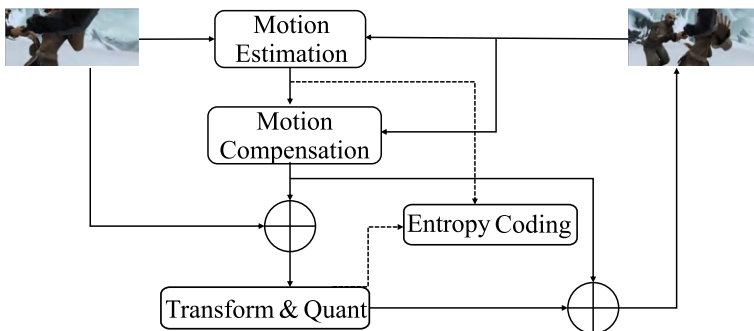


Fig. 5.1 The deep learning-based video compression architecture. Public domain open access image [162]

Following feature extraction, the algorithm proceeds to motion estimation. This phase is particularly important in video compression, as it deals with the temporal dynamics of video. Motion estimation seeks to understand how objects in the video move and change over time. Deep learning models, particularly those trained for optical flow estimation, can accurately predict motion vectors that describe the movement of objects between consecutive frames. This step is crucial for efficiently compressing video sequences, as it identifies which parts of the frame are static and which are dynamic. By understanding these motion patterns, the algorithm can focus on compressing the dynamic parts more effectively, as they typically contain more new information than static parts.

Building upon the motion estimates, motion compensation is the process of predicting future frames based on the current frame and the estimated motion vectors. This step is about creating a prediction for each frame that will come next, using the information about how objects are moving. The better the prediction, the less new information needs to be encoded for the next frame, leading to higher compression efficiency. Deep learning models can generate highly accurate predictions, which means that the differences between the actual frame and the predicted frame (often referred to as the “residual”) are minimized.

After predicting future frames, the next step is to quantize the video data. Quantization involves reducing the precision of the video data, which in turn reduces the amount of data that needs to be encoded. This is a delicate balance because overly aggressive quantization can lead to significant loss of quality, while too little quantization can result in insufficient compression. Deep learning-based models can be particularly useful in determining the optimal level of quantization for different parts of the video, based on the content’s complexity and importance to overall video quality [163, 164].

Finally, the compressed data is processed through entropy coding, which is about encoding the quantized data into a bitstream. The goal is to use as fewer bits as possible to represent data. Deep learning can optimize this process by identifying and exploiting patterns in quantized data that traditional methods might miss. By applying deep learning to entropy coding, the algorithm can achieve higher compression ratios without sacrificing video quality.

In essence, deep learning-based video compression algorithms encompass a cohesive framework that integrates feature extraction, motion estimation and compensation, quantization, and entropy coding. Each component is enhanced by the capabilities of deep learning, leading to efficient and high-quality video compression.

5.2.2 Motion Estimation

In the realm of deep learning-based video compression algorithms, the motion estimation module plays a pivotal role in enhancing the efficiency and accuracy of compression. Fundamentally, motion estimation is a process designed to determine

the movement of objects or frames between successive video frames. The core functionality of the motion estimation module in deep learning-based video compression algorithms is to predict and encode motion vectors. These vectors represent the displacement of pixels between consecutive frames, thereby facilitating the reduction of temporal redundancy. Utilizing convolutional neural networks (CNNs), the motion estimation process can effectively capture intricate motion patterns and complex textures, which traditional block-based methods might overlook. One of the key advantages of deep learning-based motion estimation is its ability to adapt to diverse video content through learned representations. Unlike traditional methods that rely on fixed heuristics, deep learning models can be trained on extensive datasets, enabling them to generalize well across various motion scenarios and video characteristics. This results in more precise motion predictions and consequently, more effective compression.

Currently, the most prevalent deep-learning-based motion-estimation methods center around the concept of optical flow. Optical flow represents the movement trajectory of individual pixels between adjacent frames. Historically, calculating optical flow was treated as an energy—minimization problem. In the application of optical flow techniques, three fundamental constraints are usually adhered to: the assumption of constant brightness in neighboring frames, the presumption of either temporal continuity or minimal motion variation between frames, and the expectation of consistent motion within a given sub-image area.

In recent developments, methods rooted in deep learning have emerged as the predominant choice for estimating optical flow. This shift is largely attributed to the time efficiency and enhanced performance characteristics inherent in these methods. Networks designed for deep learning-based optical flow estimation typically encompass three integral components: a feature extractor that identifies and processes key visual elements, a correlation encoder that maps and encodes the relationships between these elements, and an optical flow decoder that interprets and reconstructs the motion trajectories.

Furthermore, when delving into the practical application of these algorithms, we observe that deep learning-based optical flow estimation can generally be categorized into two distinct paradigms. The first, known as the coarse-to-fine approach, involves processing the motion estimation in stages, starting from a broader, less detailed perspective and progressively refining the analysis. The second paradigm, iterative refinement, operates on a repetitive adjustment process, where the motion estimation is continuously refined through successive iterations. These paradigms represent the underlying frameworks guiding the application and development of optical flow estimation in the realm of deep learning. The ensuing sections will delve deeper into the specific components of optical flow estimation networks and elaborate on these two strategic pipeline approaches.

The encoder component of optical flow estimation networks, based on deep learning, is designed to map the original RGB images into latent feature representations. This aspect of the encoder does not significantly differ from encoders used in other visual tasks. Previous works have commonly employed Convolutional Neural Networks (CNNs) as encoders, referencing studies such as [166–173]. These structures

are often based on foundational architectures like ResNet [174] or DenseNet [175]. Recently, due to the tremendous success of Transformers, many researchers have begun utilizing Transformer-based encoders in optical flow estimation, as evidenced in studies [176–178].

Through the feature extraction process of the encoder, the original RGB images are projected into a higher-dimensional space, which facilitates more efficient subsequent training and estimation. This enhancement in feature representation is crucial for capturing the nuanced motion patterns and dynamic changes between frames, which are essential for accurate optical flow estimation. The encoder's ability to effectively abstract and encode these patterns significantly contributes to the overall performance of the optical flow estimation network, particularly in complex and dynamically varied visual environments.

After extracting features from successive frames, optical flow estimation networks, whether they follow a coarse-to-fine strategy or iterative refinement, often engage in similarity computation to enhance the accuracy of flow estimation. This similarity calculation is derived from the dot product of feature matrices from the reference frame and the target frame (or the reference frame and the warped version of the reference frame) [179], namely,

$$C^k(i, j, m, n) = \frac{1}{2^{2k}} \sum_u^{2^k} \sum_v^{2^k} \langle F_1(i, j), F_2(2^k m + u, 2^k n + v) \rangle, \quad (5.1)$$

where the derived $C^k(i, j, m, n)$ could be viewed as the average over the correlation in the local $2^k \times 2^k$ window. k refers to the k th correlation level. u and v are the pixel motion in the horizontal and vertical directions, respectively. $\langle, \rangle$ denotes the inner product function. In summary, $C^k(i, j, m, n)$ means the correlation response of F_1 and F_2 pooled with a $2^k \times 2^k$ kernel.

The essence of this process lies in measuring the degree of alignment or correlation between the features of consecutive frames. This approach, often akin to cosine similarity, is instrumental in determining the extent to which two frames share common features, thereby facilitating the estimation of motion vectors. The use of cosine similarity, which compares the angle between two feature vectors, helps in capturing the underlying motion patterns with a high degree of precision. This method is particularly effective in scenarios where the frames exhibit subtle changes or when dealing with complex motion dynamics. The computed similarity score thus becomes a critical factor in accurately predicting the direction and magnitude of pixel displacement between frames, which is central to the task of optical flow estimation.

In the decoder segment of optical flow networks, features of the reference frame are typically combined with the correlation vector obtained from the previous stage and input into convolutional layers or Transformer layers to ultimately decode the optical flow. In fact, during the decoding process, existing network architectures often define an initial optical flow and continuously maintain it. The output of the

decoder is usually the residual of the final optical flow in comparison to this initial flow [179], namely,

$$\mathbf{t}, \Delta \mathbf{f}_{i+1} = H(\mathbf{t}, \mathbf{o}, \mathbf{f}_i), \quad (5.2)$$

$$\mathbf{f}_{i+1} = \Delta \mathbf{f}_{i+1} + \mathbf{f}_i \quad (5.3)$$

where H is the flow updater, $\mathbf{o}$ is the correlation vector, and $\mathbf{t}$ is the extracted context motion feature from the reference frame.

By initializing an optical flow and iteratively refining it through the decoding process, these networks effectively capture both the global motion trends and subtle local movements. The use of residuals, representing the difference between the estimated flow and the initial guess, allows for more precise adjustments, enhancing the accuracy of the flow estimation. This method is especially beneficial in complex scenarios where the motion patterns are intricate or rapidly changing. The integration of convolutional or Transformer layers facilitates the extraction and processing of spatial and temporal features, further refining the flow estimation. The overall architecture [165] is illustrated in Fig. 5.2.

Besides, a key distinction between the Coarse-to-Fine approach and Iterative Refinement lies here [169], as shown in Fig. 5.3. Iterative Refinement employs a loop for updating the flow at this stage, whereas Coarse-to-Fine estimates the flow only once. This distinction highlights a fundamental difference in handling optical flow estimation. Iterative Refinement's loop-based update mechanism allows for incremental improvements in flow estimation, adapting more effectively to the complexities and nuances of the motion depicted in the images. In contrast, the Coarse-to-Fine method takes a singular estimation approach, which, while potentially less computationally intensive, may not capture the finer details of motion with the same precision. Such analysis demonstrates how method choice should align with distinct application objectives, particularly when balancing motion granularity against system efficiency.

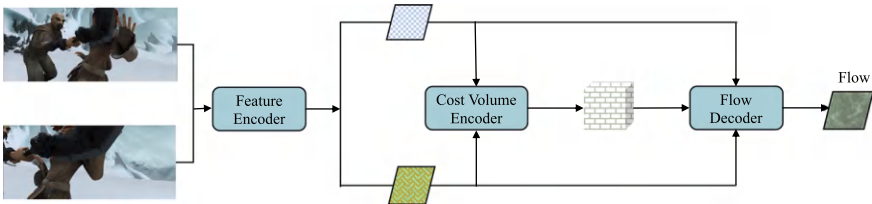


Fig. 5.2 Details of the motion estimation pipeline. Public domain open access image [165]

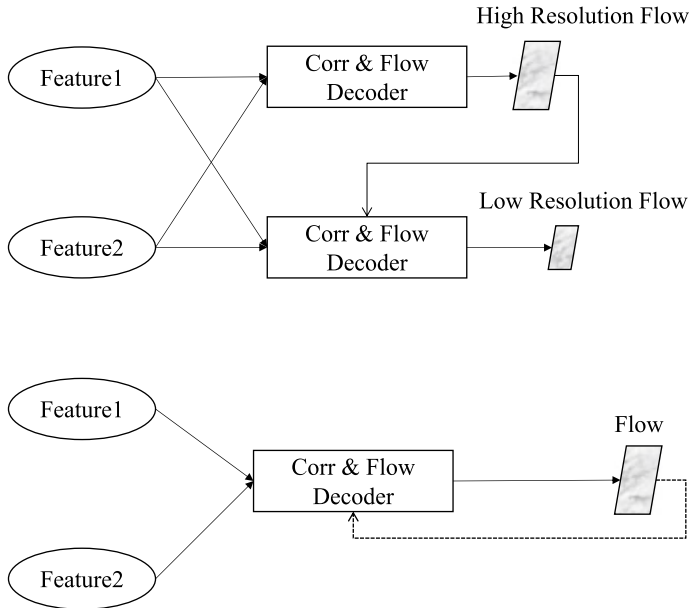


Fig. 5.3 Comparison between the coarse-to-fine and iterative refinement scheme. Public domain open access image [169]

5.2.3 Motion Compensation

Motion Compensation (MC) in video compression is a sophisticated technique designed to reduce the temporal redundancy prevalent in sequential video frames, thereby enhancing compression efficiency. This method hinges on the concept of predicting the contents of a video frame based on previously encoded frames. At its core, MC is grounded in the observation that changes between consecutive frames in a video are predominantly due to the movement of objects or the camera.

By predicting these changes, MC facilitates the encoding of the differences, or the residuals from the predicted content and the actual content, rather than encoding the entire frame, thereby achieving higher compression ratios. The process begins with motion estimation, where the algorithm identifies the movement of objects across frames using block-matching algorithms, dividing a frame into a grid of non-overlapping blocks and searching for the best match for each block in a reference frame. The output is a set of motion vectors indicating the displacement of objects. Subsequently, these motion vectors are used to generate a predictive frame by displacing blocks in the reference frame, effectively creating an estimation of the current frame. The next step involves calculating the residual frame - the difference between the actual current frame and the predictive frame, which holds the information that wasn't predictable by MC.

This residual, typically less complex than the original frame, is then encoded and transmitted along with the motion vectors, allowing for the reconstruction of the original frame at the decoder side. Implementing effective MC involves addressing challenges such as optimizing search algorithms for motion estimation, balancing accuracy and computational efficiency, mitigating error propagation through advanced error resilience techniques, adapting to varying motion complexities using variable block-size motion compensation, and achieving higher precision through sub-pixel motion estimation. MC is a vital element in modern video compression technologies, leveraging temporal redundancies to significantly reduce the data required to represent a video, thus playing a crucial role in efficient video storage and transmission.

5.2.4 *Transform and Quantization*

The purpose of a transform is to convert signals into the frequency domain, thereby facilitating subsequent processes like quantization and entropy coding. Operating in the frequency domain aligns with the characteristics of human vision, offering superior perceptual quality compared to direct operations in the pixel domain. In traditional video coding, the Discrete Cosine Transform (DCT) is universally adopted. In deep learning-based video coding, non-linear transform provided by CNNs is preferred. Thanks to the non-linear representation capabilities, original signals can be more effectively converted into forms that are easier to compress via entropy coding, with minimal distortion. Furthermore, the parameters of the transform modules can be jointly trained with those of the entropy modules, leading to improved rate-distortion performance. Quantization is the next step of the transform. Floating-point numbers are challenging to compress via entropy coding; hence, quantization is necessary to convert these signals into discrete symbols. Quantization is categorized into uniform and adaptive quantization; the former processes signals indiscriminately, while the latter discriminately quantizes signals to better preserve important ones. It is important to note that quantization prevents gradients from backpropagating through neural networks. Researchers suggest adding random noise to replace quantization during training, reverting to regular quantization during testing. This minor discrepancy indeed results in a domain shift.

5.2.5 *Entropy Coding*

Entropy coding is designed to exploit information redundancy among received symbols to achieve compression. The more this redundancy can be exploited, the more efficient the entropy coding becomes. During the entropy coding process, there is no loss of information and therefore entropy coding is lossless. The entropy of received symbols can be mathematically described as follows [180]:

$$H(X) = -\sigma_{i=1}^n P(x_i) \log_b P(x_i), \quad (5.4)$$

where X is the received symbols, n is the number of received symbols, $P(x_i)$ is the probability of X being equal to x_i , and b is the base of the logarithm used, typically 2 for binary systems. The entropy measures the difficulty of coding received symbols. Based on (5.4), we can infer that the closer the predicted probability aligns with the actual probability of the received symbols, the lower the entropy. In the area of deep learning-based entropy coding, the predicted probability is given by neural networks.

Entropy coding compresses discrete symbols following quantization. To facilitate gradient back-propagation, Ballé et al. [181] proposed to mimic the quantization process by adding random noise during training. This technique was first proposed for image compression and has since been broadly applied in video compression. To enhance information redundancy exploitation, Ballé et al. [182] further proposed improving the prediction of received symbols by using smaller hyper priors. Introducing hyper priors helps capture the spatial dependencies among adjacent symbols, leading to improved compression performance. Autoregressive entropy coding [183], which predicts symbols sequentially, leverages context information more effectively, but has a high computational cost. The channel-wise autoregressive coding [184] is a relatively compromise choice. The channel-wise autoregressive processes channels individually, capturing dependencies between channels without spatial sequence processing. These techniques are first proposed for the image process but can be adopted in video compression. In deep learning-based video compression, exploiting inter-frame dependencies enhances prediction accuracy for symbols to be encoded. Liu et al. [185] proposed to directly use adjacent frames to make a better prediction during the entropy coding of the current frame, which proves inter-frame contents provide valuable context for entropy coding without the need for explicit mechanisms. Overall, these entropy coding methods significantly advance the efficiency of deep learning-based video compression [6, 15, 23].

5.3 Typical Methods of Video Coding

This section presents a comprehensive exploration of state-of-the-art video compression algorithms rooted in deep learning over recent years. As depicted in Fig. 5.4, these algorithms are methodically arranged in chronological sequence to provide temporal context for their evolution. Building upon this temporal framework, the methods are classified along four critical dimensions: encoding manner, temporal scheme, optimization objective, and direction.

The term “encoding manner” pertains to the foundational principles governing the encoding process in these algorithms, and we distinguish two primary categories: encoding residuals and encoding conditions. This classification sheds light on the core strategies employed in representing video data for compression purposes.

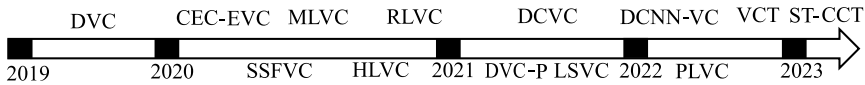


Fig. 5.4 Chronological order of various learning-based video compression method. *Source* Author

The “Temporal scheme” aspect delves into the temporal relationships incorporated in these algorithms. Specifically, it investigates whether the compression methodology involves the utilization of a dual-frame approach or extends to the compression of multiple frames. This temporal scheme delineates the intricacies of how these algorithms leverage the temporal aspect of video sequences.

Moving on to the “optimization objective,” we scrutinize the nature of constraints imposed during the training process. This dimension elucidates whether the algorithms rely solely on objective metrics for optimization or integrate perceptual factors derived from human vision, thereby reflecting a more nuanced and human-centric approach to video compression [8, 9, 22].

Lastly, the dimension of “direction” elucidates whether the algorithms leverage unidirectional or bidirectional temporal information. Unidirectional information implies a reliance on a single time sequence, while bidirectional information involves a more intricate integration of temporal relationships from both past and future frames.

This detailed examination seeks to offer an in-depth perspective on the complex landscape of deep learning-based video compression algorithms, offering insights into their encoding strategies, temporal considerations, optimization objectives, and the utilization of temporal information. The analysis is shown in Table 5.1.

DVC. DVC [186] presents a novel approach to video compression using deep learning techniques. This work proposes an integrated system for video compression, unifying motion processing, residual data encoding, and bitrate management within an end-to-end pipeline. The DVC framework operates as an end-to-end deep learning model, allowing for joint optimization of the entire compression process rather than treating each component in isolation. It utilizes an optical flow convolution network to predict inter-frame motion, which helps to accurately align reference frames with target frames to achieve motion estimation and compensation. The framework compresses the residuals (the differences between predicted and actual frames) using neural networks, optimizing for bitrate savings. To make the quantization step differentiable, the authors introduce uniform noise during training, enabling effective gradient-based optimization. A convolutional neural network (CNN) is employed to estimate the probability distributions of the latent representations, enhancing the accuracy of the entropy coding process. The DVC framework demonstrates impressive performance compared to traditional video compression methods, showcasing the effectiveness of deep learning in this domain.

DCVC. DCVC [187] presents advancements in video compression using deep learning, focusing on contextual information to improve compression efficiency and quality. It emphasizes the importance of incorporating contextual information into

Table 5.1 Various learning-based video compression methods and the category

Method	Time	Encoding	Temporal	Optimization	Direction
DVC	CVPR 2019	Res-based	Two-frame	Objective	Uni-directional
DCVC	NeurIPS 2021	Cond-based	Two-frame	Objective	Uni-directional
DVC-P	VCIP 2021	Res-based	Two-frame	Perceptual	Uni-directional
LSVC	CVPR 2021	Res-based	Two-frame	Objective	Uni-directional
SSFVC	CVPR 2020	Res-based	Two-frame	Objective	Uni-directional
HLVC	CVPR 2020	Res-based	Multi-frame	Objective	Bi-directional
RLVC	TSP 2020	Res-based	Multi-frame	Objective	Bi-directional
CEC-EVC	ECCV 2020	Cond-based	Two-frame	Objective	Uni-directional
PLVC	IJCAI 2022	Res-based	Two-frame	Perceptual	Bi-directional
MLVC	CVPR 2020	Res-based	Multi-frame	Objective	Uni-directional
VCT	NeurIP 2022	Res-based	Two-frame	Objective	Uni-directional
DCNN-VC	ICASSP 2022	Res-based	Two-frame	Objective	Uni-directional
ST-CCT	MM 2023	Res-based	Multi-frame	Objective	Uni-directional
DCVC-DC	CVPR 2023	Cond-based	Multi-frame	Objective	Uni-directional

Source Author

the compression process. This allows the model to better understand the relationships between different frames and enhance the compression of spatial and temporal features. The proposed model typically combines various neural network architectures to efficiently encode and decode video data. It often utilizes convolutional networks for spatial features and recurrent networks or attention mechanisms to capture temporal dependencies. The authors focus on optimizing the trade-off between bitrate and distortion. By incorporating contextual information, the model aims to minimize the perceptual loss while achieving high compression ratios. DCVC typically showcases experiments demonstrating the effectiveness of the proposed approach compared to traditional video compression techniques. Results often indicate improvements in compression efficiency and visual quality. The advancements made in this research could have implications for various applications, including streaming services, video conferencing, and any scenario where efficient video storage and transmission are critical.

DVC-P. DVC-P [188] addresses the need for improved perceptual quality in video compression. It recognizes that optimizing solely for traditional metrics like PSNR

may not always result in visually pleasing video outputs. Building on the foundation of the original DVC (Deep Video Compression) framework, DVC-P employs deep learning techniques to refine the compression process. This involves using neural networks to handle various aspects of video compression, including motion estimation, motion compensation, and residual encoding. Inspired by generative adversarial networks (GANs), DVC-P integrates a discriminator network into the compression framework. This addition allows the system to better assess the perceptual quality of the output, improving the overall video quality. The authors propose a mixed loss function that combines multiple loss components, including adversarial loss, Mean Squared Error (MSE) loss, VGG-based perceptual loss, and bitrate loss. This composite approach helps in balancing compression efficiency and perceptual quality. DVC-P is evaluated against traditional video compression methods, demonstrating significant improvements in both BD-rate (Bjontegaard delta rate) and visual quality, indicating its effectiveness in producing high-quality video content. The enhancements in DVC-P can be particularly beneficial for applications where visual fidelity is crucial, such as streaming services, virtual reality, and video conferencing.

SSF-VC. SSFVC [189] aims to improve the motion compensation module by addressing the limitations of conventional 2D flow-based compensation techniques. Scale-space flow techniques involve representing motion information at multiple scales, allowing the model to capture both fine details and broader motion patterns. By utilizing scale-space flow, the motion compensation module can better handle complex motion scenarios and varying motion speeds. This results in more accurate alignment of reference frames and target frames, leading to reduced residual errors. The integration of scale-space flow into an end-to-end video compression framework allows for joint optimization of the motion estimation and compensation processes alongside other components like residual compression and quantization. This holistic approach ensures that all aspects of compression work together to minimize bitrate while maintaining visual quality. Scale-space flow can help address challenges such as disocclusions (areas where objects move out of the frame) and artifacts that arise from traditional 2D flow-based compensation methods. By providing a more comprehensive understanding of motion across scales, the model can adaptively adjust to these challenges. The application of scale-space flow in video compression leads to better utilization of inter-frame correlations, allowing for more efficient encoding of frames. This ultimately results in lower bitrates while preserving higher visual fidelity. The use of scale-space flow in end-to-end optimized video compression represents a significant advancement in motion estimation and compensation techniques. By capturing motion at multiple scales, this approach enhances the accuracy and efficiency of video compression, paving the way for better quality and lower bitrate outputs.

HLVC. HLVC [190] presents a hierarchical learning-based video compression framework. The framework features three hierarchical quality modules alongside a recurrent enhancement network. High-quality frames in layer 1 enhance compression at the encoder, improving overall performance. During decoding, low-quality frames benefit from high-quality references, which boost quality without additional bitrate. For example, frames 3 and 8 in layer 3, initially compressed at low quality, receive

substantial improvements from the recurrent enhancement network by leveraging information from better frames. This approach achieves efficient video compression that surpasses the performance of x265's mode. HLVC utilizes image compression for layer 1, proposes the Bi-Directional Deep Compression (BDDC) network for layer 2, and introduces the Single Motion Deep Compression (SMDC) network for layer 3, which reduces the bitrate for motion maps. The Weighted Recurrent Quality Enhancement (WRQE) network enhances multi-frame information, and experiments and ablation studies validate its effectiveness, demonstrating the best performance.

LSVC. LSVC [191] leverages a GAN-based autoencoder architecture for efficient time-series data compression. The codec achieves a versatile rate-distortion trade-off, surpassing recent learning-based codecs. A ConvLSTM network is employed for learned video prediction in the latent domain, substituting traditional motion compensation. This approach reduces entropy by storing differences in a low-dimensional latent space, enhancing prediction quality through adversarial training against a discriminator. The proposed codec demonstrates superior performance in video compression evaluations. Additionally, the latent space prediction algorithm is applied to anomaly detection tasks, showcasing its effectiveness through unsupervised learning on event detection datasets. Key contributions include the innovative codec architecture, latent domain video prediction, and the application of perceptual tasks in the latent space.

RLVC. In the realm of video compression, Recurrent Learned Video Compression (RLVC) [192] stands out with its innovative approach, introducing recurrent structures to overcome the limitations inherent in non-recurrent methodologies. By incorporating the Recurrent AutoEncoder (RAE) and Recurrent Probability Model (RPM), RLVC recurrently taps into the information embedded in sequences of frames, offering a substantial enhancement in compression efficiency. Unlike non-recurrent counterparts, the RLVC method embraces a more extensive set of reference frames, enabling the capture of a diverse array of temporal correlations. The notable contributions of RLVC include the introduction of recurrent structures to exploit temporal correlations, the proposition of RAE to expand reference frames, and RPM for estimating temporally conditional Probability Mass Functions (PMF). Through rigorous experimental evaluations, RLVC not only demonstrates its superiority over existing methods but also validates the effectiveness of each recurrent component, as evidenced by comprehensive ablation studies.

PLVC. PLVC [193] focuses on addressing the limitations inherent in existing end-to-end learning-based image and video compression models, primarily tailored to optimize distortion metrics. The proposed system implements an adversarial training paradigm between a recurrent generator and a recurrent conditional discriminator to optimize perceptual compression performance. The generator architecture employs recurrent auto-encoders that simultaneously achieve efficient video compression and high-quality reconstruction. A key innovation is the recurrent conditional discriminator, which performs spatiotemporal feature analysis to assess both original and compressed videos, ensuring the output maintains both visual fidelity and temporal consistency throughout the adversarial learning process. Comprehensive evaluations, incorporating both objective metrics and subjective user studies, demonstrate that this

Perceptual Learned Video Compression (PLVC) approach surpasses existing video coding standards in visual quality. The framework makes three primary advances: development of a new perceptual video compression methodology, integration of adversarial loss functions to optimally trade off between bitrate efficiency, distortion metrics, and perceptual quality, and extensive experimental validation of the method's superior performance characteristics.

M-LVC. Unlike DVC, which employs a single reference frame, M-LVC [194] uses multiple frames, enhancing prediction by providing multiple hypotheses and reducing coding costs through multiple motion vector (MV) fields. This scheme explicitly addresses multiple frames prediction, offering scalability, interpretability, and improved training efficiency. Additionally, M-LVC introduces MV and residual refinement networks to compensate for compression errors, enhancing reconstruction quality. M-LVC integrates four key modules, including: (1) multi-frame motion vector (MV) prediction, (2) multi-frame motion compensation, (3) MV refinement, and (4) residual refinement. Using the unified rate-distortion loss function and the progressive training approach, M-LVC optimizes all components collectively, ensuring seamless integration and improved overall performance.

VCT. Departing from conventional methods relying on intricate components like flow prediction and warping, VCT (Video Compression Transformer) [195] employs a transformer-based temporal entropy model, offering simplicity and superior performance. Unlike existing techniques burdened by architectural biases, VCT excels across diverse video types, demonstrating its versatility. The two-step compression process involves lossy transform coding and leveraging temporal redundancies using a transformer. This innovative approach eliminates the need for complex operations and mitigates temporal error propagation, a common challenge in traditional methods. Drawing inspiration from language translation transformers, VCT faces the computational challenge posed by large input dimensions through the introduction of independence assumptions. This optimization results in a more manageable attention matrix, allowing parallel execution on subsets of symbols. The model's ease of implementation is highlighted, facilitated by compatibility with contemporary machine learning frameworks. To foster future exploration, the paper includes an extensive release of code and models, encouraging further advancements in this promising direction of neural video compression.

DCNN-VC. This work [196] addresses the critical aspects of motion estimation and compensation in video coding, particularly focusing on inter prediction. Recognizing the limitations of using original reference pictures with fractional pixel units for accurate motion compensation, DCNN-VC introduces a novel deep reference picture generator. Leveraging Convolutional Neural Network (CNN) advancements, DCNN-VC employs a dilated CNN to create a more contextually relevant reference picture for the current encoding frame. The generated deep picture is seamlessly integrated into the Versatile Video Coding (VVC) framework as a reference, and extensive experiments within the VVC Test Model–VTM demonstrate a notable 9.7% average bit saving compared to VVC under low-delay P configuration. The contributions include the development of a generator tailored for P frames, challenging due to limited reference frames, and the incorporation of the generated reference

picture into VVC, significantly reducing temporal redundancy and enhancing compression efficiency. This paper builds upon the success of CNNs in image processing, drawing inspiration from deep learning advancements to improve inter prediction in video coding.

ST-CCT. This work proposes an innovative Spatio-Temporal Cross-Covariance Transformer (ST-XCT) module, efficiently integrating spatial and temporal data for video compression. Departing from prior Neural Video Compression (NVC) methods, which separately handle spatial and temporal information, ST-XCT effectively fuses both dimensions through a unique 3D-based approach. By generating joint spatio-temporal features and applying 3D convolutions, it locally mixes data while leveraging the attention mechanism across feature channels. Notably, ST-XCT ensures global spatio-temporal correlation modeling without resorting to feature decomposition, maintaining linear complexity. Additionally, a 3D feed-forward gating mechanism enhances information flow by regulating and updating data based on spatio-temporal pixel correlation. The proposed ST-XCT module is systematically embedded throughout the neural video compression (NVC) pipeline, enhancing three critical aspects: feature representation, frame synthesis, and entropy estimation. Rigorous testing across multiple benchmarks (UVG, MCL-JCV, and HEVC sequences) demonstrates consistent outperformance against conventional standards (H.266/VVC) and leading learning-based approaches (DCVC, VCT), achieving remarkable compression efficiency.

VCII. VCII [197] is a groundbreaking work in the area of deep learning-based video compression. Rather than totally relying on hand-crafted modules, VCII utilizes image interpolation to reduce the temporal redundancy between different video frames. It also integrates traditional video coding tools and introduces a hierarchical reference structure to improve video compression performance.

Specifically, VCII first uses image compression technologies to compress every n -th frame as an I-frame, with n set to 12. To efficiently eliminate the temporal redundancy, a trained context network extracts feature maps from these I-frames. These feature maps then guide the compression of other frames, i.e. the R-frames. For instance, an R-frame uses the feature maps from its adjacent I-frames to aid its encoding and decoding process. To further exploit temporal motion, VCII tested two motion information alternatives, i.e. the calculated flow and the motion vectors from block motion estimation. Experimental results demonstrate that VCII surpasses traditional H.261 codecs and is comparable to H.264. Although VCII is innovative in deep learning-based video compression, it still incorporates some hand-craft elements like motion vectors. The inclusion of hand-crafted elements impairs its further end-to-end optimization.

CEC-EVC. Liu et al. [185] proposed Conditional Entropy Coding for Efficient Video Compression (CEC-EVC), exploiting the temporal redundancy between adjacent frames directly through conditional entropy coding, instead of depending on explicit motion estimation and compensation. Backed by extensive experimental results, they demonstrate that using conditional entropy with priors from adjacent frames sufficiently captures this redundancy. Traditional explicit motion modeling,

they argue, is not only unnecessary. Furthermore, Liu et al. [185] introduced an internal learning extension to adapt trained models for encoding test videos. This extension fixes the network parameters and applies back-propagation to latent codes, which effectively captures the distribution shift between training samples and testing samples. Notably, this extension increases encoding complexity without affecting the decoding process. By capturing the internal unique characteristics of each individual video, the internal learning extension further boosts the video compression performance of the proposed method. The proposed approach achieves better results than H.265 and other deep learning-based baselines in MS-SSIM on UVG video.

DCVC-DC. Li et al. [198] suggested leveraging features of the motion vector of the previous frame for significant coding gains. Specifically, they focused their model on learning hierarchical quality patterns throughout various frames. Additionally, they incorporated optical flow to augment the efficiency of their proposed deep learning-based video compression algorithm. To capture complex motion, they proposed using group-based offset diversity. This grouping operation occurs along the channel dimension of the feature maps. For spatial context, they introduced a quadtree-based partitioning method. Their proposed quadtree-based partitioning improves the efficiency of entropy coding and supports parallel operations. The proposed DCVC-DC offers significantly better rate-distortion performance than previous deep learning-based video compression algorithms, including next-generation codecs. The encoding and decoding times of DCVC-DC are approximately 1000 ms and 700 ms per frame for 1080p videos, which, while currently impractical, show potential.

5.4 Summary

This chapter explores the realm of video compression algorithms powered by deep learning, examining their historical development, future prospects, methodological classifications, overall architecture, specific components, and seminal methods. A forward-looking perspective is additionally provided regarding the future of deep learning-based video compression algorithms, which have demonstrated widespread applications across various sectors including streaming media, digital television, video conferencing, and mobile video communications, indicative of a promising and expansive future. The comprehensive framework and individual components of deep learning-based video compression algorithms are introduced, encompassing the entire training pipeline, feature extraction, motion estimation, motion compensation, quantization, and entropy coding. Classical methods in this domain are discussed through a categorical analysis based on encoding manner, temporal scheme, optimization objectives, and direction. Beginning with the pioneering deep learning-based video compression algorithm DVC, methods including DCVC, HLVC, and VCT are classified and sequentially summarized in chronological order, accompanied by detailed descriptions of each method's specific nuances. Video compression algorithms have been revolutionizing the way digital video content is processed

and transmitted, significantly enhancing efficiency while maintaining high-quality standards. Deep learning, with its prowess in handling complex data structures, has introduced groundbreaking advancements in this field. Looking ahead, the integration of advanced neural network architectures and innovative optimization techniques promises to further refine the performance of video compression algorithms. This evolution is likely to lead to more adaptive, intelligent compression systems capable of handling an ever-increasing need for higher definition video content. The future of deep learning in video compression heralds a new era of smart, efficient, and high-fidelity digital media transmission, paving the way for a more connected and visually rich world.

Exercises

1. Please explain how deep learning can be utilized to improve video compression efficiency. Discuss the basic principles behind it and how it differs from traditional video coding methods.
2. Please compare and contrast the two main schemes of the optical flow method in the motion estimation module, and discuss the advantages and disadvantages in terms of computational complexity, coding efficiency, and video quality.
3. In the previously introduced work, besides the classification criteria proposed in the text, please describe what additional criteria do you think can be used for classification.
4. Please provide the formulaic expression for entropy coding.
5. Consider a scenario where bandwidth is limited but high video quality is required. Please propose a deep learning-based solution for adaptive bitrate streaming and explain how it would optimize video delivery.
6. Given a set of symbols with individual information entropies denoted as H_i for each symbol, and the probability of each symbol occurring as p_i , please determine the lower bound of the minimum coding length for the entire set of symbols.
7. Please explain what is the hyperprior in the area of deep learning-based image/video compression.
8. Please explain what is Video Coding for Machine (VCM).
9. Please explain what is the motion estimation and motion compensation in the area of video coding.
10. Please describe what is Perceptual Coding.

References

1. W. Gao, G. Li, H. Yuan, R. Hamzaoui, Z. Li, S. Liu, Apccpa'22: 1st international workshop on advances in point cloud compression, processing and analysis, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 7392–7393
2. W. Gao, G. Li, Point cloud compression, enhancement and applications: from 3d perception to large models, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 292–11 293
3. H. Sun, G. Li, S. Chen, L. Li, W. Gao, Special topic on 3d point cloud processing and applications. *ZTE Commun.* **21**(4), 1 (2023)
4. W. Gao, W. Gao, X. Mu, C. Peng, G. Li, Overview and comparison of avs point cloud compression standard. *APSIPA Trans. Signal Inf. Process.* (2024)
5. W. Gao, S. Wang, X. Zhang, S. Kwong, Editorial for special issue on deep learning-based data compression. *APSIPA Trans. Signal Inf. Process.* **13**(6) (2024)
6. W. Gao, H. Ye, G. Li, H. Zheng, Y. Wu, L. Xie, OpenPointCloud: an open-source algorithm library of deep learning based point cloud compression, in *ACM International Conference on Multimedia* (2022), pp. 7347–7350
7. Y. Zhang, W. Gao, G. Li, Openpointcloud-v2: a deep learning based open-source algorithm library of point cloud processing, in *Proceedings of the 1st International Workshop on Advances in Point Cloud Compression, Processing and Analysis* (2022), pp. 51–55
8. L. Xie, W. Gao, Learningpcc: a pytorch library for learning-based point cloud compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 234–11 238
9. L. Xie, W. Gao, Pchmvision: an open-source library of point cloud compression for human and machine vision, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 239–11 243
10. L. Xie, W. Gao, H. Zheng, G. Li, Spcgc: scalable point cloud geometry compression for machine vision, in *2024 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2024), pp. 17 272–17 278
11. L. Xie, W. Gao, H. Zheng, H. Ye, Semantic-aware visual decomposition for point cloud geometry compression, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 595–595
12. Z. Qi, W. Gao, Variable-rate point cloud geometry compression based on feature adjustment and interpolation, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 63–72
13. Z. Yu, W. Gao, When dynamic neural network meets point cloud compression: computation-aware variable rate and checkerboard context, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 600–600
14. L. Xie, W. Gao, S. Fan, Z. Yao, Pdnet: parallel dual-branch network for point cloud geometry compression and analysis, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 596–596
15. L. Xie, W. Gao, H. Zheng, End-to-end point cloud geometry compression and analysis with sparse tensor, in *International Workshop on Advances in Point Cloud Compression, Processing and Analysis* (2022), pp. 27–32
16. C. Fu, G. Li, R. Song, W. Gao, S. Liu, Octattention: octree-based large-scale contexts model for point cloud compression. *Proc. AAAI Conf. Artif. Intell.* **36**(1), 625–633 (2022)
17. C. Zhang, W. Gao, Adadpcc: adaptive rate control and rate-distortion-complexity optimization for dynamic point cloud compression, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
18. K. Wang, W. Gao, Unipccg: towards practical point cloud geometry compression via an efficient unified approach, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
19. K. Zheng, W. Gao, H. Zheng, Octree-based learned point cloud geometry compression: A lossy perspective, in *2025 Data Compression Conference (DCC)* (IEEE, 2025)
20. Z. Qi, H. Liu, W. Gao, Variable-rate point cloud compression with channel-wise entropy modeling, in *2025 Data Compression Conference (DCC)* (IEEE, 2025)

21. H. Zheng, W. Gao, Z. Yu, T. Zhao, G. Li, Viewpcgc: view-guided learned point cloud geometry compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 7152–7161
22. L. Xie, W. Gao, H. Zheng, G. Li, Roi-guided point cloud geometry compression towards human and machine vision, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 3741–3750
23. Y. Wu, L. Xie, S. Sun, W. Gao, Y. Yan, Adaptive intra period size for deep learning-based screen content video coding, in *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2024)
24. H. Zheng, W. Gao, End-to-end rgb-d image compression via exploiting channel-modality redundancy. *Proc. AAAI Conf. Artif. Intell.* **38**(7), 7562–7570 (2024)
25. L. Tao, W. Gao, G. Li, C. Zhang, Adanic: towards practical neural image compression via dynamic transform routing, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 16 879–16 888
26. Y. Wu, W. Gao, End-to-end lossless compression of high precision depth maps guided by pseudo-residual (2022). [arXiv:2201.03195](https://arxiv.org/abs/2201.03195)
27. Y. Wu, Z. Qi, H. Zheng, L. Tao, W. Gao, Deep image compression with latent optimization and piece-wise quantization approximation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 1926–1930
28. W. Gao, L. Tao, L. Zhou, D. Yang, X. Zhang, Z. Guo, Low-rate image compression with super-resolution learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2020), pp. 154–155
29. W. Gao, S. Sun, H. Zheng, Y. Wu, H. Ye, Y. Zhang, Opendmc: an open-source library and performance evaluation for deep-learning-based multi-frame compression, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9685–9688
30. W. Gao, H. Zheng, C. Zhang, K. Zheng, Z. Yu, Y. Li, H. Ye, Y. Zhang, Opendic: an open-source library and performance evaluation for deep-learning-based image compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 202–11 205
31. Y. Li, W. Gao, G. Li, S. Ma, Saliency segmentation oriented deep image compression with novel bit allocation. *IEEE Trans. Image Process.* **34**, 16–29 (2025)
32. C. Zhang, W. Gao, Learned rate control for frame-level adaptive neural video compression via dynamic neural network, in *European Conference on Computer Vision* (Springer, 2024)
33. W. Liu, W. Gao, X. Mu, Fast inter-frame motion prediction for compressed dynamic point cloud attribute enhancement. *Proc. AAAI Conf. Artif. Intell.* **38**(4), 3720–3728 (2024)
34. Z. Yang, W. Gao, X. Lu, Danet: density-adaptive network for geometry-based point cloud compression artifacts removal, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
35. X. Fan, G. Li, D. Li, Y. Ren, W. Gao, T.H. Li, Deep geometry post-processing for decompressed point clouds, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
36. X. Zhang, G. Liao, W. Gao, G. Li, Tdrnet: transformer-based dual-branch restoration network for geometry based point cloud compression artifacts, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
37. Z. Li, G. Li, T.H. Li, S. Liu, W. Gao, Semantic point cloud upsampling. *IEEE Trans. Multimedia* (2022)
38. Z. Li, S. Liu, W. Gao, G. Li, G. Li, S4r: rethinking point cloud sampling via guiding upsampling-aware perception. *IEEE Trans. Multimedia* (2025)
39. R. Zhang, W. Gao, G. Li, T.H. Li, Qinet: decision surface learning and adversarial enhancement for quasi-immune completion of diverse corrupted point clouds. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–14 (2022)
40. R. Bao, Y. Ren, G. Li, W. Gao, S. Liu, Flow-based point cloud completion network with adversarial refinement, in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2559–2563

41. J. Chen, G. Li, R. Zhang, T.H. Li, W. Gao, Pointvae: invertible variational autoencoder framework for 3d point cloud generation, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 3216–3220
42. R. Zhang, J. Chen, W. Gao, G. Li, T.H. Li, Pointot: interpretable geometry-inspired point cloud generative model via optimal transport. *IEEE Trans. Circuits Syst. Video Technol.* **32**(10), 6792–6806 (2022)
43. Z. Li, G. Li, T. Li, S. Liu, W. Gao, Information-growth attention network for image super-resolution, in *Proceedings of the 29th ACM International Conference on Multimedia* (2021), pp. 544–552
44. L. Zhou, W. Gao, G. Li, H. Yuan, T. Zhao, G. Yue, Disentangled feature distillation for light field super-resolution with degradations, in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2023), pp. 116–121
45. L. Zhou, W. Gao, G. Li, End-to-end spatial-angular light field super-resolution using parallax structure preservation strategy, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 3396–3400
46. W. Gao, L. Zhou, L. Tao, A fast view synthesis implementation method for light field applications. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* **17**(4), 1–20 (2021)
47. X. Zhang, W. Gao, G. Li, Q. Jiang, R. Cong, Image quality assessment-driven reinforcement learning for mixed distorted image restoration. *ACM Trans. Multimedia Comput. Commun. Appl.* **19**(1s), 1–23 (2023)
48. X. Zhang, W. Gao, H. Yuan, G. Li, Je 2 net: joint exploitation and exploration in reinforcement learning based image restoration, in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2090–2094
49. X. Zhang, W. Gao, Hirl: hybrid image restoration based on hierarchical deep reinforcement learning via two-step analysis, in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2445–2449
50. S. Wang, Y. Lu, W. Xia, P. Xia, Z. Wang, W. Gao, Light field angular super-resolution by view-specific queries. *Visual Comput.* 1–16 (2024)
51. Y. Sun, Z. Li, S. Wang, W. Gao, Depth-assisted calibration on learning-based factorization for a compressive light field display. *Opt. Express* **31**(4), 5399–5413 (2023)
52. Y. Sun, Z. Li, L. Li, S. Wang, W. Gao, Optimization of compressive light field display in dual-guided learning, in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2075–2079
53. W. Gao, S. Fan, G. Li, W. Lin, A thorough benchmark and a new model for light field saliency detection. *IEEE Trans. Pattern Anal. Mach. Intell.* (2023)
54. Z. Guo, W. Gao, H. Wang, J. Wang, S. Fan, No-reference deep quality assessment of compressed light field images, in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6
55. T. Qin, G. Li, W. Gao, S. Liu, Multi-grained point cloud geometry compression via dual-model prediction with extended octree. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(9), 1–30 (2024)
56. Y. Shao, W. Gao, S. Liu, G. Li, Advanced patch-based affine motion estimation for dynamic point cloud geometry compression. *Sensors* **24**(10), 3142 (2024)
57. Y. Shao, F. Song, W. Gao, S. Liu, G. Li, Texture-guided graph transform optimization for point cloud attribute compression. *Appl. Sci.* **14**(10), 4094 (2024)
58. Y. Shao, X. Yang, W. Gao, S. Liu, G. Li, 3d point cloud attribute compression using diffusion-based texture-aware intra prediction. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
59. J. Zhang, Y. Chen, G. Liu, W. Gao, G. Li, Efficient point cloud attribute compression framework using attribute-guided graph fourier transform, in *ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2024), pp. 8426–8430
60. W. Gao, H. Yuan, G. Li, Z. Li, H. Yuan, Low complexity coding unit decision for video-based point cloud compression. *IEEE Trans. Image Process.* **33**, 149–162 (2023)

61. Y. Shao, G. Li, Q. Zhang, W. Gao, S. Liu, Nonrigid registration-based progressive motion compensation for point cloud geometry compression. *IEEE Trans. Geosci. Remote Sens.* (2023)
62. F. Song, G. Li, X. Yang, W. Gao, S. Liu, Block-adaptive point cloud attribute coding with region-aware optimized transform. *IEEE Trans. Circuits Syst. Video Technol.* (2023)
63. Y. An, Y. Shao, G. Li, W. Gao, S. Liu, A fast motion estimation method with hamming distance for lidar point cloud compression, in *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2022), pp. 1–5
64. H. Yuan, W. Gao, G. Li, Z. Li, Rate-distortion-guided learning approach with cross-projection information for v-pcc fast cu decision, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 3085–3093
65. F. Song, G. Li, W. Gao, T.H. Li, Rate-distortion optimized graph for point cloud attribute coding. *IEEE Signal Process. Lett.* **29**, 922–926 (2022)
66. F. Song, G. Li, X. Yang, W. Gao, T.H. Li, Fine-grained correlation representation for graph-based point cloud attribute compression, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
67. F. Shen, W. Gao, A rate control algorithm for video-based point cloud compression, in *2021 International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2021), pp. 1–5
68. F. Song, Y. Shao, W. Gao, H. Wang, T. Li, Layer-wise geometry aggregation framework for lossless lidar point cloud compression. *IEEE Trans. Circuits Syst. Video Technol.* **31**(12), 4603–4616 (2021)
69. Z. Cai, W. Gao, G. Li, W. Gao, Distortion propagation model-based v-pcc rate control for 3d point cloud broadcast. *IEEE Trans. Broadcast.* 1–13 (2024)
70. X. Mu, W. Gao, H. Yuan, S. Wang, G. Li, Adaptive lpu decision for dynamic point cloud compression. *IEEE Signal Process. Lett.* (2024)
71. C. Peng, W. Gao, Laplacian matrix learning for point cloud attribute compression with ternary search-based adaptive block partition, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 10 412–10 420
72. Y. Wang, W. Gao, X. Mu, H. Yuan, Rate control optimization for joint geometry and attribute coding of lidar point clouds, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
73. T. Guo, H. Yuan, T. Wang, W. Gao, Graph filter-based fast motion matching for inter frame coding of mpeg g-pcc, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 1151–1155
74. Z. Yang, W. Gao, G. Li, Y. Yan, Sur-driven video coding rate control for jointly optimizing perceptual quality and buffer control. *IEEE Trans. Image Process.* (2023)
75. F. Shen, Z. Cai, W. Gao, An efficient rate control algorithm for intra frame coding in avs3, in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2021), pp. 3164–3169
76. H. Yuan, W. Gao, J. Wang, Dynamic computational resource allocation for fast inter frame coding in video conferencing applications, in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6
77. W. Gao, Q. Jiang, R. Wang, S. Ma, G. Li, S. Kwong, Consistent quality oriented rate control in hevc via balancing intra and inter frame coding. *IEEE Trans. Ind. Inform.* **18**(3), 1594–1604 (2021)
78. H. Yuan, W. Gao, A new coding unit partitioning mode for screen content video coding, in *Proceedings of the 2021 5th International Conference on Digital Signal Processing* (2021), pp. 66–72
79. W. Gao, On the performance evaluation of state-of-the-art rate control algorithms for practical video coding and transmission systems, in *Proceedings of the 2020 4th International Conference on Video and Image Processing* (2020), pp. 179–185
80. W. Gao, S. Kwong, Q. Jiang, C.-K. Fong, P.H. Wong, W.Y. Yuen, Data-driven rate control for rate-distortion optimization in hevc based on simplified effective initial qp learning. *IEEE Trans. Broadcast.* **65**(1), 94–108 (2018)

81. W. Gao, A multi-objective optimization perspective for joint consideration of video coding quality, in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (IEEE, 2019), pp. 986–991
82. W. Gao, S. Kwong, Y. Jia, Joint machine learning and game theory for rate control in high efficiency video coding. *IEEE Trans. Image Process.* **26**(12), 6074–6089 (2017)
83. W. Gao, S. Kwong, Y. Zhou, H. Yuan, Ssim-based game theory approach for rate-distortion optimized intra frame ctu-level bit allocation. *IEEE Trans. Multimedia* **18**(6), 988–999 (2016)
84. W. Gao, S. Kwong, H. Yuan, X. Wang, Dct coefficient distribution modeling and quality dependency analysis based frame-level bit allocation for hevcc. *IEEE Trans. Circuits Syst. Video Technol.* **26**(1), 139–153 (2015)
85. W. Gao, S. Kwong, Phase congruency based edge saliency detection and rate control for perceptual image and video coding, in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2016), pp. 000 264–000 269
86. H. Yuan, W. Gao, Openfastvc: an open source library for video coding fast algorithm implementation, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9660–9663
87. H. Yuan, W. Gao, S. Ma, Y. Yan, Divide-and-conquer-based rdo-free cu partitioning for 8k video compression. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(4), 1–20 (2024)
88. L. Tao, W. Gao, A hardware implementation of entropy encoder for 8k video coding, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
89. Y. Guo, W. Gao, S. Ma, G. Li, Accelerating transform algorithm implementation for efficient intra coding of 8k uhd videos. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* **18**(4), 1–20 (2022)
90. Z. Cai, W. Gao, Efficient fast algorithm and parallel hardware architecture for intra prediction of avs3, in *IEEE International Symposium on Circuits and Systems (ISCAS)* (IEEE, 2021), pp. 1–5
91. W. Gao, H. Yuan, Y. Guo, L. Tao, Z. Cai, G. Li, Openhardwarevc: an open source library for 8k uhd video coding hardware implementation, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 7339–7342
92. W. Gao, H. Yuan, G. Liao, Z. Guo, J. Chen, Pp8k: a new dataset for 8k uhd video compression and processing. *IEEE MultiMedia* (2023)
93. W. Liu, W. Gao, G. Li, S. Ma, T. Zhao, H. Yuan, Enlarged motion-aware and frequency-aware network for compressed video artifact reduction. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
94. G. Li, G. Wei, W. Gao, *Point Cloud Compression: Technologies and Standardization* (Springer Nature, 2024)
95. G. Li, W. Gao, W. Gao, Introduction, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 1–28
96. G. Li, W. Gao, W. Gao, Background knowledge, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 29–51
97. G. Li, W. Gao, W. Gao, Predictive coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 53–70
98. G. Li, W. Gao, W. Gao, Transform coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 71–96
99. G. Li, W. Gao, W. Gao, Quantization techniques, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 97–112
100. G. Li, W. Gao, W. Gao, Entropy coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 113–133
101. G. Li, W. Gao, W. Gao, Mpeg geometry-based point cloud compression (g-pcc) standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 135–165
102. G. Li, W. Gao, W. Gao, Avs point cloud compression standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 167–197
103. G. Li, W. Gao, W. Gao, Mpeg video-based point cloud compression (v-pcc) standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 199–218

104. G. Li, W. Gao, W. Gao, Mpeg ai-based 3d graphics coding standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 219–241
105. G. Li, W. Gao, W. Gao, Future work, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 243–250
106. L. Xie, X. Mu, G. Li, W. Gao et al., Pku-dpcc: a new dataset for dynamic point cloud compression. *APSIPA Trans. Signal Inf. Process.* **13**(6) (2024)
107. H. Yuan, W. Gao, W. Gao, Opensep: an open source subjective experiment platform, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 248–11 251
108. S. Fan W. Gao, Screen-based 3d subjective experiment software, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9672–9675
109. J. Wang, W. Gao, G. Li, Zoom to perceive better: No-reference point cloud quality assessment via exploring effective multiscale feature. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
110. J. Wang, W. Gao, G. Li, Applying collaborative adversarial learning to blind point cloud quality measurement. *IEEE Trans. Instrum. Meas.* (2023)
111. H. Li, W. Gao, Information exploration of projected views for point cloud quality measurement. *IEEE Trans. Instrum. Meas.* (2024)
112. B. Qu, X. Liang, S. Sun, W. Gao, Exploring aigc video quality: a focus on visual harmony, video-text consistency and domain distribution gap, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2024)
113. B. Qu, H. Li, W. Gao, Bringing textual prompt to ai-generated image quality assessment, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024)
114. S. Sun, X. Liang, S. Fan, W. Gao, W. Gao, Ve-bench: subjective-aligned benchmark suite for text-driven video editing quality assessment, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
115. S. Fan, W. Gao, G. Li, Salient object detection for point clouds, in *European conference on computer vision* (Springer, 2022), pp. 1–19
116. S. Luo, W. Gao, A general framework for rotation invariant point cloud analysis, in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2024), pp. 3665–3669
117. X. Lu, W. Gao, Attentivenet: Detecting small objects for lidar point clouds by attending to important points, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
118. Z. Pan, N. Zhang, W. Gao, S. Liu, G. Li, Less is more: label recommendation for weakly supervised point cloud semantic segmentation. *Proc. AAAI Conf. Artif. Intell.* **38**(5), 4397–4405 (2024)
119. Z. Pan, G. Liu, W. Gao, T.H. Li, Epcontrast: effective point-level contrastive learning for large-scale point cloud understanding, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024), pp. 1–6
120. N. Zhang, Z. Pan, T.H. Li, W. Gao, G. Li, Improving graph representation for point cloud segmentation via attentive filtering, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 1244–1254
121. K. Wen, N. Zhang, G. Li, W. Gao, Mpvnn: multi-resolution point-voxel non-parametric network for 3d point cloud processing, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024), pp. 1–6
122. Z. Pan, N. Zhang, W. Gao, S. Liu, G. Li, Point cloud semantic segmentation with sparse and inhomogeneous annotations, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
123. R. Zhang, G. Li, W. Gao, S. Liu, A quantum-inspired framework in master-servant mode for large-scale multi-modal place recognition. *IEEE Trans. Intell. Transp. Syst.* (2024)
124. D. Yang, W. Gao, Pointchd: a point cloud benchmark for congenital heart disease classification and segmentation. *IEEE J. Biomed. Health Inform.* (2024)
125. S. Fan, W. Gao, G. Li, Point-mpp: point cloud self-supervised learning from masked position prediction. *IEEE Trans. Neural Netw. Learn. Syst.* (2024)

126. Z. Pan, W. Gao, G. Li, Distribution guidance network for weakly supervised point cloud semantic segmentation, in *Advances in Neural Information Processing Systems* (2024)
127. Y. Li, Z. Yu, X. Chen, S. Wang, H. Yuan, W. Gao, Zero-shot quantization for large-kernels via shape-based distribution and diversity self-distillation, in *ICASSP 2025–2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2025)
128. X. Chen, Y. Wang, Y. Li, X. Ling, M. Li, R. Liu, M. Ouyang, K. Zhao, T. Guan, Y. He, Low bit-width zero-shot quantization with soft feature-infused hints for iot systems. *IEEE Internet Things J.* (2024)
129. R. Zhang, G. Li, W. Gao, T.H. Li, Compoint: can complex-valued representation benefit point cloud place recognition? *IEEE Trans. Intell. Transp. Syst.* (2024)
130. X. Zang, W. Gao, G. Li, H. Fang, C. Ban, Z. He, H. Sun, A baseline investigation: Transformer-based cross-view baseline for text-based person search, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 7737–7746
131. G. Liao, W. Gao, G. Li, J. Wang, S. Kwong, Cross-collaborative fusion-encoder network for robust rgb-thermal salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **32**(11), 7646–7661 (2022)
132. W. Gao, G. Liao, S. Ma, G. Li, Y. Liang, W. Lin, Unified information fusion network for multimodal rgb-d and rgb-t salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **32**(4), 2091–2106 (2021)
133. Y. Chen, S. Sun, G. Li, W. Gao, T.H. Li, Closing the gap between theory and practice during alternating optimization for gans. *IEEE Trans. Neural Netw. Learn. Syst.* (2023)
134. Y. Chen, C. Jin, G. Li, T.H. Li, W. Gao, Mitigating label noise in gans via enhanced spectral normalization. *IEEE Trans. Circuits Syst. Video Technol.* (2023)
135. X. Zang, G. Li, W. Gao, Multidirection and multiscale pyramid in transformer for video-based pedestrian retrieval. *IEEE Trans. Ind. Inform.* **18**(12), 8776–8785 (2022)
136. X. Zang, G. Li, W. Gao, X. Shu, Learning to disentangle scenes for person re-identification. *Image Vis. Comput.* **116**, 104330 (2021)
137. X. Zang, G. Li, W. Gao, X. Shu, Exploiting robust unsupervised video person re-identification. *IET Image Process.* **16**(3), 729–741 (2022)
138. Z. Yue, G. Li, W. Gao, Cross-level guided attention for human-object interaction detection, in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2023), pp. 284–289
139. S. Sun, J. Liu, H. Li, G. Liu, T.H. Li, W. Gao, Streamflow: streamlined multi-frame optical flow estimation for video sequences, in *Advances in Neural Information Processing Systems* (2024)
140. Z. Yao, W. Gao, Iterative saliency aggregation and assignment network for efficient salient object detection in optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* (2024)
141. G. Liao, W. Gao, Rethinking feature mining for light field salient object detection. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(10), 1–24 (2024)
142. B. Yang, W. Gao, G. Li, Focus and adjust: progressive refinement network for human object interaction detection, in *2022 26th International Conference on Pattern Recognition (ICPR)* (IEEE, 2022), pp. 2546–2552
143. G. Liao, W. Gao, Q. Jiang, R. Wang, G. Li, Mmnet: multi-stage and multi-scale fusion network for rgb-d salient object detection, in *Proceedings of the 28th ACM International Conference on Multimedia* (2020), pp. 2436–2444
144. W. Gao, S. Kwong, Y. Zhou, Y. Jia, J. Zhang, W. Wu, Multiscale phase congruency analysis for image edge visual saliency detection, in *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 1 (IEEE, 2016), pp. 75–80
145. W. Gao, S. Kwong, Y. Zhou, X. Wang, Phase congruency analysis of down-sampled and blurring images for foreground extraction, in *2015 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)* (IEEE, 2015), pp. 152–157
146. Y. Guo, W. Gao, G. Li, Interpretable task-inspired adaptive filter pruning for neural networks under multiple constraints. *Int. J. Comput. Vis.* 1–17 (2024)

147. W. Gao, Y. Guo, S. Ma, G. Li, S. Kwong, Efficient neural network compression inspired by compressive sensing. *IEEE Trans. Neural Netw. Learn. Syst.* (2022)
148. Y. Guo, W. Gao, Semantic-driven automatic filter pruning for neural networks, in *IEEE international conference on multimedia and expo (ICME)* (IEEE, 2022), pp. 1–6
149. L. Tao, W. Gao, Efficient channel pruning based on architecture alignment and probability model bypassing, in *IEEE International Conference on systems, Man, and Cybernetics (SMC)* (IEEE, 2021), pp. 3232–3237
150. W. Gao, G. Li, *Deep Learning for 3D Point Clouds* (Springer Nature, 2024)
151. W. Gao, G. Li, Introduction to 3d point clouds: Datasets and perception, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 1–27
152. W. Gao, G. Li, Learning basics for 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 29–70
153. W. Gao, G. Li, Deep-learning-based point cloud enhancement i, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 71–97
154. W. Gao, G. Li, Deep-learning-based point cloud enhancement ii, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 99–130
155. W. Gao, G. Li, Deep-learning-based point cloud analysis i, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 131–162
156. W. Gao, G. Li, Deep-learning-based point cloud analysis ii, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 163–193
157. W. Gao, G. Li, Point cloud pre-trained models and large models, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 195–225
158. W. Gao, G. Li, Point cloud-language multi-modal learning, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 227–254
159. W. Gao, G. Li, Open-source projects for 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 255–272
160. W. Gao, G. Li, Typical engineering applications of 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2025), pp. 273–299
161. W. Gao, G. Li, Future work on deep learning-based point cloud technologies, in *Deep Learning for 3D Point Clouds* (Springer, 2025), pp. 301–315
162. G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, Z. Gao, Dvc: an end-to-end deep video compression framework (2018). [arXiv:1812.00101](https://arxiv.org/abs/1812.00101)
163. L. Xie, M. Hu, X. Bai, Towards hardware-friendly and robust facial landmark detection method, in *International Conference on Neural Information Processing* (Springer, 2022), pp. 432–444
164. L. Xie, M. Hu, X. Bai, Online improved vehicle tracking accuracy via unsupervised route generation, in *IEEE 34th International Conference on Tools with Artificial Intelligence* (IEEE, 2022), pp. 788–792
165. R. Yang, L.V. Gool, R. Timofte, Opendvc: an open source implementation of the dvc video compression method (2020) [Online]. Available: [arXiv:abs/2006.15862](https://arxiv.org/abs/2006.15862)
166. D. Sun, X. Yang, M.-Y. Liu, J. Kautz, Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 8934–8943
167. Z. Teed, J. Deng, Raft: recurrent all-pairs field transforms for optical flow, in *European Conference on Computer Vision* (2020), pp. 402–419
168. S. Zhao, Y. Sheng, Y. Dong, E.I. Chang, Y. Xu et al., Maskflownet: asymmetric feature matching with learnable occlusion mask, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 6278–6287
169. S. Sun, Y. Chen, Y. Zhu, G. Guo, G. Li, Skflow: learning optical flow with super kernels. *Adv. Neural Inf. Process. Syst.* **35**, 11 313–11 326 (2022)
170. T.-W. Hui, X. Tang, C.C. Loy, Liteflownet: a lightweight convolutional neural network for optical flow estimation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 8981–8989

171. T.-W. Hui, X. Tang, C.C. Loy, A lightweight optical flow cnn—revisiting data fidelity and regularization. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(8), 2555–2569 (2020)
172. G. Yang, D. Ramanan, Volumetric correspondence networks for optical flow. *Adv. Neural Inf. Process. Syst.* **32**, 793–803 (2019)
173. Z. Yin, T. Darrell, F. Yu, Hierarchical discrete distribution decomposition for match density estimation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 6044–6053
174. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 770–778
175. G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 4700–4708
176. Z. Huang, X. Shi, C. Zhang, Q. Wang, K. C. Cheung, H. Qin, J. Dai, H. Li, Flowformer: a transformer architecture for optical flow (2022). [arXiv:2203.16194](https://arxiv.org/abs/2203.16194)
177. X. Shi, Z. Huang, D. Li, M. Zhang, K.C. Cheung, S. See, H. Qin, J. Dai, H. Li, Flowformer++: masked cost volume autoencoding for pretraining optical flow estimation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 1599–1610
178. X. Shi, Z. Huang, W. Bian, D. Li, M. Zhang, K.C. Cheung, S. See, H. Qin, J. Dai, H. Li, Videoflow: exploiting temporal cues for multi-frame optical flow estimation (2023). [arXiv:2303.08340](https://arxiv.org/abs/2303.08340)
179. S. Sun, J. Liu, T.H. Li, H. Li, G. Liu, W. Gao, Streamflow: streamlined multi-frame optical flow estimation for video sequences (2023). [arXiv:2311.17099](https://arxiv.org/abs/2311.17099)
180. C.E. Shannon, A mathematical theory of communication. *Bell Syst. Tech. J.* **27**(3), 379–423 (1948)
181. J. Ballé, V. Laparra, E.P. Simoncelli, End-to-end optimized image compression (2016). [arXiv:1611.01704](https://arxiv.org/abs/1611.01704)
182. J. Ballé, D. Minnen, S. Singh, S.J. Hwang, N. Johnston, Variational image compression with a scale hyperprior (2018). [arXiv:1802.01436](https://arxiv.org/abs/1802.01436)
183. D. Minnen, J. Ballé, G.D. Toderici, Joint autoregressive and hierarchical priors for learned image compression. *Adv. Neural Inf. Process. Syst.* **31** (2018)
184. D. Minnen, S. Singh, Channel-wise autoregressive entropy models for learned image compression. in *IEEE International Conference on Image Processing* (2020), pp. 3339–3343
185. J. Liu, S. Wang, W.-C. Ma, M. Shah, R. Hu, P. Dhawan, R. Urtasun, Conditional entropy coding for efficient video compression, in *European Conference on Computer Vision* (Springer, 2020), pp. 453–468
186. G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, Z. Gao, Dvc: an end-to-end deep video compression framework, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 11 006–11 015
187. J. Li, B. Li, Y. Lu, Deep contextual video compression. *Adv. Neural Inf. Process. Syst.* **34**, 18 114–18 125 (2021)
188. S. Zhang, M. Mrak, L. Herranz, M.G. Blanch, S. Wan, F. Yang, Dvc-p: deep video compression with perceptual optimizations, in *International Conference on Visual Communications and Image Processing* (2021), pp. 1–5
189. E. Agustsson, D. Minnen, N. Johnston, J. Balle, S.J. Hwang, G. Toderici, Scale-space flow for end-to-end optimized video compression, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 8503–8512
190. R. Yang, F. Mentzer, L.V. Gool, R. Timofte, Learning for video compression with hierarchical quality and recurrent enhancement, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 6628–6637
191. B. Liu, Y. Chen, S. Liu, H.-S. Kim, Deep learning in latent space for video prediction and compression, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 701–710

192. R. Yang, F. Mentzer, L. Van Gool, R. Timofte, Learning for video compression with recurrent auto-encoder and recurrent probability model. *IEEE J. Sel. Top. Signal Process.* **15**(2), 388–401 (2020)
193. R. Yang, R. Timofte, L. Van Gool, Perceptual learned video compression with recurrent conditional gan, in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence* (2022), pp. 1537–1544
194. J. Lin, D. Liu, H. Li, F. Wu, M-lvc: multiple frames prediction for learned video compression, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 3546–3554
195. F. Mentzer, G.D. Toderici, D. Minnen, S. Caelles, S.J. Hwang, M. Lucic, E. Agustsson, Vct: a video compression transformer. *Adv. Neural Inf. Process. Syst.* **35**, 13 091–13 103 (2022)
196. H. Tian, P. Gao, R. Wei, M. Paul, Dilated convolutional neural network-based deep reference picture generation for video compression, in *IEEE International Conference on Acoustics, Speech and Signal Processing* (2022), pp. 2824–2828
197. C.-Y. Wu, N. Singhal, P. Krahenbuhl, Video compression through image interpolation, in *European Conference on Computer Vision* (2018), pp. 416–431
198. J. Li, B. Li, Y. Lu, Neural video compression with diverse contexts, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 22 616–22 626

Chapter 6

Deep Learning-based 3D and Multimodal Coding



6.1 Introduction

3D image technology has become much more popular in the fields of digital entertainment, scientific visualization, virtual reality, and scene perception [1–34]. By giving images a realistic 3D visual experience, 3D images provide a richer and more immersive visual experience, profoundly changing the way people interact with digital content. This technology is not only widely used in video games, film production, medical imaging, engineering design and other fields, but also constantly expanding into new applications, such as virtual reality travel, online education and remote collaboration [35–46].

With the rapid development of 3D imaging technology, an obvious problem is the dramatic increase in the amount of data [47–58]. The complexity of 3D image leads to its relatively large storage and transmission resources, which limits its wide application in real application scenes. With this background, 3D image compression technology came into being and became the key to solving this problem [21, 59–121]. The goal of 3D image compression is to achieve efficient storage and transmission of 3D images by eliminating redundant information and simplifying data representation while maintaining image quality [122–133, 133–152]. The development of this field not only involves algorithms and engineering techniques, but also requires a deep understanding of the principles of 3D image structure and human visual perception (Fig. 6.1).

This chapter explores the key concepts, application areas, and challenges of 3D images and their compression technology [153–161]. First, 3D images are categorized into stereo images, multi-modal images, light field images, and 360° images. Stereo images are based on data captured by a dual-camera system and achieve depth perception by simulating the stereo parallax of the human eye. This method finds wide application in simulated reality, computer vision, and virtual reality, providing users with a more realistic and immersive visual experience. Multi-modal images integrate information from different sensors or sources, including color, depth, texture, and

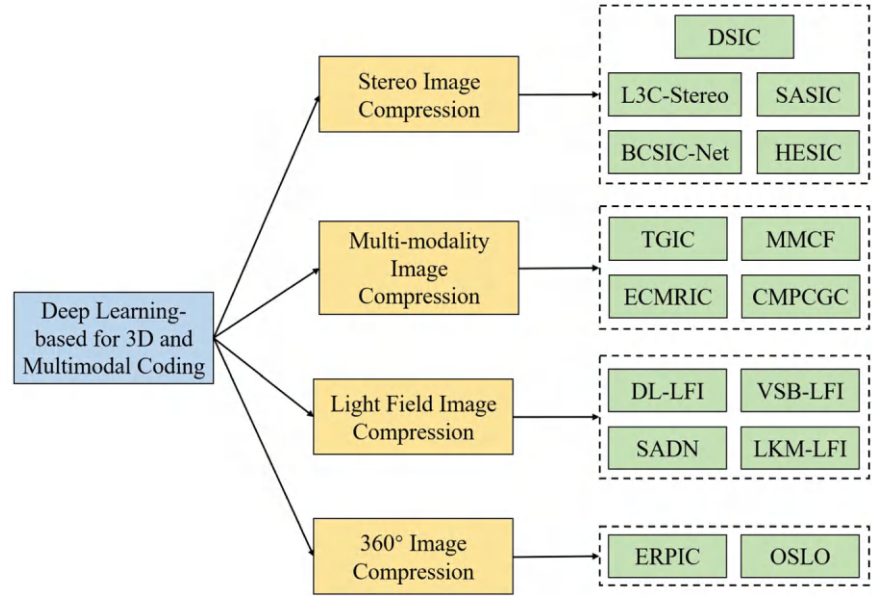


Fig. 6.1 Overview of multimodal compression frameworks. *Source* Author

other types of data. This comprehensive use of image data from various modes makes the presentation of 3D images more comprehensive and multi-dimensional, offering richer information for diverse application scenarios. Light field images record and simulate the path of light to more accurately restore optical properties. This technology not only provides highly realistic images but also enables more flexible focal length and depth of field adjustments at a later stage, giving users the freedom to explore different dimensions of the image. 360° images place the user in a panoramic view, providing an all-around viewing experience. This technology has great potential in virtual tourism, cultural heritage displays, and virtual conferences, creating an immersive feeling for users that goes beyond the limitations of traditional flat images.

This chapter thoroughly explores 3D image compression, enabling a deeper understanding of the position and function of these two technologies in the digital age. Simultaneously, such investigation facilitates insights into cutting-edge research in this field, offering novel perspectives and momentum for scientific and technological innovation. To comprehensively present the diversity and characteristics of learning-based multimodal compression frameworks, we meticulously organize various typical methods in Table 6.1. This table serves as a crucial resource for readers to quickly understand the wide range of approaches within this field.

Table 6.1 Summary of learning-based multimodal compression frameworks

Method	Author	Publication	Category
DSIC	Liu et al.	ICCV 2019	Stereo image compression
L3C-Stereo	Huang et al.	arXiv 2021	
SASIC	Wödlinger et al.	CVPR 2022	
BCSIC-Net	Lei et al.	CVPR 2022	
HESIC	Deng et al.	CVPR 2021	
TGIC	Jiang et al.	arXiv 2023	Multi-modality image compression
MMCF	Lu et al.	CVPR 2022	
ECMRIC	Zheng et al.	AAAI 2024	
CMPCGC	Lin et al.	ICASSP 2023	
DL-LFI	Zhao et al.	ICME 2018	Light field image compression
VSB-LFI	Liu et al.	Information Sciences 2021	
SADN	Tong et al.	arXiv 2022	
LKM-LFI	Shi et al.	arXiv 2023	
ERPIC	Li et al.	TIP 2022	360° image compression
OSLO	Bidgoli et al.	arXiv 2021	

Source Author

6.2 Overall Framework and Datasets

6.2.1 Immersive Media

Immersive media is a form of media that uses technology to create immersive experience. The use of virtual reality, augmented reality, 360° video and other interactive technologies brings a feeling of being fully present in the viewer. Immersive media is widely used in games, movies, and virtual reality experiences. The audience can immerse themselves in the world of the game or become part of the story of the film, enhancing the sense of engagement and realism of the entertainment experience. In the field of education, immersive media provides students with an immersive learning experience. Applications such as virtual LABS, historical re-enactments and virtual field trips can enhance students’ subject understanding and interest. Immersive media are used for medical training, surgical planning, and patient care. Doctors and students can use virtual reality technology to simulate actual operations and improve their skills.

Immersive media is closely related to 3D images, and together they create an engaging digital experience. Immersive media liberates users from the traditional two-dimensional media experience by applying technologies such as virtual reality

and augmented reality, enabling them to immerse themselves in a realistic three-dimensional environment. 3D graphics play a key role in this experience, providing a sense of space and realism to the immersive medium [162, 163]. Through carefully designed 3D models and scenes, users can feel the depth and stereoscopic, realistic visual effects, making the entire media experience closer to reality. In areas such as gaming, virtual tourism, and education, the application of 3D images makes immersive media more vivid and interesting. This combination creates a more interactive and engaging experience for users, driving a new era of digital entertainment and learning. However, the technical challenges associated with real-time rendering and the processing of high-quality graphics also drive continuous technological innovation, aiming to continuously improve the integration of immersive media with 3D images.

6.2.2 Transformation

The stereo image compression framework is shown in Fig. 6.2. First, the left image (designated as x_1) is compressed by an auto-encoder. After a spatial transformation, x_1 undergoes compression by the auto-encoder again. Subsequently, x_2 is compressed, leveraging priors from the align module. To save bit-rate, considering the correlation between them, the probability function of the second auto-encoder's latent representation (denoted as y_2) is modeled based on the latent representation of x_1 . Finally, the compressed feature output from the second auto-encoder is combined with the transformed x_1 to reconstruct x_2 . It should be noted that the align module can be placed either before the encoder, after the encoder, or within the entropy model. Its main function is to offer additional prior information for improved reconstruction, as mentioned in [158].

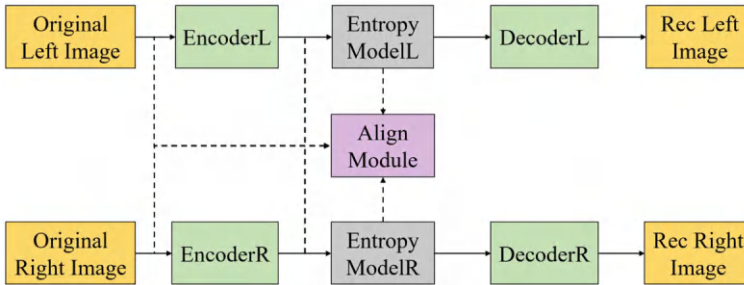


Fig. 6.2 The schematic diagram of stereo image compression frameworks. *Source* Author

6.2.3 Entropy Model

Conditional entropy model is widely adapted in stereo image compression. The equation can be expressed as [164]:

$$R = \mathbb{E}_{\tilde{\mathbf{y}}_1 \sim p_{\tilde{\mathbf{y}}_1 | \tilde{\mathbf{z}}_1}} \left[-\log_2 q_{\tilde{\mathbf{y}}_1 | \tilde{\mathbf{z}}_1}(\tilde{\mathbf{y}}_1 | \tilde{\mathbf{z}}_1) \right] + \mathbb{E}_{\tilde{\mathbf{y}}_2 \sim p_{\tilde{\mathbf{y}}_2 | \tilde{\mathbf{y}}_1, \tilde{\mathbf{z}}_2}} \left[-\log_2 q_{\tilde{\mathbf{y}}_2 | \tilde{\mathbf{y}}_1, \tilde{\mathbf{z}}_2}(\tilde{\mathbf{y}}_2 | \tilde{\mathbf{y}}_1, \tilde{\mathbf{z}}_2) \right], \quad (6.1)$$

where $q_{\tilde{\mathbf{y}}_1 | \tilde{\mathbf{z}}_1}$ and $q_{\tilde{\mathbf{y}}_2 | \tilde{\mathbf{y}}_1, \tilde{\mathbf{z}}_2}$ denote the conditional probability functions. p stands for true PMF. The image compression method models the PMF of $\tilde{\mathbf{y}}$ based on the hyper-prior $\tilde{\mathbf{z}}$. Because of the strong correlation of the left image and the right image, the entropy of $\tilde{\mathbf{y}}_2$ is expected to decrease when conditioned on $\tilde{\mathbf{y}}_1$, leading to a lower bit-rate. We can use context-based model or GMM-based model to estimate single probability mass function.

6.2.4 Related Datasets

InStereo2K [165]¹ The dataset encompasses 2050 pairs of stereo images, accompanied by highly accurate ground truth disparity maps. This collection comprises 2000 pairs designated for training and an additional 50 pairs designated for testing purposes. Notably, its integration into various networks, such as StereoNet and PSMNet, demonstrates a substantial enhancement in performance when applied to the Middlebury 2014 dataset. The InStereo2K dataset stands out due to its expansive scale, superior accuracy, and diverse content, offering researchers in the realm of stereo matching and related fields unprecedented opportunities. Its substantial size, coupled with meticulous ground truth annotations, opens avenues for advancements in stereo matching and extends the applicability of end-to-end stereo matching methods to practical scenarios (Fig. 6.3).

KITTI [166]² The dataset contains images acquired through two high-resolution color video cameras mounted on a conventional station wagon. The dataset comprises the 388 pairs of training images, and the 390 pairs of testing images, each with dimensions of 376×1240 pixels (Fig. 6.4).

Scene Flow [167]³ A substantial artificial dataset, of which a particular segment, FlyThings3D, encompasses the 22,390 pairs of training images, and the 4,370 pairs of testing images. The dimensions of each image within this subset are 540×960 . The selection of this subset for the dataset was deliberate, opting for FlyThings3D over the complete Scene Flow dataset due to its inclusion of a testing set, distinguishing it from other subsets in Scene Flow (Fig. 6.5).

¹ <https://github.com/YuhuaXu/StereoDataset>.

² <http://www.cvlibs.net/datasets/kitti/>.

³ <https://lmb.informatik.uni-freiburg.de/resources/datasets/SceneFlowDatasets.en.html>.



Fig. 6.3 Samples in InStereo2K dataset. Public domain open access image [165]



Fig. 6.4 Samples in KITTI dataset. Public domain open access image [166]



Fig. 6.5 Samples in scene flow dataset. Public domain open access image [167]

KAIST Multispectral Pedestrian Dataset [168]⁴ Comprising 95,000 pairs of color-thermal images, each with dimensions of 640×480 and recorded at a frequency 20Hz, the dataset is a comprehensive collection. It includes 41 sequences originating from 12 distinct sets. Additionally, for the purpose of multi-modality image compression assessments, a subset of 18 color-thermal image pairs is selected from the KAIST dataset. In this particular subset, the infrared images are resized to 320×240 (Fig. 6.6).

SADN [169]⁵ Comprising 7,549 light field images (LIs) captured by the Lytro Illum, this dataset surpasses the size of existing databases. Each image is meticulously annotated into 178 categories based on WordNet, including entities like cat, camel, bottle, fans, and more. The dataset contains registered depth maps for each

⁴ <http://multispectral.kaist.ac.kr/>.

⁵ <https://github.com/VincentChandelier/SADN>.



Fig. 6.6 Samples in KAIST dataset. Public domain open access image [168]



Fig. 6.7 Samples in SADN. Public domain open access image [169]

image, which can be useful for tasks such as depth estimation, 3D reconstruction, and image segmentation. The generation process involves processing the raw light field data captured by the camera using the Light Field Toolbox v0.4. Figure 6.7 displays select central Synthetic Aperture Images (SAIs) extracted from the light field images. Notably, more than 34% of the categories boast a collection of over 20 images, with 19 categories exceeding 100 images each. This extensive and diverse natural dataset proves invaluable for training a robust and stable algorithm.

6.3 Typical Methods of 3D and Multimodal Coding

6.3.1 Stereo Image Compression

Stereo image compression is the process of compressing images captured by two cameras or sensors. The core goal of stereo image compression is to reduce the storage demands or transmission bandwidth for image data while upholding the quality crucial for stereo vision. Modern stereo image compression algorithms rooted in deep learning methodologies are presented in Table 6.2.

Table 6.2 Summary of stereo image compression methods

Title	Authors	Publication	Contributions
DSIC: deep stereo image compression	Liu et al.	ICCV 2019	Proposing the incorporation of parametric skip functions for feeding disparity-warped features and modeling the probabilistic interdependence between images captured from the left and right views
L3C-stereo: lossless compression for stereo images	Huang et al.	arXiv 2021	Proposing a lossless compression model at multiple scales, comprising a warping module and a probability estimation module
SASIC: stereo image compression with latent shifts and stereo attention	Wödlinger et al.	CVPR 2022	Encoding only the residuals of the latent representation into the optimally modified the left latent representation
Deep stereo image compression via bi-directional coding	Lei et al.	CVPR 2022	Introducing a dual-directional contextual transform module (Bi-CTM) that conducts a nonlinear transformation by inter-view context, with the goal of effectively reducing redundancy between stereo views
Deep homography for efficient stereo image compression	Deng et al.	CVPR 2021	Introducing a network for compressing stereo images, founded on homography transformation, incorporating two conditional entropy models, and formulating a cross-quality enhancement module

Source Author

Liu et al. [170] proposed a deep learning-based stereo image compression network (DSIC) (shown in Fig. 6.8), utilizing deep learning techniques within a singular image compression model. This innovative approach introduces the parameter skip function and the conditional entropy model to effectively capture two modalities interdependence. Employing volumetric warping results in a warped feature map, which exhibits improved alignment by the second image feature map. The aggregation function serves as the parametric skip connection's terminal operation, generating the predicted feature map as its ultimate output. Consequently, any vertical shift can potentially disrupt the proper functioning of the DSIC method.

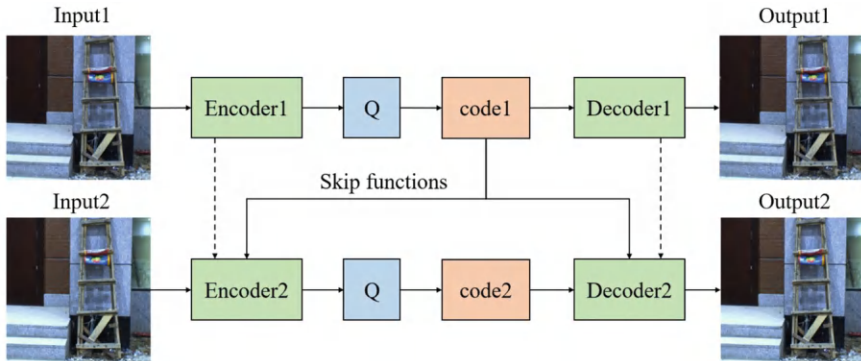


Fig. 6.8 The DSIC network architecture. Public domain open access image [170]

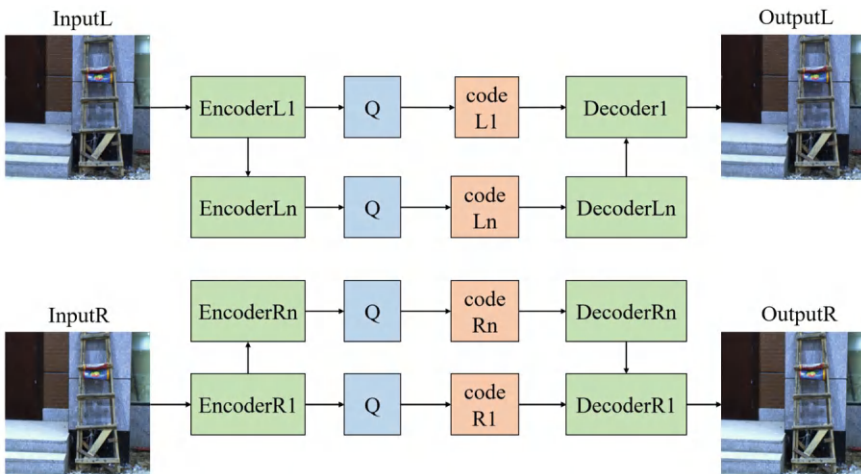


Fig. 6.9 The L3C-Stereo network architecture. Public domain open access image [171]

Huang et al. [171] proposed a novel model termed as L3C-Stereo (shown in Fig. 6.9) for lossless compression that uses deep learning technology to predict disparity maps from both views' features and includes a module for adaptive arithmetic coding.

Wödlinger et al. [172] presented a novel approach for image compression termed as SASIC (shown in Fig. 6.10), leveraging the similarity between left and right images derived from an overlapping field of view in stereo image pairs. This scheme employs the autoencoder model of hyperprior entropy, which compresses the left image. The compression process of the right image leverages data from the previously encoded left image throughout the encoding and decoding phases. It encodes solely the residual of its potential representation, reflecting the optimally adjusted version of the left image's representation.

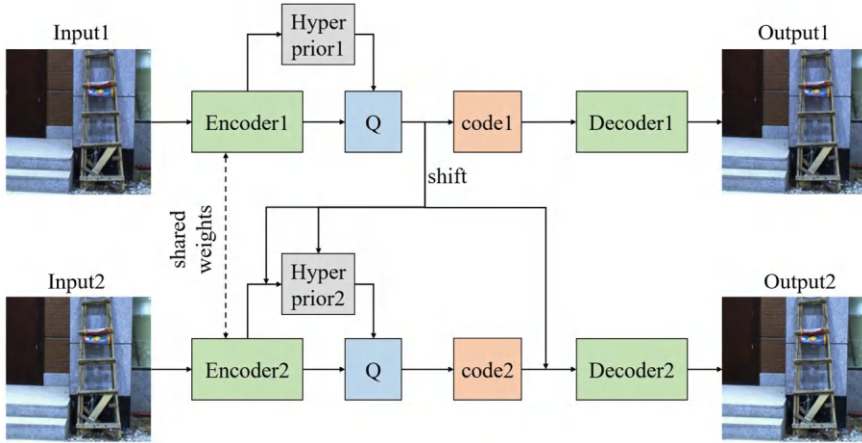


Fig. 6.10 The SASIC network architecture. ©2022 IEEE. Reprinted, with permission, from Ref. [172]

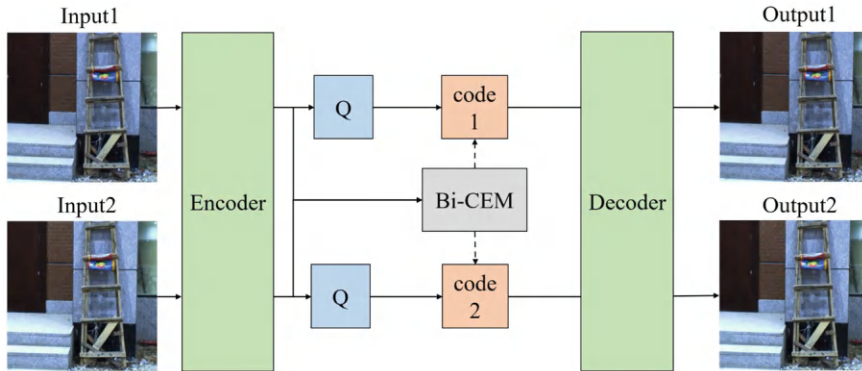


Fig. 6.11 The BCSIC-Net architecture. ©2022 IEEE. Reprinted, with permission, from Ref. [173]

Lei et al. [173] introduced an end-to-end stereo image compression network via bi-directional coding dubbed BCSIC-Net (shown as Fig. 6.11). BCSIC-Net incorporates a novel bidirectional context transform module designed to conduct a nonlinear transformation in the underlying space. This transformation is based on inter-view context, aiming to diminish inter-view redundancy. Moreover, a novel conditional entropy model is introduced, leveraging the correspondence of different view information as conditional priors to enhance coding efficiency.

Deng et al. [164] introduced a stereo image compression network based on homography estimation, dubbed as HESIC. In the HESIC method, the reference image is aligned to the target image using an estimated homography matrix. As a result, only the residual information from the target image requires encoding. The computational

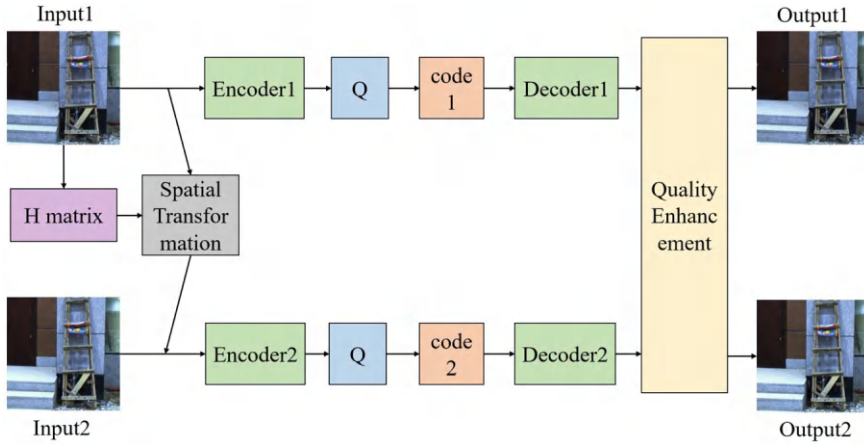


Fig. 6.12 The HESIC network architecture. ©2021 IEEE. Reprinted, with permission, from Ref. [164]

efficiency of homography matrix computation and coding in this scheme is noteworthy, placing minimal strain on the overall coding rate. The network structure diagram of HESIC is depicted in Fig. 6.12. To leverage the mutual information between the two stereo images, HESIC utilizes a deep regression model for estimating the homography matrix, denoted as the H-matrix. Subsequently, the spatial transformation using the H matrix is applied to the left image. Importantly, only the residual information needs the encoding, resulting in a more efficient use of the bit rate. This approach helps to capture and represent the differences between the two stereo images while minimizing redundant information, contributing to more effective compression.

6.3.2 Multi-modality Image Compression

Multi-modality image compression refers to advanced techniques that simultaneously process and compress multiple correlated imaging formats by exploiting their cross-modal redundancies while preserving modality-specific features. These methods typically employ joint encoding architectures and alignment modules to achieve superior compression efficiency compared to independent single-modality approaches. Some representative methods are listed in Table 6.3.

Jiang et al. [174] presented a novel method employing multi-modal machine learning techniques (shown in Fig. 6.13) to improve text-guided image compression. This approach uses text semantic information as a guiding factor during image compression, resulting in a notable enhancement in compression performance. The study examines the role of text descriptions in various codec segments and verifies their efficacy. Additionally, the research introduces an image-text focus module

Table 6.3 Summary of multi-modality image compression methods

Title	Authors	Publication	Contributions
Multi-modality deep network for extreme learned image compression	Jiang et al.	arXiv 2023	Examining the influence of text descriptions on various codec components, integrating image-text attention (ITA) module to merge image and text features, and suggesting image-request complement (IRC) module to dynamically acquire essential guidance knowledge from textual information
Learning based multi-modality image and video compression	Lu et al.	CVPR 2022	Presenting alignment modules that operate both channel-wise and spatial-wise, effectively harnessing correlations between diverse modalities within the feature space
End-to-end RGB-D image compression via exploiting channel-modality redundancy	Zheng et al.	AAAI 2024	Introducing a two branch RGB-D image compression framework, featuring IMA and CMA to eliminate channel and modality redundancies
Your camera improves your point cloud compression	Lin et al.	ICASSP 2023	Introducing the first framework for multi-modality compression dedicated to point cloud data and enhancing the compression efficiency of LiDAR point clouds by leveraging cross-modal information from cameras

Source Author

and an image-request supplement module to strengthen the integration of image and text features. For comprehensive semantic reconstruction, an enhanced multi-modal semantic consistency loss function is proposed. Through a series of comprehensive experiments, the results indicate that their method produces visually compelling outcomes even at extremely low bit rates. Moreover, the performance of this method is comparable to, and in certain instances, surpasses state-of-the-art technologies, even when these technologies operate at two to four times the bit rate of alternative methods.

Lu et al. [175] proposed an innovative multi-modal compression framework (shown in Fig. 6.14) tailored for paired infrared and visible images, harnessing cross-modal redundancy. The framework functions by utilizing a reference mode image (e.g., infrared) to generate aligned features through a channel alignment module, employing affine transformations. These aligned features serve as contextual information for compressing the current mode image, such as the visible image, while

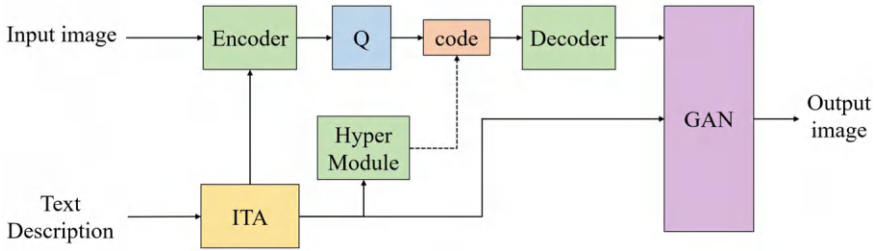


Fig. 6.13 The text-guided image compression (TGIC) model. Public domain open access image [174]

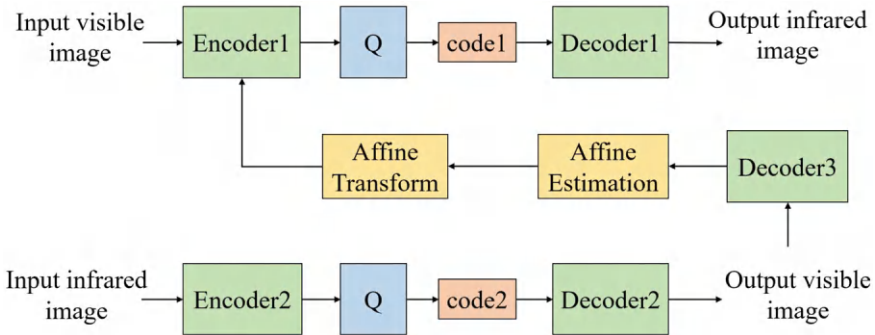


Fig. 6.14 The multi-modality compression framework. ©2022 IEEE. Reprinted, with permission, from Ref. [175]

the corresponding affine coefficients are compressed at minimal cost. Additionally, the framework integrates a Transformer-based spatial alignment module to leverage correlations among intermediate features during the decoding of various modes. Importantly, this framework exhibits adaptability and can be easily extended to scenarios involving multi-modal video compression. Experimental results underscore the framework's superiority over both traditional methods and learning-based single-mode compression techniques when assessed on FLIR and KAIST datasets. The proposed approach surpasses existing methods, demonstrating its effectiveness in multi-modal image compression.

Zheng et al. [93] proposed a dual-branch framework for RGB-D image compression using a learning-based approach (shown in Fig. 6.15). It substitutes traditional RGB techniques with a YUV domain scheme to reduce spatial redundancy. It incorporates intra-modal and cross-modal attention mechanisms to tackle modal redundancy. To utilize inter-modal prior knowledge, the framework features context prediction and fusion modules, enhancing the accuracy of context probability estimates. Experimental results demonstrate that this approach surpasses the recently methods on two datasets, achieving over 15% bit rate savings for RGB images compared to BPG.

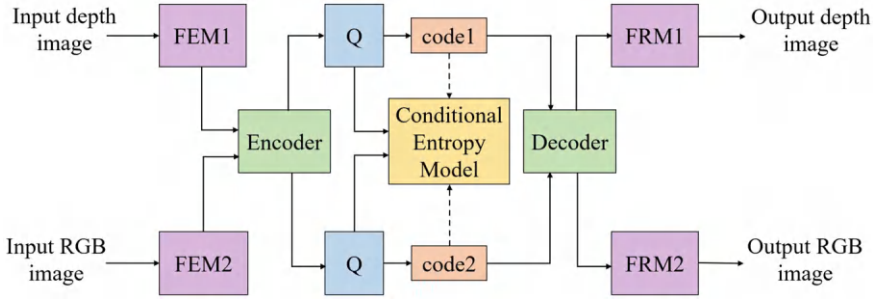


Fig. 6.15 The RGB-D compression framework. Public domain open access image [93] (<https://github.com/VincentChandelier/SADN>)

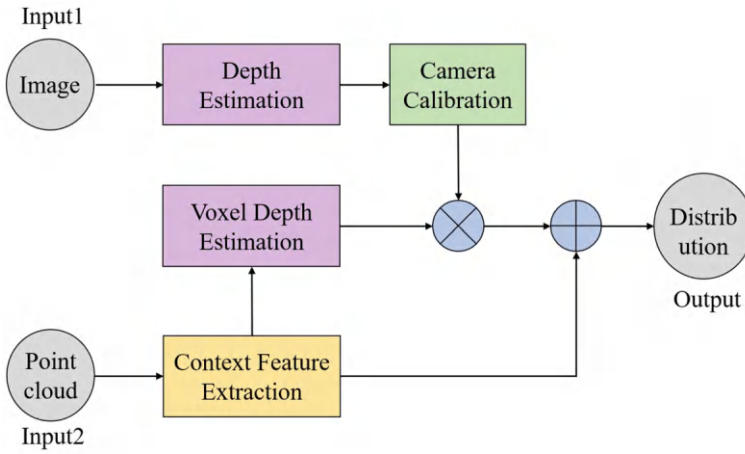


Fig. 6.16 The multi-modality compression framework. ©2023 IEEE. Reprinted, with permission, from Ref. [176]

Lin et al. [176] introduced an innovative multimodal compression framework (shown in Fig. 6.16) specifically designed for LiDAR point clouds, incorporating depth information to predict their associated images. Remarkably, their model represents the pioneering multimodal compression framework tailored for point clouds. The methodology initiates by utilizing an octree-based approach to represent point clouds, effectively mitigating spatial redundancy [83, 86, 91, 160, 161].⁶ Subsequently, a cross-modal fusion structure is introduced to enhance the compression of these octrees. This fusion structure extracts depth distributions from camera pixels, serving as supplementary information. In comparison to previous state-of-the-art (SOTA) techniques, their approach demonstrates an improvement in compression rates of up to 8.10% in LiDAR point cloud compression. This signifies a significant

⁶ <https://ojs.aaai.org/index.php/AAAI/article/view/28588>.

advancement over existing methods, underscoring the effectiveness of their proposed framework in compressing LiDAR point clouds.

6.3.3 Light Field Image Compression

Light field image compression refers to specialized techniques that exploit the inherent redundancy among multiple subviews of a light field to achieve efficient data reduction. These methods typically employ sparse sampling of subviews combined with deep learning-based reconstruction, often incorporating multi-view enhancement networks to maintain geometric accuracy while minimizing compression artifacts. Some representative methods are listed in Table 6.4.

Zhao et al. [177] proposed a compression scheme (shown in Fig. 6.17) for optical field images that capitalizes on the intrinsic similarity within such images through the integration of deep learning techniques. Diverging from the conventional approach of transmitting the complete set of Light Field subviews, the method strategically transmits sparsely sampled LF subviews. Utilizing Convolutional Neural Networks, the

Table 6.4 Summary of light field image compression methods

Title	Authors	Publication	Contributions
Light field image compression based on deep learning	Zhao et al.	ICME 2018	Introducing an innovative compression scheme for light field images that capitalizes on the inherent similarity within the images through the application of deep learning techniques
View synthesis-based light field image compression using a generative adversarial network	Liu et al.	Information Sciences 2021	Introducing a method for compressing light field images based on view synthesis, employing a Generative Adversarial Network (GAN) in the compression process
SADN: learned light field image compression with spatial-angular decorrelation	Tong et al.	arXiv 2022	Introducing an innovative end-to-end Spatial-Angular-Decorrelated Network (SADN) aimed at achieving highly efficient compression of light field images
Learning kernel-modulated neural representation for efficient light field compression	Shi et al.	arXiv 2023	Presenting mechanisms for modulator allocation and kernel tensor decomposition to preserve the high quality of the decoded light field

Source Author

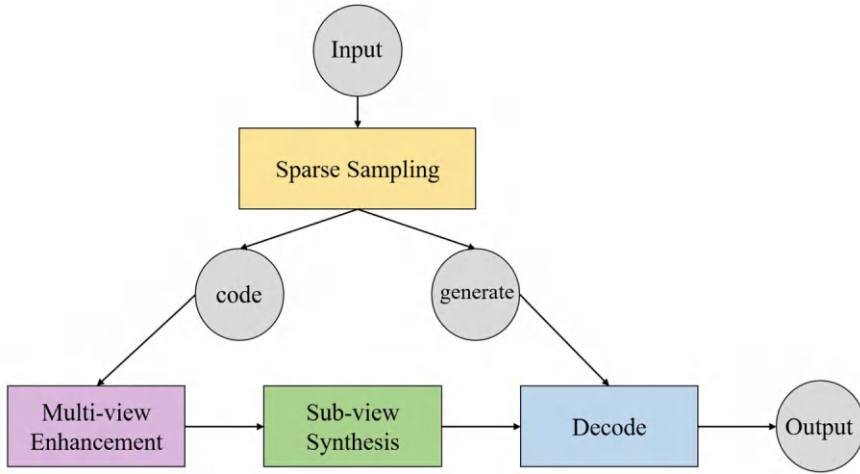


Fig. 6.17 The deep learning-based light field image compression framework. ©2018 IEEE. Reprinted, with permission, from Ref. [177]

system reconstructs the missing subviews by leveraging encoded information from neighboring views. To mitigate artifacts arising from the compression process and ensure high geometric precision in reconstructing the uncoded view, they introduce a multi-view joint enhancement network. Through an extensive series of experiments, the scheme exhibits notably superior compression performance compared to the most advanced existing methods. This underscores the effectiveness and efficiency of the approach in compressing optical field images.

Liu et al. [178] proposed a Light Field image compression method focused on view synthesis, leveraging Generative Adversarial Networks (GAN) to address this challenge. The fundamental concept behind the compression revolves around the substantial correlation among adjacent subaperture images. Consequently, the transmission involves only sparsely sampled subaperture images (SAI), with the remaining SAIs being reconstructed within the decoder. The paper introduces a sparse SAI sampling technique, enhancing the quality of reconstructed SAIs by carefully balancing SAI redundancy against the available number of SAIs for synthesis. The refinement of reconstructed SAIs involves a GAN-based synthesis process that decomposes synthesis into parallax estimation and unsampled SAI estimation components. Moreover, the adversarial training of the generative and discriminative networks within the GAN framework enhances the accuracy of the generated model. Furthermore, integrating a quality-based loss function helps retain more texture details in the synthesized SAIs. Comprehensive experimental results reveal that this approach outperforms various leading compression techniques, showing improvements in both standard quality metrics and the perceived quality of synthesized SAIs at the decoder side.

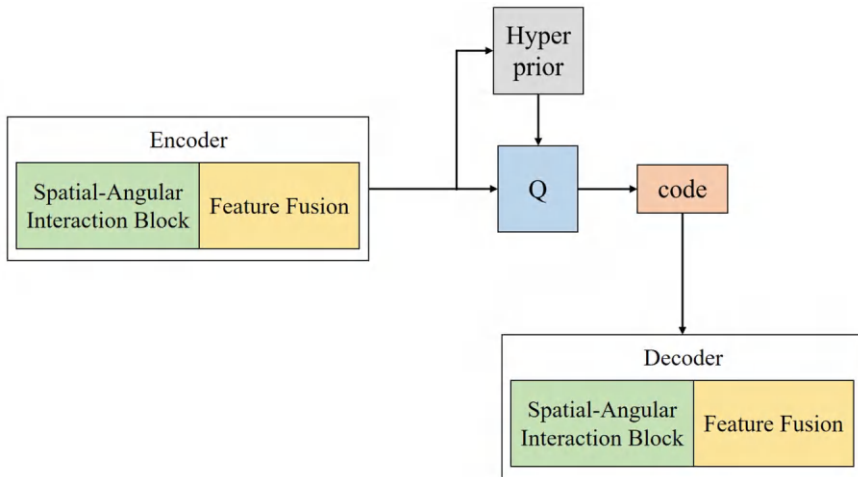


Fig. 6.18 Overview of the SADN architecture. Public domain open access image [179]

Tong et al. [179] developed the End-to-End Spatial Angle Declination Network (SADN), a novel method for efficiently compressing light field images. Unlike traditional approaches that focus on spatial or angular consistency, SADN uniquely separates angular information from spatial details using expanded and step convolutions. This separation facilitates feature fusion, allowing the concurrent compression of spatial and angular information through spatial-angle interaction. To ensure the stability and robustness of the algorithm, a substantial dataset⁷ comprising 7549 light field images was meticulously curated for comprehensive training. In comparison to industry standards such as H.266/VVC and H.265/HEVC, this method demonstrates a significant enhancement in compression efficiency, outperforming these benchmarks by 2.137 times and 2.849 times, respectively. Additionally, it accomplishes an average bit rate reduction of 79.6% relative to the other end-to-end image compression networks, all while ensuring enhanced subjective quality and maintaining light field consistency. The architecture is shown in Fig. 6.18.

Shi et al. [180] introduced an innovative method for light field compression (shown in Fig. 6.19) by leveraging the visual characteristics of Sub-Aperture Images (SAIs) within a concise neural network framework. The network architecture, inspired by optical field features, employs randomly initialized noise as input and trains to supervise the spatial-angular resolution (SAR) of the targeted light field. This network consists of two distinct nuclei: descriptive nuclei, capturing scene description information learned during training, and modulators, regulating the presentation of various SAIs from the queried perspective. To enhance the network's compactness without compromising the quality of the decoded light field, the paper introduces modulator allocation and nuclear tensor decomposition mechanisms. Additionally, non-uniform

⁷ <https://github.com/VincentChandelier/SADN>.

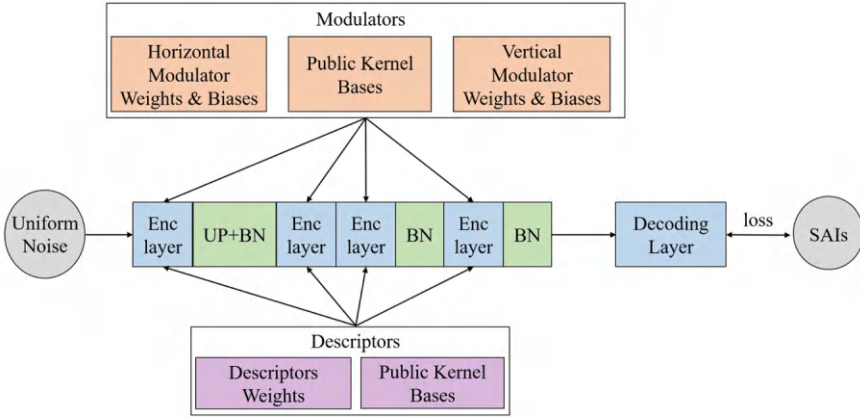


Fig. 6.19 The light field image compression framework. Public domain open access image [180]

quantization and lossless entropy coding techniques are incorporated, forming an efficient compression pipeline. The method shows excellent performance compared to the other methods in light field compression tasks. Furthermore, after aligning descriptors, a modulator learned from one light field can transfer to a new light field for comprehensive view synthesis, offering a potential solution for view synthesis tasks. This transferability indicates the adaptability and potential applicability of the method beyond compression, contributing to denser view generation for diverse optical field applications.

6.3.4 360° Image Compression

Spherical image compression involves specialized coding techniques that adaptively allocate bits across different latitude regions while maintaining spherical geometry through domain-specific convolutional operations. Some representative methods are listed in Table 6.5.

Li et al. [181] devised a novel approach (shown in Fig. 6.20) termed latitude adaptive coding for deep neural networks, tailoring code allocation based on the specific latitude in a spherical representation. This strategy involves assigning varying quantities of codes to distinct regions, considering the region's assigned codes and entropy. A dynamic area-specific rate loss is presented, allowing for rate control across different sections. Additionally, location-based constraints are applied to limit excessive code usage in oversampled regions. The authors also utilize frame-based distortion losses, averaging distortion across various frames. The model is refined and tested using a comprehensive 360° dataset⁸ of 19,790 images sourced online. The

⁸ <https://github.com/limuhit/360-Image-Compression>.

Table 6.5 Summary of 360° image compression methods

Title	Authors	Publication	Contributions
End-to-end optimized 360° image compression	Li et al.	TIP 2022	Presenting the first end-to-end optimized image compression codec specifically designed for 360° images, along with the creation of a large, high-quality dataset for compressing such images
OSLO: on-the-sphere learning for omnidirectional images and its application to 360° image compression	Bidgoli et al.	arXiv 2021	Introducing an innovative learning approach for spherical images and Defining a novel convolution operation on the sphere that retains high expressiveness while ensuring low complexity, similar to conventional 2D convolution

Source Author

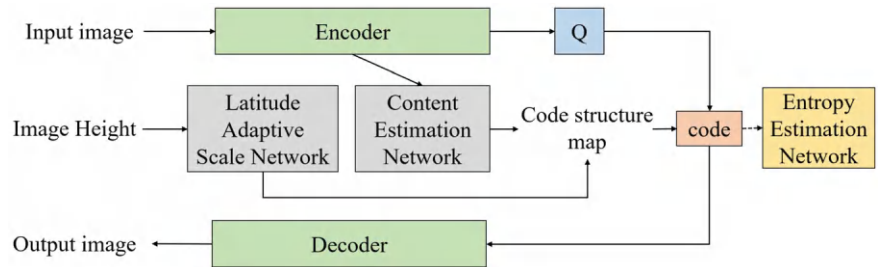


Fig. 6.20 Overview of the 360° image compression architecture. ©2022 IEEE. Reprinted, with permission, from Ref. [181]

effectiveness of the proposed location-adaptive coding method is validated. Overall, this model exceeds existing image compression standards, such as JPEG, JPEG2000, HEVC, and VVC, achieving roughly 15% bit savings compared to traditional models for flat images.

Bidgoli et al. [182] introduced a novel convolution operation tailored for the spherical domain, aiming to preserve expressive power while maintaining low complexity similar to classical 2D convolution. By leveraging standard Convolutional Neural Network (CNN) techniques such as sequential processing, repeated aggregation, and pixel manipulation, these methods adapt for the spherical domain. This novel framework is utilized for omni-directional image compression tasks. Experimental results demonstrate that the proposed spherical method achieves significant compression improvements, reducing the bit rate by 13.7% compared to a similar learning model applied to rectangular images. Furthermore, in contrast to graph convolutional network-based approaches, the proposed solution utilizes more advanced filters that maintain finer details, resulting in enhanced visual quality of the compressed images.

This outcome not only validates the effectiveness of the framework but also paves the way for a new research field, enabling the effective implementation of various omnidirectional vision tasks on spherical manifolds.

6.4 Summary

This chapter examines typical coding methods for 3D and multimodal data, addressing the specific needs of stereo and multi-modality image compression. In addition, it discusses specialized compression techniques for light field and 360° images, with the aim of maintaining high-quality visual experiences despite the increased data complexity. The focus is on achieving efficient compression while preserving perceptual quality for diverse applications.

Exercises

1. Please explain what is learned stereo image compression.
2. Please explain how deep learning-based compression differ from traditional methods.
3. Please explain the advantages of learned stereo image compression.
4. Please explain the challenges of learned stereo image compression.
5. Please explain how deep learning models handle the stereo disparity in stereo images during compression.
6. Please explain what types of deep learning architectures are commonly used for stereo image compression.
7. Please explain how do researchers evaluate the performance of learned stereo image compression algorithms.
8. Please explain how deep learning-based compression can be applied in real-time applications.
9. Please describe how learned stereo image compression can be further improved.
10. Please describe the potential future directions for learned stereo image compression research.

References

1. X. Zang, W. Gao, G. Li, H. Fang, C. Ban, Z. He, H. Sun, A baseline investigation: Transformer-based cross-view baseline for text-based person search, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 7737–7746
2. G. Liao, W. Gao, G. Li, J. Wang, S. Kwong, Cross-collaborative fusion-encoder network for robust rgb-thermal salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **32**(11), 7646–7661 (2022)

3. W. Gao, G. Liao, S. Ma, G. Li, Y. Liang, W. Lin, Unified information fusion network for multi-modal rgb-d and rgb-t salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **32**(4), 2091–2106 (2021)
4. Y. Chen, S. Sun, G. Li, W. Gao, T.H. Li, Closing the gap between theory and practice during alternating optimization for gans. *IEEE Trans. Neural Netw. Learn. Syst.* (2023)
5. Y. Chen, C. Jin, G. Li, T.H. Li, W. Gao, Mitigating label noise in gans via enhanced spectral normalization. *IEEE Trans. Circuits Syst. Video Technol.* (2023)
6. X. Zang, G. Li, W. Gao, Multidirection and multiscale pyramid in transformer for video-based pedestrian retrieval. *IEEE Trans. Ind. Inform.* **18**(12), 8776–8785 (2022)
7. X. Zang, G. Li, W. Gao, X. Shu, Learning to disentangle scenes for person re-identification. *Image Vis. Comput.* **116**, 104330 (2021)
8. X. Zang, G. Li, W. Gao, X. Shu, Exploiting robust unsupervised video person re-identification. *IET Image Process.* **16**(3), 729–741 (2022)
9. Z. Yue, G. Li, W. Gao, Cross-level guided attention for human-object interaction detection, in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2023), pp. 284–289
10. S. Sun, J. Liu, H. Li, G. Liu, T.H. Li, W. Gao, Streamflow: streamlined multi-frame optical flow estimation for video sequences, in *Advances in Neural Information Processing Systems* (2024)
11. Z. Yao, W. Gao, Iterative saliency aggregation and assignment network for efficient salient object detection in optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* (2024)
12. G. Liao, W. Gao, Rethinking feature mining for light field salient object detection. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(10), 1–24 (2024)
13. B. Yang, W. Gao, G. Li, Focus and adjust: Progressive refinement network for human object interaction detection, in *2022 26th International Conference on Pattern Recognition (ICPR)* (IEEE, 2022), pp. 2546–2552
14. G. Liao, W. Gao, Q. Jiang, R. Wang, G. Li, Mmnet: multi-stage and multi-scale fusion network for rgb-d salient object detection, in *Proceedings of the 28th ACM International Conference on Multimedia* (2020), pp. 2436–2444
15. W. Gao, S. Kwong, Y. Zhou, Y. Jia, J. Zhang, W. Wu, Multiscale phase congruency analysis for image edge visual saliency detection, in *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 1 (IEEE, 2016), pp. 75–80
16. W. Gao, S. Kwong, Y. Zhou, X. Wang, Phase congruency analysis of down-sampled and blurring images for foreground extraction, in *2015 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)* (IEEE, 2015), pp. 152–157
17. Y. Guo, W. Gao, G. Li, Interpretable task-inspired adaptive filter pruning for neural networks under multiple constraints. *Int. J. Comput. Vis.* 1–17 (2024)
18. W. Gao, Y. Guo, S. Ma, G. Li, S. Kwong, Efficient neural network compression inspired by compressive sensing. *IEEE Trans. Neural Netw. Learn. Syst.* (2022)
19. Y. Guo, W. Gao, Semantic-driven automatic filter pruning for neural networks, in *IEEE International Conference on Multimedia and t(ICME)* (IEEE, 2022), pp. 1–6
20. L. Tao, W. Gao, Efficient channel pruning based on architecture alignment and probability model bypassing, in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2021), pp. 3232–3237
21. Y. Li, Z. Yu, X. Chen, S. Wang, H. Yuan, W. Gao, Zero-shot quantization for large-kernels via shape-based distribution and diversity self-distillation, in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2025)
22. S. Fan, W. Gao, G. Li, Salient object detection for point clouds, in *European Conference on Computer Vision* (Springer, 2022), pp. 1–19
23. S. Luo, W. Gao, A general framework for rotation invariant point cloud analysis, in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2024), pp. 3665–3669
24. X. Lu, W. Gao, Attentivenet: Detecting small objects for lidar point clouds by attending to important points, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5

25. Z. Pan, N. Zhang, W. Gao, S. Liu, G. Li, Less is more: label recommendation for weakly supervised point cloud semantic segmentation. *Proc. AAAI Conf. Artif. Intell.* **38**(5), 4397–4405 (2024)
26. Z. Pan, G. Liu, W. Gao, T.H. Li, Epcontrast: effective point-level contrastive learning for large-scale point cloud understanding, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024), pp. 1–6
27. N. Zhang, Z. Pan, T.H. Li, W. Gao, G. Li, Improving graph representation for point cloud segmentation via attentive filtering, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 1244–1254
28. K. Wen, N. Zhang, G. Li, W. Gao, Mpvnn: multi-resolution point-voxel non-parametric network for 3d point cloud processing, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024), pp. 1–6
29. Z. Pan, N. Zhang, W. Gao, S. Liu, G. Li, Point cloud semantic segmentation with sparse and inhomogeneous annotations, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
30. R. Zhang, G. Li, W. Gao, S. Liu, A quantum-inspired framework in master-servant mode for large-scale multi-modal place recognition. *IEEE Trans. Intell. Transp. Syst.* (2024)
31. D. Yang, W. Gao, Pointchd: a point cloud benchmark for congenital heart disease classification and segmentation. *IEEE J. Biomed. Health Inform.* (2024)
32. S. Fan, W. Gao, G. Li, Point-mpp: point cloud self-supervised learning from masked position prediction. *IEEE Trans. Neural Netw. Learn. Syst.* (2024)
33. Z. Pan, W. Gao, G. Li, Distribution guidance network for weakly supervised point cloud semantic segmentation, in *Advances in Neural Information Processing Systems* (2024)
34. R. Zhang, G. Li, W. Gao, T.H. Li, Compoint: can complex-valued representation benefit point cloud place recognition? *IEEE Trans. Intell. Transp. Syst.* (2024)
35. W. Gao, G. Li, *Deep Learning for 3D Point Clouds* (Springer Nature, 2024)
36. W. Gao, G. Li, Introduction to 3d point clouds: datasets and perception, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 1–27
37. W. Gao, G. Li, Learning basics for 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 29–70
38. W. Gao, G. Li, Deep-learning-based point cloud enhancement i, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 71–97
39. W. Gao, G. Li, Deep-learning-based point cloud enhancement ii, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 99–130
40. W. Gao, G. Li, Deep-learning-based point cloud analysis i, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 131–162
41. W. Gao, G. Li, Deep-learning-based point cloud analysis ii, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 163–193
42. W. Gao, G. Li, Point cloud pre-trained models and large models, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 195–225
43. W. Gao, G. Li, Point cloud-language multi-modal learning, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 227–254
44. W. Gao, G. Li, Open-source projects for 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 255–272
45. W. Gao, G. Li, Typical engineering applications of 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2025), pp. 273–299
46. W. Gao, G. Li, Future work on deep learning-based point cloud technologies, in *Deep Learning for 3D Point Clouds* (Springer, 2025), pp. 301–315
47. G. Li, G. Wei, W. Gao, *Point Cloud Compression: Technologies and Standardization* (Springer Nature, 2024)
48. G. Li, W. Gao, W. Gao, Introduction, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 1–28
49. G. Li, W. Gao, W. Gao, Background knowledge, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 29–51

50. G. Li, W. Gao, W. Gao, Predictive coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 53–70
51. G. Li, W. Gao, W. Gao, Transform coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 71–96
52. G. Li, W. Gao, W. Gao, Quantization techniques, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 97–112
53. G. Li, W. Gao, W. Gao, Entropy coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 113–133
54. G. Li, W. Gao, W. Gao, Mpeg geometry-based point cloud compression (g-pcc) standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 135–165
55. G. Li, W. Gao, W. Gao, Avs point cloud compression standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 167–197
56. G. Li, W. Gao, W. Gao, Mpeg video-based point cloud compression (v-pcc) standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 199–218
57. G. Li, W. Gao, W. Gao, Mpeg ai-based 3d graphics coding standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 219–241
58. G. Li, W. Gao, W. Gao, Future work, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 243–250
59. T. Qin, G. Li, W. Gao, S. Liu, Multi-grained point cloud geometry compression via dual-model prediction with extended octree. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(9), 1–30 (2024)
60. Y. Shao, W. Gao, S. Liu, G. Li, Advanced patch-based affine motion estimation for dynamic point cloud geometry compression. *Sensors* **24**(10), 3142 (2024)
61. Y. Shao, F. Song, W. Gao, S. Liu, G. Li, Texture-guided graph transform optimization for point cloud attribute compression. *Appl. Sci.* **14**(10), 4094 (2024)
62. Y. Shao, X. Yang, W. Gao, S. Liu, G. Li, 3d point cloud attribute compression using diffusion-based texture-aware intra prediction. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
63. J. Zhang, Y. Chen, G. Liu, W. Gao, G. Li, Efficient point cloud attribute compression framework using attribute-guided graph fourier transform, in *ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2024), pp. 8426–8430
64. W. Gao, H. Yuan, G. Li, Z. Li, H. Yuan, Low complexity coding unit decision for video-based point cloud compression. *IEEE Trans. Image Process.* **33**, 149–162 (2023)
65. Y. Shao, G. Li, Q. Zhang, W. Gao, S. Liu, Nonrigid registration-based progressive motion compensation for point cloud geometry compression. *IEEE Trans. Geosci. Remote Sens.* (2023)
66. F. Song, G. Li, X. Yang, W. Gao, S. Liu, Block-adaptive point cloud attribute coding with region-aware optimized transform. *IEEE Trans. Circuits Syst. Video Technol.* (2023)
67. Y. An, Y. Shao, G. Li, W. Gao, S. Liu, A fast motion estimation method with hamming distance for lidar point cloud compression, in *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2022), pp. 1–5
68. H. Yuan, W. Gao, G. Li, Z. Li, Rate-distortion-guided learning approach with cross-projection information for v-pcc fast cu decision, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 3085–3093
69. F. Song, G. Li, W. Gao, T.H. Li, Rate-distortion optimized graph for point cloud attribute coding. *IEEE Signal Process. Lett.* **29**, 922–926 (2022)
70. F. Song, G. Li, X. Yang, W. Gao, T.H. Li, Fine-grained correlation representation for graph-based point cloud attribute compression, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
71. F. Shen, W. Gao, A rate control algorithm for video-based point cloud compression, in *2021 International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2021), pp. 1–5
72. F. Song, Y. Shao, W. Gao, H. Wang, T. Li, Layer-wise geometry aggregation framework for lossless lidar point cloud compression. *IEEE Trans. Circuits Syst. Video Technol.* **31**(12), 4603–4616 (2021)

73. Z. Cai, W. Gao, G. Li, W. Gao, Distortion propagation model-based v-pcc rate control for 3d point cloud broadcasting. *IEEE Trans. Broadcast.* 1–13 (2024)
74. X. Mu, W. Gao, H. Yuan, S. Wang, G. Li, Adaptive lpu decision for dynamic point cloud compression. *IEEE Signal Process. Lett.* (2024)
75. C. Peng, W. Gao, Laplacian matrix learning for point cloud attribute compression with ternary search-based adaptive block partition, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 10 412–10 420
76. Y. Wang, W. Gao, X. Mu, H. Yuan, Rate control optimization for joint geometry and attribute coding of lidar point clouds, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
77. T. Guo, H. Yuan, T. Wang, W. Gao, Graph filter-based fast motion matching for inter frame coding of mpeg g-pcc, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 1151–1155
78. L. Xie, W. Gao, H. Zheng, G. Li, Spcgc: scalable point cloud geometry compression for machine vision, in *2024 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2024), pp. 17 272–17 278
79. L. Xie, W. Gao, H. Zheng, H. Ye, Semantic-aware visual decomposition for point cloud geometry compression, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 595–595
80. Z. Qi, W. Gao, Variable-rate point cloud geometry compression based on feature adjustment and interpolation, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 63–72
81. Z. Yu, W. Gao, When dynamic neural network meets point cloud compression: Computation-aware variable rate and checkerboard context, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 600–600
82. L. Xie, W. Gao, S. Fan, Z. Yao, Pdnet: parallel dual-branch network for point cloud geometry compression and analysis, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 596–596
83. L. Xie, W. Gao, H. Zheng, End-to-end point cloud geometry compression and analysis with sparse tensor, in *International Workshop on Advances in Point Cloud Compression, Processing and Analysis* (2022), pp. 27–32
84. C. Fu, G. Li, R. Song, W. Gao, S. Liu, Octattention: octree-based large-scale contexts model for point cloud compression. *Proc. AAAI Conf. Artif. intell.* **36**(1), 625–633 (2022)
85. C. Zhang, W. Gao, Adapcc: adaptive rate control and rate-distortion-complexity optimization for dynamic point cloud compression, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
86. K. Wang, W. Gao, Unipccg: towards practical point cloud geometry compression via an efficient unified approach, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
87. K. Zheng, W. Gao, H. Zheng, Octree-based learned point cloud geometry compression: a lossy perspective, in *2025 Data Compression Conference (DCC)* (IEEE, 2025)
88. Z. Qi, H. Liu, W. Gao, Variable-rate point cloud compression with channel-wise entropy modeling, in *2025 Data Compression Conference (DCC)* (IEEE, 2025)
89. X. Chen, Y. Wang, Y. Li, X. Ling, M. Li, R. Liu, M. Ouyang, K. Zhao, T. Guan, Y. He, Low bit-width zero-shot quantization with soft feature-infused hints for iot systems. *IEEE Internet Things J.* (2024)
90. H. Zheng, W. Gao, Z. Yu, T. Zhao, G. Li, Viewpccg: view-guided learned point cloud geometry compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 7152–7161
91. L. Xie, W. Gao, H. Zheng, G. Li, Roi-guided point cloud geometry compression towards human and machine vision, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 3741–3750
92. Y. Wu, L. Xie, S. Sun, W. Gao, Y. Yan, Adaptive intra period size for deep learning-based screen content video coding, in *IEEE International Conference on Multimedia and Expo Workshops* (2024), pp. 1–6

93. H. Zheng, W. Gao, End-to-end rgb-d image compression via exploiting channel-modality redundancy. *Proc. AAAI Conf. Artif. Intell.* **38**(7), 7562–7570 (2024)
94. L. Tao, W. Gao, G. Li, C. Zhang, Adanic: towards practical neural image compression via dynamic transform routing, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 16 879–16 888
95. Y. Wu, W. Gao, End-to-end lossless compression of high precision depth maps guided by pseudo-residual (2022). [arXiv:2201.03195](https://arxiv.org/abs/2201.03195)
96. Y. Wu, Z. Qi, H. Zheng, L. Tao, W. Gao, Deep image compression with latent optimization and piece-wise quantization approximation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 1926–1930
97. W. Gao, L. Tao, L. Zhou, D. Yang, X. Zhang, Z. Guo, Low-rate image compression with super-resolution learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2020), pp. 154–155
98. W. Gao, S. Sun, H. Zheng, Y. Wu, H. Ye, Y. Zhang, Opendmc: an open-source library and performance evaluation for deep-learning-based multi-frame compression, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9685–9688
99. W. Gao, H. Zheng, C. Zhang, K. Zheng, Z. Yu, Y. Li, H. Ye, Y. Zhang, Opendic: an open-source library and performance evaluation for deep-learning-based image compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 202–11 205
100. Y. Li, W. Gao, G. Li, S. Ma, Saliency segmentation oriented deep image compression with novel bit allocation. *IEEE Trans. Image Process.* **34**, 16–29 (2025)
101. C. Zhang, W. Gao, Learned rate control for frame-level adaptive neural video compression via dynamic neural network, in *European Conference on Computer Vision* (Springer, 2024)
102. Z. Yang, W. Gao, G. Li, Y. Yan, Sur-driven video coding rate control for jointly optimizing perceptual quality and buffer control. *IEEE Trans. Image Process.* (2023)
103. F. Shen, Z. Cai, W. Gao, An efficient rate control algorithm for intra frame coding in avs3, in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2021), pp. 3164–3169
104. H. Yuan, W. Gao, J. Wang, Dynamic computational resource allocation for fast inter frame coding in video conferencing applications, in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6
105. W. Gao, Q. Jiang, R. Wang, S. Ma, G. Li, S. Kwong, Consistent quality oriented rate control in hevc via balancing intra and inter frame coding. *IEEE Trans. Ind. Inform.* **18**(3), 1594–1604 (2021)
106. H. Yuan, W. Gao, A new coding unit partitioning mode for screen content video coding, in *Proceedings of the 2021 5th International Conference on Digital Signal Processing* (2021), pp. 66–72
107. W. Gao, On the performance evaluation of state-of-the-art rate control algorithms for practical video coding and transmission systems, in *Proceedings of the 2020 4th International Conference on Video and Image Processing* (2020), pp. 179–185
108. W. Gao, S. Kwong, Q. Jiang, C.-K. Fong, P.H. Wong, W.Y. Yuen, Data-driven rate control for rate-distortion optimization in hevc based on simplified effective initial qp learning. *IEEE Trans. Broadcast.* **65**(1), 94–108 (2018)
109. W. Gao, A multi-objective optimization perspective for joint consideration of video coding quality, in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (IEEE, 2019), pp. 986–991
110. W. Gao, S. Kwong, Y. Jia, Joint machine learning and game theory for rate control in high efficiency video coding. *IEEE Trans. Image Process.* **26**(12), 6074–6089 (2017)
111. W. Gao, S. Kwong, Y. Zhou, H. Yuan, Ssim-based game theory approach for rate-distortion optimized intra frame ctu-level bit allocation. *IEEE Trans. Multimedia* **18**(6), 988–999 (2016)
112. W. Gao, S. Kwong, H. Yuan, X. Wang, Dct coefficient distribution modeling and quality dependency analysis based frame-level bit allocation for hevc. *IEEE Trans. Circuits Syst. Video Technol.* **26**(1), 139–153 (2015)

113. W. Gao, S. Kwong, Phase congruency based edge saliency detection and rate control for perceptual image and video coding, in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2016), pp. 000 264–000 269
114. H. Yuan, W. Gao, Openfastvc: an open source library for video coding fast algorithm implementation, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9660–9663
115. H. Yuan, W. Gao, S. Ma, Y. Yan, Divide-and-conquer-based rdo-free cu partitioning for 8k video compression. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(4), 1–20 (2024)
116. L. Tao, W. Gao, A hardware implementation of entropy encoder for 8k video coding, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
117. Y. Guo, W. Gao, S. Ma, G. Li, Accelerating transform algorithm implementation for efficient intra coding of 8k uhd videos. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* **18**(4), 1–20 (2022)
118. Z. Cai, W. Gao, Efficient fast algorithm and parallel hardware architecture for intra prediction of avs3, in *IEEE International Symposium on Circuits and Systems (ISCAS)* (IEEE, 2021), pp. 1–5
119. W. Gao, H. Yuan, Y. Guo, L. Tao, Z. Cai, G. Li, Openhardwarevc: an open source library for 8k uhd video coding hardware implementation, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 7339–7342
120. W. Gao, H. Yuan, G. Liao, Z. Guo, J. Chen, Pp8k: a new dataset for 8k uhd video compression and processing. *IEEE MultiMedia* (2023)
121. W. Liu, W. Gao, G. Li, S. Ma, T. Zhao, H. Yuan, Enlarged motion-aware and frequency-aware network for compressed video artifact reduction. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
122. Z. Li, G. Li, T. Li, S. Liu, W. Gao, Information-growth attention network for image super-resolution, in *Proceedings of the 29th ACM International Conference on Multimedia* (2021), pp. 544–552
123. L. Zhou, W. Gao, G. Li, H. Yuan, T. Zhao, G. Yue, Disentangled feature distillation for light field super-resolution with degradations, in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2023), pp. 116–121
124. L. Zhou, W. Gao, G. Li, End-to-end spatial-angular light field super-resolution using parallax structure preservation strategy, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 3396–3400
125. W. Gao, L. Zhou, L. Tao, A fast view synthesis implementation method for light field applications. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* **17**(4), 1–20 (2021)
126. X. Zhang, W. Gao, G. Li, Q. Jiang, R. Cong, Image quality assessment-driven reinforcement learning for mixed distorted image restoration. *ACM Trans. Multimedia Comput. Commun. Appl.* **19**(1s), 1–23 (2023)
127. X. Zhang, W. Gao, H. Yuan, G. Li, Je 2 net: joint exploitation and exploration in reinforcement learning based image restoration, in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2090–2094
128. X. Zhang, W. Gao, Hirl: hybrid image restoration based on hierarchical deep reinforcement learning via two-step analysis, in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2445–2449
129. S. Wang, Y. Lu, W. Xia, P. Xia, Z. Wang, W. Gao, Light field angular super-resolution by view-specific queries. *Vis. Comput.* 1–16 (2024)
130. Y. Sun, Z. Li, S. Wang, W. Gao, Depth-assisted calibration on learning-based factorization for a compressive light field display. *Opt. Express* **31**(4), 5399–5413 (2023)
131. Y. Sun, Z. Li, L. Li, S. Wang, W. Gao, Optimization of compressive light field display in dual-guided learning, in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2075–2079
132. W. Gao, S. Fan, G. Li, W. Lin, A thorough benchmark and a new model for light field saliency detection. *IEEE Trans. Pattern Anal. Mach. Intell.* (2023)

133. Z. Guo, W. Gao, H. Wang, J. Wang, S. Fan, No-reference deep quality assessment of compressed light field images. in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6
134. L. Xie, X. Mu, G. Li, W. Gao et al., Pku-dpcc: a new dataset for dynamic point cloud compression. *APSIPA Trans. Signal Inf. Process.* **13**(6) (2024)
135. H. Yuan, W. Gao, W. Gao, Opensep: an open source subjective experiment platform, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 248–11 251
136. S. Fan, W. Gao, Screen-based 3d subjective experiment software, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9672–9675
137. J. Wang, W. Gao, G. Li, Zoom to perceive better: No-reference point cloud quality assessment via exploring effective multiscale feature. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
138. J. Wang, W. Gao, G. Li, Applying collaborative adversarial learning to blind point cloud quality measurement. *IEEE Trans. Instrum. Meas.* (2023)
139. H. Li, W. Gao, Information exploration of projected views for point cloud quality measurement. *IEEE Trans. Instrum. Meas.* (2024)
140. B. Qu, X. Liang, S. Sun, W. Gao, Exploring aigc video quality: a focus on visual harmony, video-text consistency and domain distribution gap, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2024)
141. B. Qu, H. Li, W. Gao, Bringing textual prompt to ai-generated image quality assessment, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024)
142. S. Sun, X. Liang, S. Fan, W. Gao, W. Gao, Ve-bench: subjective-aligned benchmark suite for text-driven video editing quality assessment, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
143. W. Liu, W. Gao, X. Mu, Fast inter-frame motion prediction for compressed dynamic point cloud attribute enhancement. *Proc. AAAI Conf. Artif. Intell.* **38**(4), 3720–3728 (2024)
144. Z. Yang, W. Gao, X. Lu, Danet: density-adaptive network for geometry-based point cloud compression artifacts removal, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
145. X. Fan, G. Li, D. Li, Y. Ren, W. Gao, T.H. Li, Deep geometry post-processing for decompressed point clouds, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
146. X. Zhang, G. Liao, W. Gao, G. Li, Tdrnet: transformer-based dual-branch restoration network for geometry based point cloud compression artifacts, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
147. Z. Li, G. Li, T.H. Li, S. Liu, W. Gao, Semantic point cloud upsampling. *IEEE Trans. Multimedia* (2022)
148. Z. Li, S. Liu, W. Gao, G. Li, G. Li, S4r: rethinking point cloud sampling via guiding upsampling-aware perception. *IEEE Trans. Multimedia* (2025)
149. R. Zhang, W. Gao, G. Li, T.H. Li, Qinet: decision surface learning and adversarial enhancement for quasi-immune completion of diverse corrupted point clouds. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–14 (2022)
150. R. Bao, Y. Ren, G. Li, W. Gao, S. Liu, Flow-based point cloud completion network with adversarial refinement, in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2559–2563
151. J. Chen, G. Li, R. Zhang, T.H. Li, W. Gao, Pointvae: invertible variational autoencoder framework for 3d point cloud generation, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 3216–3220
152. R. Zhang, J. Chen, W. Gao, G. Li, T.H. Li, Pointot: interpretable geometry-inspired point cloud generative model via optimal transport. *IEEE Trans. Circuits Syst. Video Technol.* **32**(10), 6792–6806 (2022)
153. W. Gao, G. Li, H. Yuan, R. Hamzaoui, Z. Li, S. Liu, Apccpa’22: 1st international workshop on advances in point cloud compression, processing and analysis, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 7392–7393

154. W. Gao, G. Li, Point cloud compression, enhancement and applications: from 3d perception to large models, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 292–11 293
155. H. Sun, G. Li, S. Chen, L. Li, W. Gao, Special topic on 3d point cloud processing and applications. *ZTE Commun.* **21**(4), 1 (2023)
156. W. Gao, W. Gao, X. Mu, C. Peng, G. Li, Overview and comparison of avs point cloud compression standard. *APSIPA Trans. Signal Inf. Process.* (2024)
157. W. Gao, S. Wang, X. Zhang, S. Kwong, Editorial for special issue on deep learning-based data compression. *APSIPA Trans. Signal Inf. Process.* **13**(6) (2024)
158. W. Gao, H. Ye, G. Li, H. Zheng, Y. Wu, L. Xie, OpenPointCloud: an open-source algorithm library of deep learning based point cloud compression, in *ACM International Conference on Multimedia* (2022), pp. 7347–7350
159. Y. Zhang, W. Gao, G. Li, Openpointcloud-v2: a deep learning based open-source algorithm library of point cloud processing, in *Proceedings of the 1st International Workshop on Advances in Point Cloud Compression, Processing and Analysis* (2022), pp. 51–55
160. L. Xie, W. Gao, Learningpcc: a pytorch library for learning-based point cloud compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 234–11 238
161. L. Xie, W. Gao, Pchmvision: an open-source library of point cloud compression for human and machine vision, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 239–11 243
162. L. Xie, M. Hu, X. Bai, Towards hardware-friendly and robust facial landmark detection method, in *International Conference on Neural Information Processing* (Springer, 2022), pp. 432–444
163. L. Xie, M. Hu, X. Bai, Online improved vehicle tracking accuracy via unsupervised route generation, in *IEEE 34th International Conference on Tools with Artificial Intelligence* (IEEE, 2022), pp. 788–792
164. X. Deng, W. Yang, R. Yang, M. Xu, E. Liu, Q. Feng, R. Timofte, Deep homography for efficient stereo image compression, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 1492–1501
165. W. Bao, W. Wang, Y. Xu, Y. Guo, S. Hong, X. Zhang, Instereo2k: a large real dataset for stereo matching in indoor scenes. *Sci. China Inf. Sci.* **63**, 1–11 (2020)
166. A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: the kitti dataset. *Int. J. Robot. Res.* **32**(11), 1231–1237 (2013)
167. N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, T. Brox, A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 4040–4048
168. Y. Choi, N. Kim, S. Hwang, K. Park, J.S. Yoon, K. An, I.S. Kweon, Kaist multi-spectral day/night data set for autonomous and assisted driving. *IEEE Trans. Intell. Transp. Syst.* **19**(3), 934–948 (2018)
169. K. Tong, X. Jin, C. Wang, F. Jiang, Sadn: learned light field image compression with spatial-angular decorrelation, in *IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE, 2022), pp. 1870–1874
170. J. Liu, S. Wang, R. Urtasun, Dsic: deep stereo image compression (2019). [arXiv:1908.03631](https://arxiv.org/abs/1908.03631)
171. Z. Huang, Z. Sun, F. Duan, A. Cichocki, P. Ruan, C. Li, L3c-stereo: lossless compression for stereo images (2021). [arXiv:2108.09422](https://arxiv.org/abs/2108.09422)
172. M. Wödlinger, J. Kotera, J. Xu, R. Sablatnig, Sasic: stereo image compression with latent shifts and stereo attention, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 661–670
173. J. Lei, X. Liu, B. Peng, D. Jin, W. Li, J. Gu, Deep stereo image compression via bi-directional coding, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 19 669–19 678

174. X. Jiang, W. Tan, T. Tan, B. Yan, L. Shen, Multi-modality deep network for extreme learned image compression (2023). [arXiv:2304.13583](https://arxiv.org/abs/2304.13583)
175. G. Lu, T. Zhong, J. Geng, Q. Hu, D. Xu, Learning based multi-modality image and video compression, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 6083–6092
176. Y. Lin, T. Xu, Z. Zhu, Y. Li, Z. Wang, Y. Wang, Your camera improves your point cloud compression, in *IEEE International Conference on Acoustics, Speech and Signal Processing* (2023), pp. 1–5
177. Z. Zhao, S. Wang, C. Jia, X. Zhang, S. Ma, J. Yang, Light field image compression based on deep learning, in *IEEE International Conference on Multimedia and Expo* (2018), pp. 1–6
178. D. Liu, X. Huang, W. Zhan, L. Ai, X. Zheng, S. Cheng, View synthesis-based light field image compression using a generative adversarial network. *Inf. Sci.* **545**, 118–131 (2021)
179. K. Tong, X. Jin, C. Wang, F. Jiang, Sadn: learned light field image compression with spatial-angular decorrelation (2022). [arXiv:2202.10837](https://arxiv.org/abs/2202.10837)
180. J. Shi, Y. Xu, C. Guillemot, Learning kernel-modulated neural representation for efficient light field compression, p. 2307 (2023)
181. M. Li, J. Li, S. Gu, F. Wu, D. Zhang, End-to-end optimized 360° image compression. *IEEE Trans. Image Process.* **31**, 6267–6281 (2022)
182. N.M. Bidgoli, R.G.D.A. Azevedo, T. Maugey, A. Roumy, P. Frossard, Oslo: on-the-sphere learning for omnidirectional images and its application to 360-degree image compression (2021). [arXiv:2107.09179](https://arxiv.org/abs/2107.09179)

Chapter 7

Human and Machine Perception Oriented Image and Video Coding



7.1 Introduction

With the increasing demand for efficient storage and transmission of visual data, deep learning-based image and video compression methods [1–9] have gained significant attention in recent years. Traditional compression techniques, such as JPEG, JPEG2000, and BPG, focus primarily on optimizing image quality based on mathematical metrics, often overlooking how humans and machines perceive visual information differently. As both human perception and machine vision [10–21] become key recipients of compressed visual content, modern compression algorithms [22–33] must account for both objectives, balancing visual quality for human users and features preservation for machine-based analysis.

Recent studies [34–97] have demonstrated that human visual perception is highly sensitive to specific spatial and semantic features of an image, leading to the developing of perceptually optimized compression techniques. These approaches often incorporate loss functions, semantic coding strategies that emphasize perceptually significant regions, and Just-Noticeable-Difference (JND) modeling to minimize perceptible distortions. Furthermore, region-of-interest (ROI) coding schemes have been proposed to allocate bits more efficiently by preserving crucial details in important regions while compressing less significant areas aggressively.

On the other hand, machine vision-oriented compression techniques [98–107, 107–128] aim to retain essential features for downstream analysis tasks. Unlike human viewers, machine vision models rely on deep feature representations rather than pixel-wise fidelity. Consequently, new frameworks, such as Latent Space Masking Networks (LSMnet) and preprocessing-enhanced image compression methods, have emerged to optimize compressed representations for specific machine vision applications. These methods integrate deep learning-based preprocessing-enhanced image compression methods, have emerged to optimize compressed representations for specific machine vision applications. These methods [68–97] integrate deep learning-based preprocessing, adaptive quantization, and task-aware latent space

modifications to enhance compression efficiency while maintaining the integrity of features critical for analysis.

The field of VCM has similarly emerged as a prominent area of research, particularly in applications such as industrial automation. Unlike traditional video compression, which prioritizes subjective visual quality, VCM focuses on encoding compact visual representations that facilitate accurate and efficient video compression tailored to machine vision tasks.

7.2 Typical Methods of Human Perception-based Coding

7.2.1 *Loss Function for Human Perception*

A simple approach to leveraging human visual perception in deep learning-based image compression is to create a loss function that closely aligns with how the human eye perceives images. Snell et al. [129] propose a reconstruction metric called Multi-Scale Structural Similarity (MS-SSIM) which is more in tune with human vision. MS-SSIM is an improved version of the Structural Similarity Index (SSIM) used for measuring the perceptual quality of images or videos. SSIM appraises the image similarity by gauging changes in the aspects of luminance, contrast, and structure. However, SSIM operates at a single scale, which limits its ability to capture image details at different resolutions. MS-SSIM improves upon SSIM by computing similarity at multiple scales. It first creates a series of downsampled versions of the input images and then computes SSIM for each scale. This multi-scale approach captures image details at various resolutions, allowing for a more comprehensive assessment of image quality. MS-SSIM is especially useful for comparing images in image compression, restoration, or enhancement tasks, where different levels of detail may be affected. Similarly, Dosovitskiy and Brox [130] proposed deep perceptual similarity metrics (DeePSiM) as a category of loss functions, designed for generating sharp, high-resolution images based on compressed abstract representations. The primary insight is to compute distances between images in the feature space of deep neural networks rather than in the pixel space, which better captures perceptual similarities between images and leads to more visually appealing results. Their approach involves computing distances between high-level CNN-extracted image features. This method demonstrates greater sensitivity to perceptually relevant attributes (e.g., edges and textures) while maintaining robustness against minor deformations that preserve visual similarity. Using only a feature-space distance metric often leads to suboptimal results, as this approach tends to generate undesirable high-frequency artifacts in the output. Therefore, the authors introduce an adversarial loss to encourage the generation of realistic images. The composite loss function comprises three elements, including a feature loss component that assesses generated-target image similarity through adversarial loss in high-level feature space. It attempts to guarantee that the

generated images look realistic to a discriminator network, as in Generative Adversarial Network (GAN), and image-space loss, which uses squared error in pixel space to stabilize training and reduce noise in the outputs.

7.2.2 Semantic Coding for Human Perception

CNN-based image compression methods usually focus on the reconstruction of low level information, while the reconstruction of perceptual information is generally ineffective. Generative models such as GAN are more effective in reconstructing perceptual information. Wang et al. [131] proposed a method that breaks down an image into multiple scales using a Laplacian pyramid. The image is then progressively reconstructed from lower to higher resolution. For each scale, a dedicated neural network is trained to capture the compressed representation, with the optimization focusing on different aspects such as pixel fidelity, semantic meaning, distribution, and entropy. This approach ensures a well-balanced trade-off between rate, distortion, and perception, allowing for the retention of high-quality visual details even at lower bit rates. The proposed framework leverages a multi-scale architecture based on a Laplacian pyramid, progressively transmitting and recovering image details. A deep neural network is trained for each scale, containing components such as an encoder, quantizer, decoder, and discriminator. Each scale optimizes for specific goals, which helps to progressively build up the image from coarse to fine details.

The process starts with the image being decomposed into sub-images of varying resolutions. At each level of the pyramid, the model refines the image, adding more detailed layers until the final image is reconstructed at full resolution. The mathematical formulation for building the Laplacian pyramid is as follows:

$$y_m = x_m - U(x_{m+1}), \quad (7.1)$$

where x_m and y_m represent images at different scales, and $U(\cdot)$ is the upsampling function. Quantization, which is key to compressing image features, is defined by:

$$\hat{z}_i = Q(z_i) = \operatorname{argmin}_{c_k \in C} \|z_i - c_k\|, \quad (7.2)$$

where z_i represents the i th continuous value obtained from the encoder. The encoder processes the original image and encodes it into a high-dimensional continuous space. c_k is a quantization center from the set C , which is a collection of predefined or learned discrete values. The quantizer Q maps the continuous value z_i to one of the discrete quantization centers c_k .

The optimization in this framework balances rate, distortion, and perception. The total loss function combines multiple objectives, such as Mean Squared Error (MSE) for pixel-level accuracy, feature matching loss for perceptual quality, and adversarial loss for texture fidelity:

$$J = \frac{1}{N} \sum_{n=1}^N (\lambda_{MSE} L_{MSE} + \lambda_{FM} L_{FM} + \lambda_G L_G + \gamma H), \quad (7.3)$$

where the hyper-parameters λ_{MSE} , λ_{FM} , λ_G and γ balance the trade-off among four terms. In this equation, L_{MSE} denotes the pixel-level constraint to keep the consistency between two images. The MSE loss is calculated as $L_{MSE}(\theta) = \mathbb{E}_{(\mathbf{x}, \hat{\mathbf{x}})} \frac{1}{N_x} \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2$ where N_x denotes the size of x . Besides L_{FM} denotes the future matching loss, which relates to the perceptual loss derived from a pre-trained deep neural network. The L_{FM} is calculated as $L_{FM}(\theta) = \mathbb{E}_{(\mathbf{x}, \hat{\mathbf{x}})} \sum_{i=1}^T \frac{1}{N_{D(i)}} \|D^{(i)}(\mathbf{x}) - D^{(i)}(\hat{\mathbf{x}})\|_1$ where $D^{(i)}(\cdot)$ denotes the output of the discriminator's i th layer, and $N_{D(i)}$ denotes the size of $D^{(i)}(\cdot)$. The L_G loss is the discriminator loss of the GAN network, which is described as $L_G(\theta) = \mathbb{E}_{\hat{\mathbf{x}}} \log(1 - D(\hat{\mathbf{x}}))$.

7.2.3 Regions-of-Interest-based Coding for Human Perception

The HVS has the remarkable ability to focus on specific areas of a visual scene, known as Regions-of-Interest (ROI), while giving less attention to other regions. This selective attention plays a critical role in how human process and perceive visual information. By focusing on the ROI regions of an image, i.e., the parts of an image that are more interesting to the human eye, and allocating more bit rate to these parts. This approach enables simultaneous optimization of human visual quality and bitrate efficiency [132, 133].

Cai et al. [134] introduced an innovative deep learning-based framework for ROI image compression, designed to optimize both bit allocation and rate-distortion performance in an end-to-end fashion. Traditional ROI compression methods typically split the process into two steps—ROI prediction and image compression—which often prevents them from reaching the optimal rate-distortion trade-off. The proposed architecture addresses these limitations by integrating both tasks within a unified framework that is fully trainable. The framework is shown in Fig. 7.1.

7.2.4 Just-Noticeable-Distortion Modeling

Just-Noticeable-Distortion (JND) refers to the smallest detectable difference between two stimuli that a person can perceive. In the context of image or video processing, JND typically represents the minimal variation in image or video quality that is noticeable to the human eye. This approach holds significant importance in applications including image compression, video coding, and perceptual quality assessment,

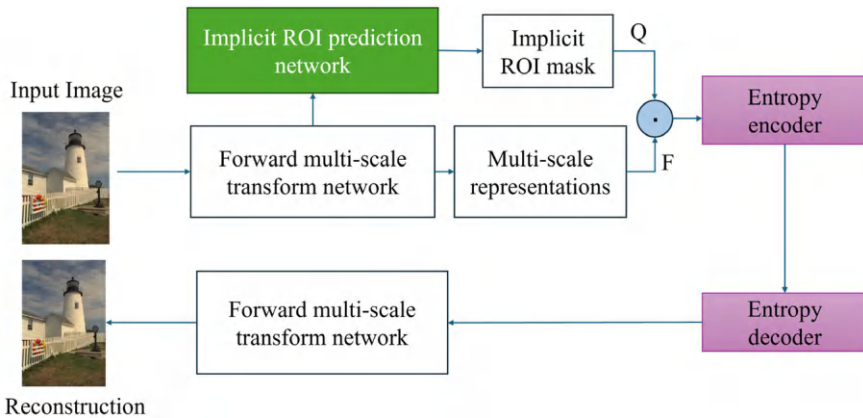


Fig. 7.1 The framework of the end-to-end ROI image compression. ©2019 IEEE Reprinted, with permission, from [134]

as it allows systems to optimize resource usage by reducing or altering information in a way that remains imperceptible to the viewer.

Liu et al. [135] developed a deep learning framework for predicting Picture-Wise Just-Noticeable-Distortion (PW-JND) to improve image compression performance. PW-JND represents the minimal detectable image distortion perceptible to human vision. Their approach preserves high visual quality while reducing data size by casting PW-JND prediction as a multi-class task, then reformulating it as a binary classification problem. A deep CNN-based binary classification network, referred to as the perceptually lossy/lossless predictor, is integrated into the framework to evaluate whether the perceptual quality of a distorted image is preserved or compromised compared to its original reference. After training the model to detect noticeable distortions, different compressed versions of each image are generated by adjusting the compression quality factor (QF). Afterward, the predictor discerns which variants show perceptual distortion, and the PW-JND is detected through a sliding window process.

The first diagram represents a deep learning-based method for determining whether an image can be compressed losslessly or with some loss in quality, depending on its perceptual quality. The process starts with a set of training samples, which consist of both high-quality reference images and distorted versions of these images. These images are used to train the system to recognize distortions and quality differences. The system takes two inputs: a reference image (the original high-quality image) and a distorted image (a lower-quality version of the reference). The next step is patch selection, where the image is divided into smaller patches or blocks. This step focuses on selecting key patches from the images for further analysis. After patch selection, the patches are passed through a CNN-based feature extractor. The convolutional neural network (CNN) extracts important features from each patch, capturing details like textures, edges, and patterns that are essential for assessing

image quality. These extracted features from both the reference and distorted images are then combined in the patch feature fusion step. This fusion process merges the features to create a comprehensive representation of each image patch. Next, the system calculates a patch-wise quality measure, which evaluates the perceptual quality of each individual patch by comparing it to the reference image patch. Finally, a decision is made in the picture-wise lossy/lossless prediction stage, where the system predicts whether the entire image can be compressed in a lossless manner (preserving all details) or if lossy compression is acceptable (allowing quality degradations for reduced file size).

As shown in the second diagram, a tripartite process identifies the PW-JND, representing the threshold of image distortion that escapes visual detection. The process starts in Step 1 with a test image being subjected to different levels of compression, leading to in a series of a distorted images. These images vary in quality due to different levels of lossy compression being applied, creating a range of images from slightly distorted to highly compressed versions of the test image. In Step 2, the system uses a lossy/lossless predictor to analyze the distorted images. This predictor helps to identify which versions of the image can be compressed without noticeable visual degradation. The final step, Step 3, involves a search strategy that fine-tunes the selection process to find the exact level of distortion that reaches the PW-JND threshold. This threshold is the point at which any further compression would result in noticeable quality loss to human viewers. The process concludes with the identification of the PW-JND image, which is the most highly compressed image that maintains perceptual quality, meaning the viewer would not notice the difference between the original and the compressed version under normal viewing conditions.

7.2.5 *Latent Space Masking Network*

With the rapid development of machine-to-machine data communication, the increased image and video data is no longer consumed directly only by humans but is instead used for data analysis and processing between machines. This trend is especially prominent in fields such as transportation, industrial automation, and smart city surveillance, where data needs not only to be efficiently transmitted but also to be optimized for machine vision tasks. As a result, while compressed images may appear visually good, they do not necessarily perform well when used by machines for analysis tasks. To address this challenge, there is an urgent need for a new image compression method that is specifically optimized for machine vision tasks. In this context, Fischer et al. [136] introduced Latent Space Masking Network (LSMNet) to offer a more efficient compression network tailored for machine vision tasks, as depicted in Fig. 7.2.

The core concept behind LSMNet is to eliminate irrelevant latent information during compression, reducing the data amount, while preserving the accuracy of machine tasks. The LSMNet architecture runs parallel to the encoder, generating masks to filter out parts of the latent space that are unimportant for visual tasks.

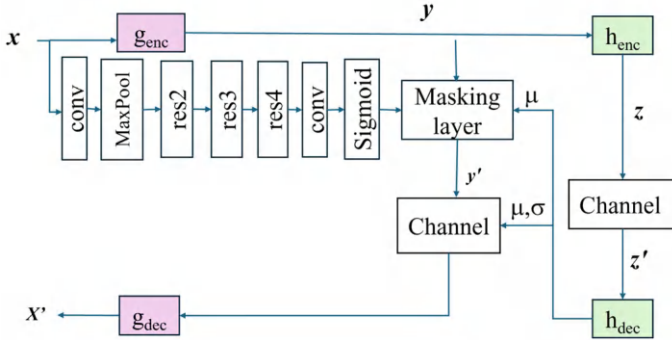


Fig. 7.2 The overall structure of LSMNet. Public domain open access image [136]

Specifically, LSMNet leverages early features extracted by the analysis network to determine which information is critical and which can be reduced or masked. In implementation, LSMNet employs a “soft masking” mechanism, which adjusts latent representations by shifting their values toward estimated means, thereby reducing the number of bits required for transmission. Unlike traditional hard masking, which completely removes data, this approach makes slight adjustments to reduce redundant information without negatively impacting machine vision performance. Moreover, LSMNet is highly modular, meaning it can be easily integrated into existing neural network compression architectures without requiring a full retraining of the model, significantly lowering the complexity and the cost of implementation.

In conclusion, LSMNet provides an innovative solution for machine-oriented image compression by intelligently filtering out non-essential information, achieving significant bitrate savings. This technological breakthrough not only excels in machine perception tasks, but also lays the foundation for more efficient image and video compression in future data-intensive scenarios. This flexibility and efficiency of LSMNet suggest its potential for widespread application in other fields, making it a promising technology for a variety of industrial and commercial contexts in the future.

7.3 Machine Vision-Oriented Codec Design and Optimization

7.3.1 Encoder-Decoder for Machine Vision Tasks

This category can be split into two directions based on different strategies for reconstructing images or videos to incorporate information beneficial for machine vision tasks, i.e., machine vision constraints and semantic information preservation. Under the machine vision constraints approach, end-to-end learned codecs are trained with



Fig. 7.3 The overall structure of encoder-decoder for machine vision tasks. *Source* Author

loss functions dedicated to machine vision optimization. In [137–140], a task-driven loss is employed, meaning the codec is directly connected to a downstream machine vision network, and both components are jointly trained. Furthermore, in [138], a perceptual loss is introduced as part of the content-adaptive fine-tuning strategy to boost performance, thereby enhancing compression efficiency for machine processing. Yang et al. [141] employ a maximum mean discrepancy metric to align feature distributions, leading to more consistent feature-domain perceptions and enhanced recognition accuracy by pre-trained machine vision models (Fig. 7.3).

To meet the demands of machine vision with lower computational cost, Kim et al. [142] developed a multi-scale feature compression technique that effectively removes redundancy and improves overall compression effectiveness. Before [142], Versatile Video Coding (VVC) showed superior performance in feature compression, although it was designed for natural images, indicating potential for further improvement. Kim et al. [142] aimed to explore feature redundancy and proposed a multi-scale feature compression algorithm with reduced computational complexity. It is important to note that [142] is exclusively designed for machine vision and is not suitable for human viewers. Kim et al. delved into feature fusion and learned compression, topics not previously explored in combination in prior research. This method achieves efficient feature fusion and encoding by first converting large-scale features into latent representations, which are then progressively integrated with features from smaller scales. The core modules proposed in [142] are the Feature Fusion and Encoding Network (FENet) and the Feature Decoding and Reconstruction Network (DRNet). The FENet capitalizes on the redundancy between multi-scale features, while the DRNet progressively decodes these features. Additionally, the approach in [142] demonstrates applicability across multiple tasks, encompassing both object detection and instance segmentation. Experiments show that their approach can obtain excellent compression performance for machine vision tasks, along with a lower complexity [6, 58].

Zhang et al. [143] employed competitive learning to optimize post-processing filters, addressing machine vision requirements in video encoding to improve codec performance for machine-centric visual applications. Traditional video and image compression technologies, such as Versatile Video Coding (VVC), are primarily designed for human viewing and experience, with optimization objectives typically focused on visual quality and user experience. However, with the rapid development of machine perception tasks, the existing codecs exhibit limitations when handling machine perception tasks because machines have low tolerance for compression artifacts and have different processing requirements for specific content. Therefore, there is a need for video encoding methods specifically optimized for machine vision,

a trend dubbed as “Video Coding for Machines (VCM)”. Traditional approaches typically use a single filter trained based on fixed artifact features (such as QP values), but this method cannot effectively optimize for different content, resulting in low efficiency.

To this end, a competitive learning-based training strategy is proposed for implementing content-specific filters in machine-oriented video encoding. The goal is to dynamically assign training samples to each filter, further optimizing the filters to handle specific types of content. Specifically, multiple filters are trained simultaneously, competing when processing particular training samples to determine which filter performs best. The better-performing filters are assigned higher weights and are further optimized in the next round of training. To avoid the negative effects of randomness during initialization, the Softmax function is adopted as the weight allocation function, and the temperature parameter T is gradually reduced to ensure the convergence of sample assignments during training, allowing the optimal filters to receive more optimization opportunities. The post-processing filters designed in the paper use an autoencoder structure, characterized by: (1) downsampling and feature extraction through convolution operations and upsampling and reconstruction through deconvolution; (2) using residual blocks to preserve image details and enhance the model’s representation capability; (3) inputting a QP indicator to indicate the quality of the reconstructed data, helping the filters adjust their processing strategies under different compression qualities [8, 9, 68].

Semantic information preservation techniques [144] enforce that the encoded/decoded bitstream retain meaningful information by incorporating a classifier that generates semantic label predictions for machine vision applications from the decoded bitstream. In conclusion, although machine vision-oriented codecs still produce outputs that are visually natural to humans and thus remaining amenable to machine vision pipelines, they demand a fully featured analysis model. This configuration actually heightens the computational load on both the front-end and the cloud.

7.3.2 *Signal-Semantic Joint Codec*

Rather than focusing on preserving feature fidelity, this entire compression framework is jointly optimized for machine vision tasks. Alvar et al. [145] introduced a novel loss function that explicitly incorporates feature compressibility considerations, thereby directing the training process of multi-task learning models. Likewise, Singh et al. [146] introduced the penalty mechanism to train networks end to end, balancing both the representational capacity and compressibility of deep features. Shah et al. [147] proposed an annealing representation shrinkage approach, which produces a compact network (and corresponding features) through iterative adjustments of the shrinkage model incorporating network layer reduction and annealed label techniques. Zhang et al. [148] further proposed a multi-scale feature compression

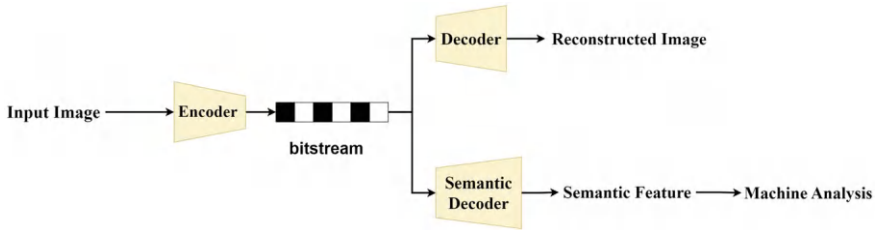


Fig. 7.4 The overall structure of joint signal-semantic Codec with the same bitstream. *Source* Author

framework, which is collaboratively trained alongside Mask R-CNN’s objectives—covering mask generation, box localization, and category classification (Fig. 7.4).

In their comprehensive review of VCM, Yang et al. [149] grouped the approaches into several categories, including feature-assisted schemes, scalable coding strategies, intermediate feature compression/optimization, and vision-oriented codec designs. They propose a novel approach to unify intermediate features from diverse vision tasks into a generalized form, which is effective in removing redundancies among various types of features. The algorithm they developed demonstrates the capability to efficiently support a range of vision tasks, even at low bit rates. Furthermore, Yang et al. proposed an innovative codebook hyperprior model, aiming to significantly improve the compression efficiency of network-generated features. This innovative design could potentially uncover underlying patterns within network-generated features. Experimental evaluations demonstrate that the proposed algorithm significantly enhances machine vision analysis at low bit rates, outperforming previous work. This work suggests the potential to find commonalities among features for different tasks, inspiring future research.

In summary, leveraging intermediate feature compression and optimization facilitates collaborative intelligence. However, such methods cannot provide the complete image and video reconstruction.

7.3.3 Scalable Codec for Collaborative Perception

This line of research addresses the joint compression and transmission of both feature streams and image/video signal streams, typically following one of two approaches: dual-stream scalable encoding or multi-stream scalable encoding. Dual-stream scalable encoding often adopts a hierarchical strategy, where the feature stream not only facilitates image/video reconstruction but also meets the demands of machine vision applications. In [150–152], for instance, a dual-stream configuration is employed to balance the accuracy of facial image analysis with high-quality reconstruction. The work in [153] models and conveys semantic priors under a conceptual coding framework, enabling low-bitrate image compression while employing the reconstructed

images for facial landmark detection. In [154], there exists a mutual enhancement between semantic segmentation maps and reconstructed images, resulting in better semantic and visual performance. Additionally, [155] learns motion patterns and transmits key points along with their trajectories to support both action recognition and the reconstruction of video frames (Fig. 7.5).

Multi-stream scalable encoding transmits features through multiple streams. In [156], an auxiliary stream, a base stream, and an enhancement stream are sent simultaneously. Reference [157] proposed a feature compression method for intermediate representations, where distinct layers cater to tasks requiring different levels of semantic detail. Additionally, [158] introduced an enhancement structure based on trainable and reversible transformations, partitioning images into multiple bands to support a variety of machine vision applications.

Duan et al. [159] focused on the concept of VCM, building on traditional video coding and feature compression works. In their definition, they outlined three critical aspects of VCM: tasks, features, and resources, and emphasized the importance of collaborative and scalable modes catering to both human and machine vision. Furthermore, they introduced exemplary paradigms and solutions within VCM to illustrate its superior performance over conventional video and feature compression methods. Their proposed dual-stream architecture features two separate streams, specifically designed to cater to human viewers and machine processing, respectively. These streams are collaboratively managed through a feedback mechanism, adapting to the real-time needs of both human visual experience and machine vision tasks. The complete architecture, tailored for surveillance purposes, effectively embodies the innovative concept of a digital retina.

In summary, by leveraging scalable mechanisms and feature streams, these methods effectively and flexibly accommodate both human and machine vision. Nonetheless, when conducting intelligent collaborative analysis of large-scale data across diverse tasks/applications, vital challenges persist—particularly regarding feature transferability between different tasks and the generalizability of the extracted features (Fig. 7.6).

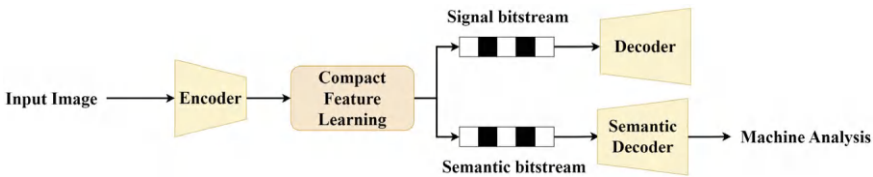


Fig. 7.5 The overall structure of joint signal-semantic codec with the separate bitstream. *Source* Author

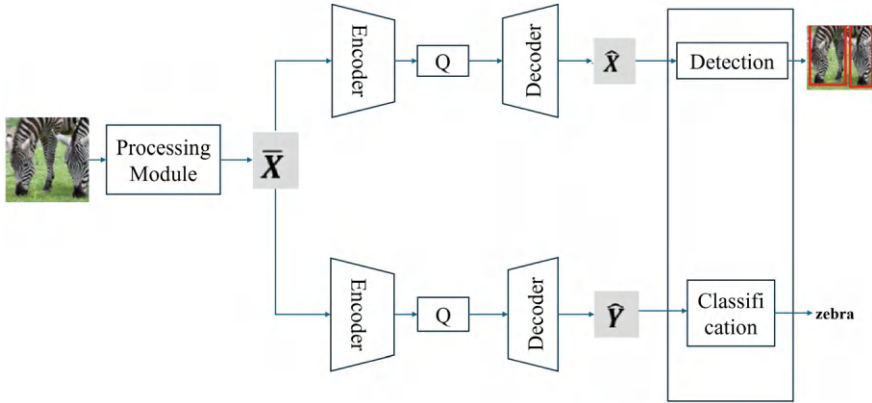


Fig. 7.6 An overview of image compression methods improved by preprocessing for machine vision. Public domain open access image, from [160]

7.4 Summary

The domain of image and video compression has traditionally been dominated by techniques optimized for the HVS, aiming to maximize visual fidelity and minimize perceptible distortions. Standards such as JPEG, HEVC, and VVC employ methods like discrete cosine transforms, motion compensation, and entropy coding to efficiently reduce data size while preserving image quality as perceived by humans. These approaches prioritize pixel-level accuracy and visual aesthetics, ensuring compressed content remains visually indistinguishable from the original to the human eye.

Conversely, the rise of machine vision applications necessitates a paradigm shift in compression strategies. Machine perception-based compression methods, including VCM, LSMNet, and feature-based compression algorithms, are designed to preserve semantic and structural information critical for machine analysis tasks. These techniques often leverage deep learning models to identify and retain task-relevant features, employing specialized loss functions that prioritize feature fidelity over pixel-perfect reconstruction. For instance, LSMNet enhances compression efficiency by selectively masking irrelevant latent information, thereby reducing data redundancy without compromising the integrity of features essential for machine tasks.

Human and machine perception-friendly compression emerges as an integrative approach that seeks to balance the distinct requirements of both human and machine vision systems. This hybrid methodology involves designing multi-objective loss functions that simultaneously optimize for visual quality and feature preservation. Techniques such as semantic coding and ROI are adapted to ensure that salient regions important for both human perception and machine analysis receive higher bit allocation. Additionally, JND modeling is employed to determine acceptable

levels of distortion that remain imperceptible to humans while keeping the efficacy of machine perception tasks.

A key distinction between human-centric and machine-centric compression lies in their optimization goals. Human-oriented methods excel in maintaining visual aesthetics by minimizing artifacts and preserving color fidelity, whereas machine-oriented methods focus on retaining high-level semantic information and feature robustness necessary for accurate machine analysis. Human and machine perception-friendly compression aims to harness the strengths of both paradigms, guaranteeing the reconstructed content visually appealing to humans and being analytically effective for machines. However, achieving this balance introduces complexities in codec design, as it requires sophisticated algorithms capable of dynamically adjusting compression parameters based on the dual objectives.

Furthermore, the comparative analysis highlights inherent trade-offs between the two approaches. While human-centric compression may inadvertently discard information vital for machine tasks, machine-centric compression might degrade visual quality below acceptable thresholds for human perception. Joint optimization methods seek to alleviate these trade-offs by implementing advanced techniques such as dynamic bitrate allocation and content-aware encoding, enabling context-dependent balancing of human and machine vision needs.

In conclusion, by integrating the principles of both human and machine perception, collaborative compression methods offer a versatile solution that enhances the utility of compressed content across diverse use cases. Future research in this area is poised to explore more sophisticated optimization frameworks, leveraging large-scale models and cross-domain feature integration to further refine the balance between visual quality and machine analysis performance, thereby unlocking new potentials in the realm of multimedia compression (Table 7.1).

Table 7.1 Summary of learning-based multimodal compression frameworks

Method	Author	Publication	Category
ColorCNN	Hou et al.	CVPR 2020	Encoder-decoder for machine vision tasks
ICM	Le et al.	ICASSP 2021	
LCAIC	Le et al.	ICME 2021	
IPRIC	Su et al.	ICIP 2019	
DIC	Yang et al.	ACM MM 2020	
BAMCI	Alvar et al.	ICASSP 2020	Signal-semantic joint codec
LCF	Singh et al.	ICIP 2020	
DCFAC	Shah et al.	ICASSP 2020	
MSFC	Zhang et al.	ICME 2021	
VCM-SCA	Xia et al.	ICME 2020	Scalable codec for collaborative perception
SSIC-CFM	Yan et al.	ICIP 2020	
S2S-SIC	Liu et al.	IJCV 2021	

Source Author

Conventional compression methods are specifically designed for HVS to maintain high visual quality and minimize perceptual distortions. However, as machine vision applications like object detection, semantic segmentation, and automated surveillance become increasingly prevalent, the inability of traditional compression methods to retain essential features affects the accuracy of machine analysis. To address this, the human and machine perception-friendly compression has emerged, aiming to balance human perceptual quality and machine analysis efficiency. Optimization strategies for human perception include perceptual loss functions (e.g., MS-SSIM, DeePSiM), semantic coding (leveraging GANs to reconstruct perceptual information), ROI compression (allocating more bitrate to important regions), and JND modeling (optimizing compression based on human visual thresholds). Meanwhile, VCM focuses on enhancing feature extraction for deep learning models, with approaches such as LSMNet to remove irrelevant features and improve compression efficiency, Preprocessing-Enhanced Compression to transform images in a task-aware manner before encoding, Multi-Scale Feature Compression to reduce redundancy in deep feature representations, and Competitive Learning to adaptively optimize post-processing filters for machine analysis. Unlike traditional methods that solely prioritize human visual experience, the human and machine perception oriented coding achieves a balance between visual quality and feature preservation through multi-objective optimization, content-aware encoding, and dynamic bitrate allocation. In the future, advancements in large-scale models, cross-model compression, and adaptive encoding strategies will further drive the evolution of this field, making compression technologies more effective for video streaming, AI-driven analysis, and automation, positioning them as a crucial direction in multimedia technology development.

Exercises

1. Please explain why the volume of image and video data has exploded in recent years.
2. Please state the primary goal of traditional image and video compression.
3. Please describe how the development of deep learning has changed image and video compression methods.
4. Please outline the requirements for image and video compression in machine vision systems.
5. Please explain what Video Coding for Machines (VCM) is and state its purpose.
6. Please outline the main advantages of the multi-scale feature compression algorithm proposed by End-to-end learnable multi-scale feature compression for VCM.
7. Please detail how Video-based detection of generalized tonic-clonic seizures using deep learning improved the compression efficiency of intermediate features.
8. Please describe the “rate-distortion-perception” trade-off in image compression.

9. Please explain how human visual perception is utilized in designing the loss function in deep learning-based image compression methods.
10. Please describe what Just-Noticeable-Distortion (JND) modeling is and explain how it is applied in image and video compression.

References

1. W. Gao, G. Li, H. Yuan, R. Hamzaoui, Z. Li, S. Liu, Apccpa'22: 1st international workshop on advances in point cloud compression, processing and analysis, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 7392–7393
2. W. Gao, G. Li, Point cloud compression, enhancement and applications: from 3d perception to large models, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 292–11 293
3. H. Sun, G. Li, S. Chen, L. Li, W. Gao, Special topic on 3d point cloud processing and applications. *ZTE Commun.* **21**(4), 1 (2023)
4. W. Gao, W. Gao, X. Mu, C. Peng, G. Li, Overview and comparison of avs point cloud compression standard. *APSIPA Trans. Signal Inf. Process.* (2024)
5. W. Gao, S. Wang, X. Zhang, S. Kwong, Editorial for special issue on deep learning-based data compression. *APSIPA Trans. Signal Inf. Process.* **13**(6) (2024)
6. W. Gao, H. Ye, G. Li, H. Zheng, Y. Wu, L. Xie, OpenPointCloud: an open-source algorithm library of deep learning based point cloud compression, in *ACM International Conference on Multimedia* (2022), pp. 7347–7350
7. Y. Zhang, W. Gao, G. Li, Openpointcloud-v2: a deep learning based open-source algorithm library of point cloud processing, in *Proceedings of the 1st International Workshop on Advances in Point Cloud Compression, Processing and Analysis* (2022), pp. 51–55
8. L. Xie, W. Gao, Learningpcc: a pytorch library for learning-based point cloud compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 234–11 238
9. L. Xie, W. Gao, Pchmvision: an open-source library of point cloud compression for human and machine vision, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 239–11 243
10. W. Gao, G. Li, *Deep Learning for 3D Point Clouds* (Springer Nature, 2024)
11. W. Gao, G. Li, Introduction to 3d point clouds: Datasets and perception, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 1–27
12. W. Gao, G. Li, Learning basics for 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 29–70
13. W. Gao, G. Li, Deep-learning-based point cloud enhancement i, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 71–97
14. W. Gao, G. Li, Deep-learning-based point cloud enhancement ii, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 99–130
15. W. Gao, G. Li, Deep-learning-based point cloud analysis i, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 131–162
16. W. Gao, G. Li, Deep-learning-based point cloud analysis ii, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 163–193
17. W. Gao, G. Li, Point cloud pre-trained models and large models, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 195–225
18. W. Gao, G. Li, Point cloud-language multi-modal learning, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 227–254
19. W. Gao, G. Li, Open-source projects for 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 255–272

20. W. Gao, G. Li, Typical engineering applications of 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2025), pp. 273–299
21. W. Gao, G. Li, Future work on deep learning-based point cloud technologies, in *Deep Learning for 3D Point Clouds* (Springer, 2025), pp. 301–315
22. G. Li, G. Wei, W. Gao, *Point Cloud Compression: Technologies and Standardization* (Springer Nature, 2024)
23. G. Li, W. Gao, W. Gao, Introduction, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 1–28
24. G. Li, W. Gao, W. Gao, Background knowledge, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 29–51
25. G. Li, W. Gao, W. Gao, Predictive coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 53–70
26. G. Li, W. Gao, W. Gao, Transform coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 71–96
27. G. Li, W. Gao, W. Gao, Quantization techniques, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 97–112
28. G. Li, W. Gao, W. Gao, Entropy coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 113–133
29. G. Li, W. Gao, W. Gao, Mpeg geometry-based point cloud compression (g-pcc) standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 135–165
30. G. Li, W. Gao, W. Gao, Avs point cloud compression standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 167–197
31. G. Li, W. Gao, W. Gao, Mpeg video-based point cloud compression (v-pcc) standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 199–218
32. G. Li, W. Gao, W. Gao, Mpeg ai-based 3d graphics coding standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 219–241
33. G. Li, W. Gao, W. Gao, Future work, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 243–250
34. T. Qin, G. Li, W. Gao, S. Liu, Multi-grained point cloud geometry compression via dual-model prediction with extended octree. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(9), 1–30 (2024)
35. Y. Shao, W. Gao, S. Liu, G. Li, Advanced patch-based affine motion estimation for dynamic point cloud geometry compression. *Sensors* **24**(10), 3142 (2024)
36. Y. Shao, F. Song, W. Gao, S. Liu, G. Li, Texture-guided graph transform optimization for point cloud attribute compression. *Appl. Sci.* **14**(10), 4094 (2024)
37. Y. Shao, X. Yang, W. Gao, S. Liu, G. Li, 3d point cloud attribute compression using diffusion-based texture-aware intra prediction. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
38. J. Zhang, Y. Chen, G. Liu, W. Gao, G. Li, Efficient point cloud attribute compression framework using attribute-guided graph fourier transform, in *ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2024), pp. 8426–8430
39. W. Gao, H. Yuan, G. Li, Z. Li, H. Yuan, Low complexity coding unit decision for video-based point cloud compression. *IEEE Trans. Image Process.* **33**, 149–162 (2023)
40. Y. Shao, G. Li, Q. Zhang, W. Gao, S. Liu, Nonrigid registration-based progressive motion compensation for point cloud geometry compression. *IEEE Trans. Geosci. Remote Sens.* (2023)
41. F. Song, G. Li, X. Yang, W. Gao, S. Liu, Block-adaptive point cloud attribute coding with region-aware optimized transform. *IEEE Trans. Circuits Syst. Video Technol.* (2023)
42. Y. An, Y. Shao, G. Li, W. Gao, S. Liu, A fast motion estimation method with hamming distance for lidar point cloud compression, in *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2022), pp. 1–5
43. H. Yuan, W. Gao, G. Li, Z. Li, Rate-distortion-guided learning approach with cross-projection information for v-pcc fast cu decision, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 3085–3093

44. F. Song, G. Li, W. Gao, T.H. Li, Rate-distortion optimized graph for point cloud attribute coding. *IEEE Signal Process. Lett.* **29**, 922–926 (2022)
45. F. Song, G. Li, X. Yang, W. Gao, T.H. Li, Fine-grained correlation representation for graph-based point cloud attribute compression, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
46. F. Shen, W. Gao, A rate control algorithm for video-based point cloud compression, in *2021 International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2021), pp. 1–5
47. F. Song, Y. Shao, W. Gao, H. Wang, T. Li, Layer-wise geometry aggregation framework for lossless lidar point cloud compression. *IEEE Trans. Circuits Syst. Video Technol.* **31**(12), 4603–4616 (2021)
48. Z. Cai, W. Gao, G. Li, W. Gao, Distortion propagation model-based v-pcc rate control for 3d point cloud broadcasting. *IEEE Trans. Broadcast.* 1–13 (2024)
49. X. Mu, W. Gao, H. Yuan, S. Wang, G. Li, Adaptive lpu decision for dynamic point cloud compression. *IEEE Signal Process. Lett.* (2024)
50. C. Peng, W. Gao, Laplacian matrix learning for point cloud attribute compression with ternary search-based adaptive block partition, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 10 412–10 420
51. Y. Wang, W. Gao, X. Mu, H. Yuan, Rate control optimization for joint geometry and attribute coding of lidar point clouds, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
52. T. Guo, H. Yuan, T. Wang, W. Gao, Graph filter-based fast motion matching for inter frame coding of mpeg g-pcc, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 1151–1155
53. L. Xie, W. Gao, H. Zheng, G. Li, Spcgc: scalable point cloud geometry compression for machine vision, in *2024 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2024), pp. 17 272–17 278
54. L. Xie, W. Gao, H. Zheng, H. Ye, Semantic-aware visual decomposition for point cloud geometry compression, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 595–595
55. Z. Qi, W. Gao, Variable-rate point cloud geometry compression based on feature adjustment and interpolation, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 63–72
56. Z. Yu, W. Gao, When dynamic neural network meets point cloud compression: computation-aware variable rate and checkerboard context, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 600–600
57. L. Xie, W. Gao, S. Fan, Z. Yao, Pdnet: parallel dual-branch network for point cloud geometry compression and analysis, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 596–596
58. L. Xie, W. Gao, H. Zheng, End-to-end point cloud geometry compression and analysis with sparse tensor, in *International Workshop on Advances in Point Cloud Compression, Processing and Analysis* (2022), pp. 27–32
59. C. Fu, G. Li, R. Song, W. Gao, S. Liu, Octattention: octree-based large-scale contexts model for point cloud compression. *Proc. AAAI Conf. Artif. Intell.* **36**(1), 625–633 (2022)
60. C. Zhang, W. Gao, Adadpcc: adaptive rate control and rate-distortion-complexity optimization for dynamic point cloud compression, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
61. K. Wang, W. Gao, Unipcgc: towards practical point cloud geometry compression via an efficient unified approach, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
62. K. Zheng, W. Gao, H. Zheng, Octree-based learned point cloud geometry compression: a lossy perspective, in *2025 Data Compression Conference (DCC)* (IEEE, 2025)
63. Y. Li, Z. Yu, X. Chen, S. Wang, H. Yuan, W. Gao, Zero-shot quantization for large-kernels via shape-based distribution and diversity self-distillation, in *ICASSP 2025–2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2025)

64. X. Chen, Y. Wang, Y. Li, X. Ling, M. Li, R. Liu, M. Ouyang, K. Zhao, T. Guan, Y. He, Low bit-width zero-shot quantization with soft feature-infused hints for iot systems. *IEEE Internet Things J.* (2024)
65. Z. Qi, H. Liu, W. Gao, Variable-rate point cloud compression with channel-wise entropy modeling, in *2025 Data Compression Conference (DCC)* (IEEE, 2025)
66. H. Zheng, W. Gao, Z. Yu, T. Zhao, G. Li, Viewpcgc: view-guided learned point cloud geometry compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 7152–7161
67. L. Xie, W. Gao, H. Zheng, G. Li, Roi-guided point cloud geometry compression towards human and machine vision, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 3741–3750
68. Y. Wu, L. Xie, S. Sun, W. Gao, Y. Yan, Adaptive intra period size for deep learning-based screen content video coding, in *IEEE International Conference on Multimedia and Expo Workshops* (2024), pp. 1–6
69. H. Zheng, W. Gao, End-to-end rgb-d image compression via exploiting channel-modality redundancy. *Proc. AAAI Conf. Artif. Intell.* **38**(7), 7562–7570 (2024)
70. L. Tao, W. Gao, G. Li, C. Zhang, Adanic: towards practical neural image compression via dynamic transform routing, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 16 879–16 888
71. Y. Wu, W. Gao, End-to-end lossless compression of high precision depth maps guided by pseudo-residual (2022). [arXiv:2201.03195](https://arxiv.org/abs/2201.03195)
72. Y. Wu, Z. Qi, H. Zheng, L. Tao, W. Gao, Deep image compression with latent optimization and piece-wise quantization approximation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 1926–1930
73. W. Gao, L. Tao, L. Zhou, D. Yang, X. Zhang, Z. Guo, Low-rate image compression with super-resolution learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2020), pp. 154–155
74. W. Gao, S. Sun, H. Zheng, Y. Wu, H. Ye, Y. Zhang, Opendmc: an open-source library and performance evaluation for deep-learning-based multi-frame compression, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9685–9688
75. W. Gao, H. Zheng, C. Zhang, K. Zheng, Z. Yu, Y. Li, H. Ye, Y. Zhang, Opendic: an open-source library and performance evaluation for deep-learning-based image compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 202–11 205
76. Y. Li, W. Gao, G. Li, S. Ma, Saliency segmentation oriented deep image compression with novel bit allocation. *IEEE Trans. Image Process.* **34**, 16–29 (2025)
77. C. Zhang, W. Gao, Learned rate control for frame-level adaptive neural video compression via dynamic neural network, in *European Conference on Computer Vision* (Springer, 2024)
78. Z. Yang, W. Gao, G. Li, Y. Yan, Sur-driven video coding rate control for jointly optimizing perceptual quality and buffer control. *IEEE Trans. Image Process.* (2023)
79. F. Shen, Z. Cai, W. Gao, An efficient rate control algorithm for intra frame coding in avs3, in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2021), pp. 3164–3169
80. H. Yuan, W. Gao, J. Wang, Dynamic computational resource allocation for fast inter frame coding in video conferencing applications, in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6
81. W. Gao, Q. Jiang, R. Wang, S. Ma, G. Li, S. Kwong, Consistent quality oriented rate control in hevc via balancing intra and inter frame coding. *IEEE Trans. Ind. Inform.* **18**(3), 1594–1604 (2021)
82. H. Yuan, W. Gao, A new coding unit partitioning mode for screen content video coding, in *Proceedings of the 2021 5th International Conference on Digital Signal Processing* (2021), pp. 66–72
83. W. Gao, On the performance evaluation of state-of-the-art rate control algorithms for practical video coding and transmission systems, in *Proceedings of the 2020 4th International Conference on Video and Image Processing* (2020), pp. 179–185

84. W. Gao, S. Kwong, Q. Jiang, C.-K. Fong, P.H. Wong, W.Y. Yuen, Data-driven rate control for rate-distortion optimization in hevc based on simplified effective initial qp learning. *IEEE Trans. Broadcast.* **65**(1), 94–108 (2018)
85. W. Gao, A multi-objective optimization perspective for joint consideration of video coding quality, in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (IEEE, 2019), pp. 986–991
86. W. Gao, S. Kwong, Y. Jia, Joint machine learning and game theory for rate control in high efficiency video coding. *IEEE Trans. Image Process.* **26**(12), 6074–6089 (2017)
87. W. Gao, S. Kwong, Y. Zhou, H. Yuan, Ssim-based game theory approach for rate-distortion optimized intra frame ctu-level bit allocation. *IEEE Trans. Multimedia* **18**(6), 988–999 (2016)
88. W. Gao, S. Kwong, H. Yuan, X. Wang, Dct coefficient distribution modeling and quality dependency analysis based frame-level bit allocation for hevc. *IEEE Trans. Circuits Syst. Video Technol.* **26**(1), 139–153 (2015)
89. W. Gao, S. Kwong, Phase congruency based edge saliency detection and rate control for perceptual image and video coding, in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2016), pp. 000 264–000 269
90. H. Yuan, W. Gao, Openfastvc: an open source library for video coding fast algorithm implementation, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9660–9663
91. H. Yuan, W. Gao, S. Ma, Y. Yan, Divide-and-conquer-based rdo-free cu partitioning for 8k video compression. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(4), 1–20 (2024)
92. L. Tao, W. Gao, A hardware implementation of entropy encoder for 8k video coding, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
93. Y. Guo, W. Gao, S. Ma, G. Li, Accelerating transform algorithm implementation for efficient intra coding of 8k uhd videos. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* **18**(4), 1–20 (2022)
94. Z. Cai, W. Gao, Efficient fast algorithm and parallel hardware architecture for intra prediction of avs3, in *IEEE International Symposium on Circuits and Systems (ISCAS)* (IEEE, 2021), pp. 1–5
95. W. Gao, H. Yuan, Y. Guo, L. Tao, Z. Cai, G. Li, Openhardwarevc: an open source library for 8k uhd video coding hardware implementation, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 7339–7342
96. W. Gao, H. Yuan, G. Liao, Z. Guo, J. Chen, Pp8k: a new dataset for 8k uhd video compression and processing. *IEEE MultiMedia* (2023)
97. W. Liu, W. Gao, G. Li, S. Ma, T. Zhao, H. Yuan, Enlarged motion-aware and frequency-aware network for compressed video artifact reduction. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
98. L. Xie, X. Mu, G. Li, W. Gao et al., Pku-dpcc: a new dataset for dynamic point cloud compression. *APSIPA Trans. Signal Inf. Process.* **13**(6) (2024)
99. H. Yuan, W. Gao, W. Gao, Opensep: an open source subjective experiment platform, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 248–11 251
100. S. Fan, W. Gao, Screen-based 3d subjective experiment software, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9672–9675
101. J. Wang, W. Gao, G. Li, Zoom to perceive better: No-reference point cloud quality assessment via exploring effective multiscale feature. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
102. J. Wang, W. Gao, G. Li, Applying collaborative adversarial learning to blind point cloud quality measurement. *IEEE Trans. Instrum. Meas.* (2023)
103. H. Li, W. Gao, Information exploration of projected views for point cloud quality measurement. *IEEE Trans. Instrum. Meas.* (2024)
104. B. Qu, X. Liang, S. Sun, W. Gao, Exploring aigc video quality: A focus on visual harmony, video-text consistency and domain distribution gap, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2024)

105. B. Qu, H. Li, W. Gao, Bringing textual prompt to ai-generated image quality assessment, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024)
106. S. Sun, X. Liang, S. Fan, W. Gao, W. Gao, Ve-bench: subjective-aligned benchmark suite for text-driven video editing quality assessment, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
107. Z. Guo, W. Gao, H. Wang, J. Wang, S. Fan, No-reference deep quality assessment of compressed light field images, in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6
108. W. Liu, W. Gao, X. Mu, Fast inter-frame motion prediction for compressed dynamic point cloud attribute enhancement. *Proc. AAAI Conf. Artif. Intell.* **38**(4), 3720–3728 (2024)
109. Z. Yang, W. Gao, X. Lu, Danet: density-adaptive network for geometry-based point cloud compression artifacts removal, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
110. X. Fan, G. Li, D. Li, Y. Ren, W. Gao, T.H. Li, Deep geometry post-processing for decompressed point clouds, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
111. X. Zhang, G. Liao, W. Gao, G. Li, Tdrnet: transformer-based dual-branch restoration network for geometry based point cloud compression artifacts, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
112. Z. Li, G. Li, T.H. Li, S. Liu, W. Gao, Semantic point cloud upsampling. *IEEE Trans. Multimedia* (2022)
113. Z. Li, S. Liu, W. Gao, G. Li, G. Li, S4r: rethinking point cloud sampling via guiding upsampling-aware perception. *IEEE Trans. Multimedia* (2025)
114. R. Zhang, W. Gao, G. Li, T.H. Li, Qinet: decision surface learning and adversarial enhancement for quasi-immune completion of diverse corrupted point clouds. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–14 (2022)
115. R. Bao, Y. Ren, G. Li, W. Gao, S. Liu, Flow-based point cloud completion network with adversarial refinement, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2559–2563
116. J. Chen, G. Li, R. Zhang, T.H. Li, W. Gao, Pointvae: invertible variational autoencoder framework for 3d point cloud generation, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 3216–3220
117. R. Zhang, J. Chen, W. Gao, G. Li, T.H. Li, Pointot: interpretable geometry-inspired point cloud generative model via optimal transport. *IEEE Trans. Circuits Systems Video Technol* **32**(10), 6792–6806 (2022)
118. Z. Li, G. Li, T. Li, S. Liu, W. Gao, Information-growth attention network for image super-resolution, in *Proceedings of the 29th ACM International Conference on Multimedia* (2021), pp. 544–552
119. L. Zhou, W. Gao, G. Li, H. Yuan, T. Zhao, G. Yue, Disentangled feature distillation for light field super-resolution with degradations, in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2023), pp. 116–121
120. L. Zhou, W. Gao, G. Li, End-to-end spatial-angular light field super-resolution using parallax structure preservation strategy, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 3396–3400
121. W. Gao, L. Zhou, L. Tao, A fast view synthesis implementation method for light field applications. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* **17**(4), 1–20 (2021)
122. X. Zhang, W. Gao, G. Li, Q. Jiang, R. Cong, Image quality assessment-driven reinforcement learning for mixed distorted image restoration. *ACM Trans. Multimedia Comput. Commun. Appl.* **19**(1s), 1–23 (2023)
123. X. Zhang, W. Gao, H. Yuan, G. Li, Je 2 net: joint exploitation and exploration in reinforcement learning based image restoration, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2090–2094
124. X. Zhang, W. Gao, Hirl: hybrid image restoration based on hierarchical deep reinforcement learning via two-step analysis, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2445–2449

125. S. Wang, Y. Lu, W. Xia, P. Xia, Z. Wang, W. Gao, Light field angular super-resolution by view-specific queries. *Visual Comput.* 1–16 (2024)
126. Y. Sun, Z. Li, S. Wang, W. Gao, Depth-assisted calibration on learning-based factorization for a compressive light field display. *Opt. Express* **31**(4), 5399–5413 (2023)
127. Y. Sun, Z. Li, L. Li, S. Wang, W. Gao, Optimization of compressive light field display in dual-guided learning, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2075–2079
128. W. Gao, S. Fan, G. Li, W. Lin, A thorough benchmark and a new model for light field saliency detection. *IEEE Trans. Pattern Anal. Mach. Intell.* (2023)
129. J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning. *Adv. Neural Inf. Process. Syst.* **30**, 4077–4087 (2017)
130. A. Dosovitskiy, T. Brox, Generating images with perceptual similarity metrics based on deep networks. *Adv. Neural Inf. Process. Syst.* **29**, 658–666 (2016)
131. J. Wang, Y. Duan, X. Tao, M. Xu, J. Lu, Semantic perceptual image compression with a laplacian pyramid of convolutional networks. *IEEE Trans. Image Process.* **30**, 4225–4237 (2021)
132. L. Xie, M. Hu, X. Bai, Towards hardware-friendly and robust facial landmark detection method, in *International Conference on Neural Information Processing* (Springer, 2022), pp. 432–444
133. L. Xie, M. Hu, X. Bai, online improved vehicle tracking accuracy via unsupervised route generation, in *IEEE 34th International Conference on Tools with Artificial Intelligence* (IEEE, 2022), pp. 788–792
134. C. Cai, L. Chen, X. Zhang, Z. Gao, End-to-end optimized roi image compression. *IEEE Trans. Image Process.* **29**, 3442–3457 (2019)
135. H. Liu, Y. Zhang, H. Zhang, C. Fan, S. Kwong, C.-C.J. Kuo, X. Fan, Deep learning-based picture-wise just noticeable distortion prediction model for image compression. *IEEE Trans. Image Process.* **29**, 641–656 (2019)
136. K. Fischer, F. Brand, A. Kaup, Boosting neural image compression for machines using latent space masking (2021). [arXiv:2112.08168](https://arxiv.org/abs/2112.08168)
137. Y. Hou, L. Zheng, S. Gould, Learning to structure an image with few colors, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 10 116–10 125
138. N. Le, H. Zhang, F. Cricri, R. Ghaznavi-Youvalari, E. Rahtu, Image coding for machines: an end-to-end learned approach, in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2021), pp. 1590–1594
139. N. Le, H. Zhang, F. Cricri, R. Ghaznavi-Youvalari, H.R. Tavakoli, E. Rahtu, Learned image coding for machines: a content-adaptive approach, in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6
140. S. Suzuki, M. Takagi, K. Hayase, T. Onishi, A. Shimizu, Image pre-transformation for recognition-aware image compression, in *2019 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2019), pp. 2686–2690
141. Z. Yang, Y. Wang, C. Xu, P. Du, C. Xu, Q. Tian, Discernible image compression, in *Proceedings of the 28th ACM International Conference on Multimedia* (2020), pp. 1561–1569
142. Y. Kim, H. Jeong, J. Yu, Y. Kim, J. Lee, S.Y. Jeong, H.Y. Kim, End-to-end learnable multi-scale feature compression for vcm. *IEEE Trans. Circuits Syst. Video Technol.* **34**(5), 3156–3167 (2024)
143. H. Zhang, J.I. Ahonen, N. Le, R. Yang, F. Cricri, Competitive learning for achieving content-specific filters in video coding for machines (2024). [arXiv:2406.12367](https://arxiv.org/abs/2406.12367)
144. A. Orzan, A. Bousseau, P. Barla, J. Thollot, Structure-preserving manipulation of photographs, in *Proceedings of the 5th International Symposium on Non-Photorealistic Animation and Rendering* (2007), pp. 103–110
145. S.R. Alvar, I.V. Bajić, Bit allocation for multi-task collaborative intelligence, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2020), pp. 4342–4346

146. S. Singh, S. Abu-El-Haija, N. Johnston, J. Ballé, A. Shrivastava, G. Toderici, End-to-end learning of compressible features, in *2020 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2020), pp. 3349–3353
147. M.A. Shah, B. Raj, Deriving compact feature representations via annealed contraction, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2020), pp. 2068–2072
148. Z. Zhang, M. Wang, M. Ma, J. Li, X. Fan, Msfc: deep feature compression in multi-task network, in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6
149. Y. Yang, R.A. Sarkis, R. El Atrache, T. Loddenkemper, C. Meisel, Video-based detection of generalized tonic-clonic seizures using deep learning. *IEEE J. Biomed. Health Inform.* **25**(8), 2997–3008 (2021)
150. Y. Hu, S. Yang, W. Yang, L.-Y. Duan, J. Liu, Towards coding for human and machine vision: a scalable image coding approach, in *2020 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2020), pp. 1–6
151. S. Wang, S. Wang, W. Yang, X. Zhang, S. Wang, S. Ma, W. Gao, Towards analysis-friendly face representation with scalable feature and texture compression. *IEEE Trans. Multimedia* **24**, 3169–3181 (2021)
152. S. Yang, Y. Hu, W. Yang, L.-Y. Duan, J. Liu, Towards coding for human and machine vision: scalable face image coding. *IEEE Trans. Multimedia* **23**, 2957–2971 (2021)
153. J. Chang, Z. Zhao, L. Yang, C. Jia, J. Zhang, S. Ma, Thousand to one: semantic prior modeling for conceptual coding, in *IEEE international conference on multimedia and expo (ICME)* (IEEE, 2021), pp. 1–6
154. T.M. Hoang, J. Zhou, Y. Fan, Image compression with encoder-decoder matched semantic segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2020), pp. 160–161
155. S. Xia, K. Liang, W. Yang, L.-Y. Duan, J. Liu, An emerging coding paradigm vcm: A scalable coding approach beyond feature and signal, in *2020 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2020), pp. 1–6
156. H. Choi, I.V. Bajić, Scalable image coding for humans and machines. *IEEE Trans. Image Process.* **31**, 2739–2754 (2022)
157. N. Yan, D. Liu, H. Li, F. Wu, Semantically scalable image coding with compression of feature maps, in *2020 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2020), pp. 3114–3118
158. K. Liu, D. Liu, L. Li, N. Yan, H. Li, Semantics-to-signal scalable image compression with learned revertible representations. *Int. J. Comput. Vis.* **129**(9), 2605–2621 (2021)
159. L. Duan, J. Liu, W. Yang, T. Huang, W. Gao, Video coding for machines: a paradigm of collaborative compression and intelligent analytics. *IEEE Trans. Image Process.* **29**, 8680–8695 (2020)
160. G. Lu, X. Ge, T. Zhong, J. Geng, Q. Hu, Preprocessing enhanced image compression for machine vision (2022). [arXiv:2206.05650](https://arxiv.org/abs/2206.05650)

Chapter 8

Compression Artifacts Removal for Image and Video Coding



8.1 Introduction

In the above chapters, common image and video coding methods are introduced in detail [1–73]. Its goal is to strike a balance between minimizing the size of the bit-stream file required for transmission and signal quality [74–83]. For decoded images and videos, post-processing enhancement methods at the decoding end can further improve the quality of the compressed files without adding additional transmission overhead [84–95]. This chapter will introduce in detail the image and video compression artifact removal methods [73, 96] based on deep learning. These methods focus on improving the objective quality and subjective experience [74–83] of compressed images and videos with compression distortion in a post-processing manner. Furthermore, the image and video signals can be expected to be enhanced with higher quality to bring better machine perception performances in diverse tasks [53, 97–129].

Some pioneering methods for compression artifact removal tasks study the quality enhancement problem [83, 96, 130–149] under the image compression algorithm JPEG. Based on the spatial feature extraction focused on image quality enhancement, video quality enhancement additionally considers the inter-frame temporal features. Sequence quality enhancement methods under the video compression algorithm H264/H265 can be roughly divided into the following stages according to the feature extraction paradigm: those that rely on optical flow, motion, deformable convolution, attention mechanisms, and alternative strategies. According to images and videos and their feature extraction paradigms, some representative methods are introduced, including model frameworks and analysis. Furthermore, the research trends of image and video artifact removal methods based on existing analysis are provided at the end of the chapter.

8.2 Compressed Image Artifacts Reduction

To efficiently store and transmit images under limited bandwidth, lossy image compression algorithms like JPEG and WebP are widely used. However, lossy compression algorithms unavoidably introduce undesirable artifacts, which degrade both objective quality (measured by metrics like PSNR and SSIM) and subjective quality (human perception) [74–83]. Common artifacts include blocking, ringing, and blurring effects, often due to the 8×8 block structure and quantization process in formats like JPEG. These artifacts reduce visual fidelity, making image enhancement at the decoding end necessary to restore perceptual quality.

Before the adoption of deep learning [150–161], traditional methods aimed to reduce these artifacts using signal processing techniques. Key traditional approaches include: (1) Adaptive Filtering. Adaptive filters adjust their parameters according to the local context of the image, smoothing block boundaries to mitigate blocking artifacts. (2) Non-Local Means Filtering. This approach leverages redundancy in the image by averaging similar patches across different locations to reduce noise and artifacts. (3) Wavelet-based Methods. Wavelet transforms are applied to compress and decompress data by separating frequencies. By modifying wavelet coefficients, these methods can reduce both blocking and ringing artifacts. (4) Sparse Coding. Sparse representation techniques model image patches as a combination of dictionary atoms, reconstructing compressed images with fewer artifacts. These approaches laid foundational principles, but their limitations in capturing complex features limit their effectiveness in modern high-compression scenarios.

The deep learning-based compression distortion mitigation technique seeks to boost both the quantitative quality and the perceived experience of images through post-processing means. Dong et al. [162] proposed a pioneering compression artifacts reduction method (AR-CNN), which designs a simple convolution neural network to enhance compressed images. AR-CNN consists of four network layers. Each layer contains a convolution layer and an activation layer. The feature dimensions and convolution kernel sizes are $64 \times 9 \times 9$, $32 \times 7 \times 7$, $16 \times 1 \times 1$, and $1 \times 5 \times 5$, respectively. The model takes low-quality compressed images as input and outputs enhanced images with high-quality. The main framework of the AR-CNN is shown in Fig. 8.1.

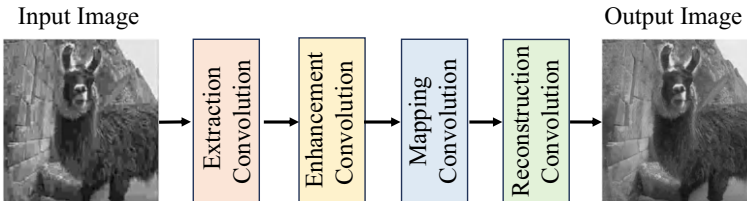
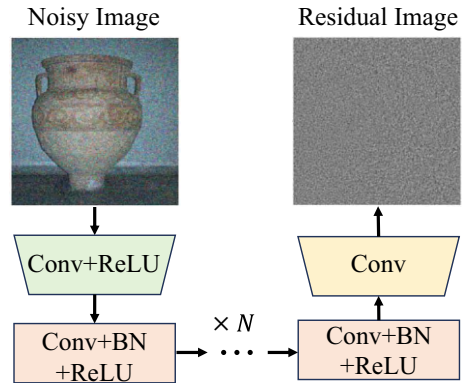


Fig. 8.1 The framework of AR-CNN. Public domain open access image [162]

Fig. 8.2 The framework of DnCNN. Public domain open access image [163]



Research in deep learning-based super-resolution has revealed that learning residual image features often proves more effective than direct image prediction. Therefore, Zhang et al. [163] designed a deep convolutional neural network (DnCNN) with batch normalization to learn the residual image between the distorted image and the raw image. The proposed DnCNN model shows good generalization ability on general image enhancement tasks, such as super-resolution, noise reduction, and deblocking. The main framework of the DnCNN is shown in Fig. 8.2.

Additionally, residual learning in CNNs has proven to be effective in reducing artifacts by predicting the residual between the compressed and the target high-quality image. Instead of reconstructing the entire image, networks learn only the details needed for enhancement, which significantly improves convergence and quality. Self-attention mechanisms allow models to focus on relevant features while ignoring less useful ones, providing a way to enhance important image regions selectively. Attention mechanisms enable networks to emphasize details like edges and textures, improving subjective quality by capturing subtle variations across images. Some models use multi-scale feature extraction, integrating features from different scales (e.g., using pyramid networks) to capture both global structures and fine details. This is particularly beneficial for reducing both blocking and ringing artifacts. Moreover, adding variability to the training data helps models generalize better, particularly in varying lighting conditions, textures, and compression levels. For real-time applications, model size and inference speed are critical. Techniques like knowledge distillation can compress models while maintaining performance, essential for deployment on devices with limited computational resources.

8.3 Compressed Video Artifacts Reduction

The distortion of compressed video mainly comes from the quantization process of prediction residual, which causes irreversible information loss. Traditional compressed video restoration methods usually use certain prior information or combine the encoding and decoding process to enhance the quality. The compressed video restoration method based on deep learning generally processes the compressed video content in a post-processing manner. With powerful feature extraction and feature fusion capabilities, deep neural networks transform the compressed video restoration task into an end-to-end, learnable prediction enhancement process.

Compared with the spatial features that image processing tasks focus on, inter-frame relationship extraction is the key to video processing. According to the form and process of inter-frame feature extraction, deep learning-based video compression artifact reduction approaches can be primarily categorized into: optical flow-based approaches, motion-based approaches, deformable convolution-based approaches, and attention-based approaches. The optical flow-based method simulates inter-frame motion by predicting pixel-level optical flow estimation. Under the guidance of optical flow, the features of adjacent frames are fused by alignment and interpolation. This strategy depends on precise optical flow calculations and employs a coarse alignment method. The motion-based method uses a deep network to implicitly establish inter-frame connections, which can be non-pixel-level, such as block level and super-pixel level. Compared with the above two categories of methods, the deformable convolution-based method is more robust to the impact of compression distortion on inter-frame motion estimation. It benefits from the deformable convolution's ability to more flexibly extract effective spatial-temporal features. Attention-based methods believe that features at a certain spatial-temporal position in compressed videos should receive more attention and determine the quality of inter-frame feature extraction. Some simple attention modeling methods include channel attention, spatial attention, and temporal attention. In addition, other creative and generalized attention mechanisms are being widely used in various image and video processing tasks, including self-attention mechanism and cross-attention mechanism [6, 41].

This section will introduce some representative models of the above methods. The pros and cons of these methods are also briefly summarized.

8.3.1 Optical Flow-based Methods

Among the methods based on explicit optical flow estimation, the early representative work is mainly introduced. Motion estimation and compensation, as well as the quantization process of residual information, lead to blurring and motion loss in compressed videos. At the same time, the compression coding framework

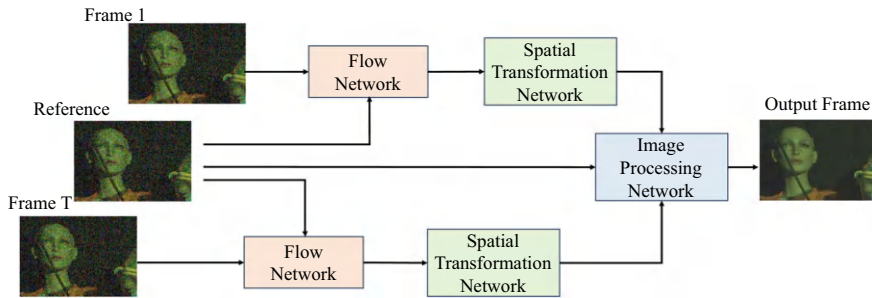


Fig. 8.3 The framework of TOFlow. Public domain open access image [164]

based on block partitioning causes block effects. Therefore, Xue et al. [164] proposed a task-oriented flow (TOFlow) model for removing video block effects, which simultaneously performs motion analysis and video processing, as shown in Fig. 8.3.

Specifically, the optical flow estimation module takes a compressed sequence with seven consecutive images as input, producing a motion field that maps the remaining six reference frames to the central target frame. TOFlow refers to the four-layer optical flow estimation model (SPynet) [165] and extracts the optical flow estimation from all reference frames to the intermediate frames. Utilizing this optical flow data, a spatial transformation network (STN) is employed to facilitate the alignment from a reference frame to an intermediate one. Unlike traditional pixel-by-pixel warping, STN accomplishes feature alignment across frames in a comprehensive, end-to-end process. Ultimately, the aligned frame's features are merged with those of the target frame to generate a high-fidelity intermediate frame [166, 167].

8.3.2 Motion-based Methods

For motion-based approaches, Yang et al. [168] introduced a multi-frame quality enhancement model known as MFQE involving implicit inter-frame motion estimation. This approach examines the variation in the objective quality of compressed video over time and suggests a method to improve the intermediate low-quality frames by utilizing high-quality frames from adjacent frames. The main framework of the MFQE is shown in Fig. 8.4, which is divided into two parts: high-quality frame detection and multi-frame enhancement. The multi-frame enhancement segment is comprised of a motion compensation network and a fusion enhancement network. The motion compensation component is designed to forecast motion estimation and compensation from high-quality frames to low-quality ones, whereas the enhancement component primarily aims to acquire residual details of the refined frames [8, 9, 42].

Video compression algorithms involving real-time requirements (such as Low Latency Mode) generally adopt the Group of Pictures mechanism, which causes

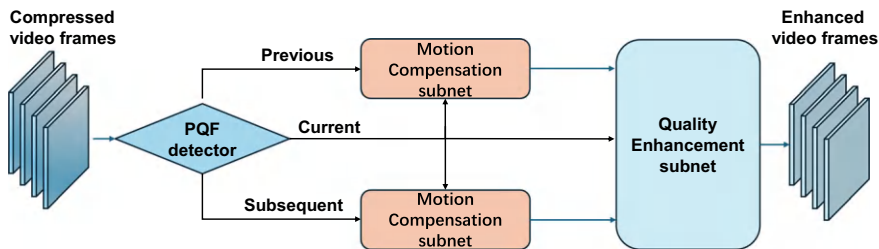


Fig. 8.4 The framework of MFQE. Public domain open access image [168]

inconsistent quality between frames. To avoid the code stream overhead occupied by the indicator of high-quality frames, MFQE first predicts the position of high-quality frames and low-quality frames through the support vector machine (SVM) algorithm. Next, the inter-frame motion relationship from high-quality adjacent frames to low-quality intermediate frames is predicted and compensated in the motion compensation subnetwork. Unlike the optical flow-based method, the compensated low-quality frames are supervised by an additional loss function. Finally, all compensated low-quality frames and intermediate target frame are fused into the quality enhancement subnetwork.

The multi-frame enhancement method mainly uses the high-quality frames in two adjacent frames to enhance the low-quality frames at the frame level, which ignores some high-quality detail areas in the low-quality frames. Therefore, the enhancement framework that only relies on high-quality frames needs to be further optimized. How to effectively utilize the high-quality detail areas in low-quality frames will help the existing methods to further improve the enhancement quality of compressed videos.

Xv et al. [169] proposed the NL-CLSTM model, which can capture hidden spatial-temporal information in adjacent frames. With the help of the non-local mean mechanism in the image denoising task and the ConvLSTM structure in the spatial-temporal sequence prediction task, NL-CLSTM can still learn effective spatial-temporal information in the absence of explicit motion estimation and compensation information. NL-CLSTM contains the encoding module, the NL-ConvLSTM module and the decoding module, which are used to extract single-frame image features, learn spatial-temporal associations between frames, and residual decoding of high-dimensional features, respectively. The general ConvLSTM structure can capture spatial-temporal information in a video sequence of any number of frames. However, it is difficult to process videos with large displacement and blurred motion. Compared with motion compensation, the introduction of non-local mechanisms can effectively capture long-range motion. The main framework of the NL-CLSTM is shown in Fig. 8.5.

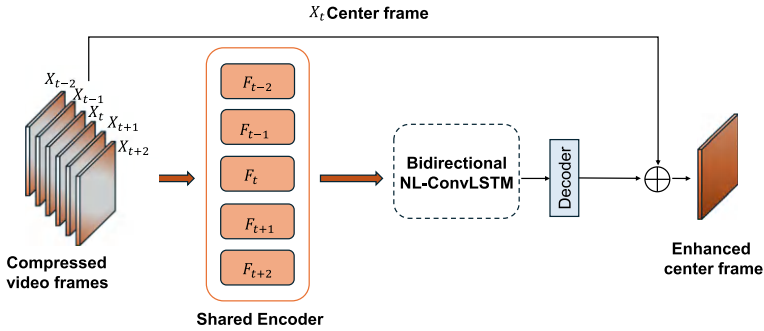


Fig. 8.5 The framework of NL-CLSTM. Public domain open access image [169]

8.3.3 Deformable Convolution-based Methods

For deformable convolution-based approaches, the EDVR model and the STDF model have taken a step forward in the quality gain of compressed videos. Previous methods based on explicit optical flow estimation or implicit motion compensation usually require an inter-frame alignment process. However, they rely heavily on accurate optical flow estimation or motion estimation, especially for videos with large displacement motion. At the same time, in the frame fusion process, previous methods rarely consider the feature differences at different spatial-temporal positions.

Wang et al. [170] introduced the EDVR approach, initially crafting a pyramid-based deformable alignment unit and a spatial-temporal fusion mechanism to tackle extensive displacement issues. Contrasting with the direct optical flow technique, EDVR's deformable alignment is executed within the feature space. In addition to video super-resolution and deblurring, EDVR can also handle video compression artifact reduction task. The main framework of the EDVR is shown in Fig. 8.6.

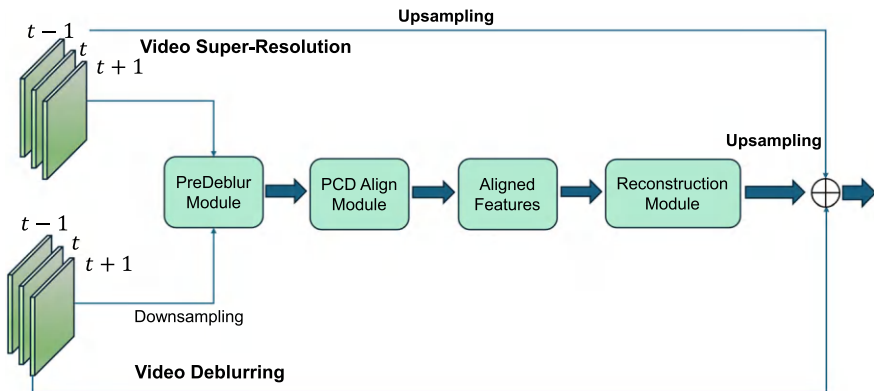


Fig. 8.6 The framework of EDVR. Public domain open access image [170]

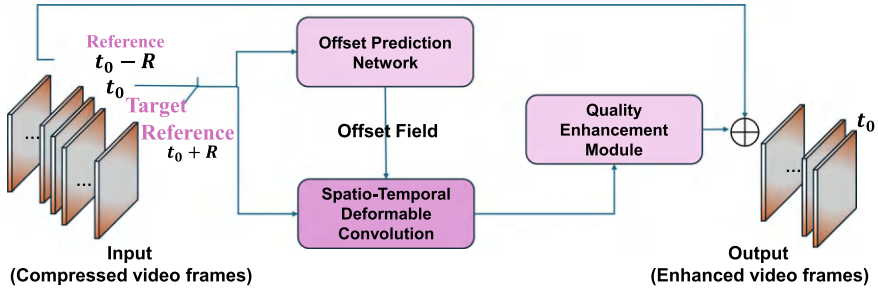


Fig. 8.7 The framework of STDF. Public domain open access image [171] (<https://ojs.aaai.org/index.php/AAAI/article/view/6697>)

Deng et al. [171] proposed the STDF method, which can quickly and effectively extract the temporal information of the sequence. STDF mainly consists of three modules, including the offset field prediction (OPN) module, the spatial-temporal deformation convolution (STDC) module and the quality enhancement (QE) module. The offset field prediction module predicts the offset between the neighborhood points of the central pixel and the similar pixels in its adjacent frames. Unlike the optical flow estimation between video frames, the offset field prediction module can utilize the information of multiple frames in a larger range. Moreover, compared with the optical flow estimation process, richer potential spatiotemporal motion is captured by OPN. The STDC module applies the obtained offset field to the convolution kernel of fixed size, which can greatly expand the convolution receptive field. Considering the feature differences at different sampling positions, the STDC module adds additional learnable feature weights. The quality enhancement module receives all deformed intra-frame features and inter-frame features to learn the residual features of the target frame. The main framework of the STDF is shown in Fig. 8.7.¹

The process of deformable convolution is briefly as follows [171]. Given a low-quality compressed video sequence $\{F_{t_0-R}^{LQ}, \dots, F_{t_0}^{LQ}, \dots, F_{t_0+R}^{LQ}\}$, the regular convolution feature extraction is defined as follows:

$$G(p) = \sum_{t_0-R}^{t_0+R} \sum_{k=1}^{K^2} V_{t,k} \cdot F_t^{LQ}(q + q_k), \quad (8.1)$$

where G and K represent the generated feature map and convolution kernel size, respectively. q and q_k record the arbitrary spatial position and the corresponding square sampling neighborhood point, respectively. Assume that $\sigma_{(t,q)} \in R^{2K^2}$ describes the predicted offset field, which is used to update the square sampling positions:

$$q_k \leftarrow q_k + \sigma_{(t,q),k}. \quad (8.2)$$

¹ <https://ojs.aaai.org/index.php/AAAI/article/view/6697>.

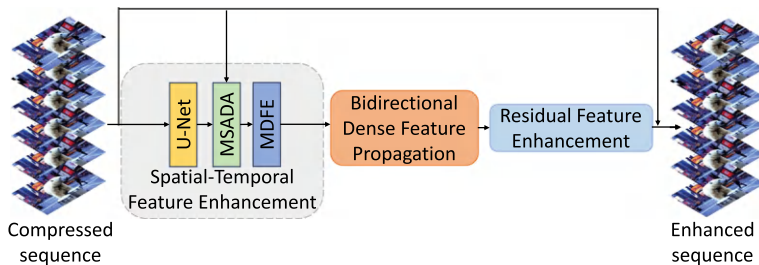


Fig. 8.8 The framework of EMAFA. ©2024 IEEE Reprinted, with permission, from Ref. [73]

The learnable offset field enables deformable convolution to flexibly sample richer features in adjacent frames.

To further expand the receptive field of deformable convolution, Liu et al. [73] replaced the initial 3×3 sampling position with a deformable convolution with a dilation factor. Although increasing the convolution kernel size can achieve the expansion of the receptive field, this will cause a sharp increase in the number of model parameters. It also found that existing methods tend to enhance the high-frequency structural information of compressed videos while ignoring non-high-frequency information or texture information. Combining the above analysis, Liu et al. [73] adopted directional convolution instead of general convolution and propose the EMAFA model. The framework is shown in Fig. 8.8. It includes three components: spatio-temporal feature alignment (STFA), bidirectional dense feature propagation (BDFP) and residual feature enhancement (RFE). In STFA, dilated deformable convolutions are applied to all adjacent frames to achieve spatiotemporal feature extraction. Specifically, three deformable convolution operators with different dilation factors are designed to generate coarse inter-frame features. Next, BDFP refers to the common bidirectional inter-frame feature propagation framework to refine the coarse features. The proposed pixel-level and patch-level deformable convolution modules further expand the feature receptive field. Finally, RFE is mainly used to predict the residual between the compressed frame and the original frame, which includes the quantization error and motion error generated by the compression process.

8.3.4 Attention-based Methods

Most methods enhance the quality of compressed videos at the pixel level or frame level, which ignore the block-level settings in video compression algorithms. The existing video compression framework is based on the block division mechanism, that is, the distortion degree of each sub-block is uneven. Therefore, these sub-blocks should be adaptively selected and utilized in the quality enhancement process. For example, given a low-quality target sub-block, blocks with similar structure and quality in adjacent frames are searched. Intuitively, the weights of these reference blocks

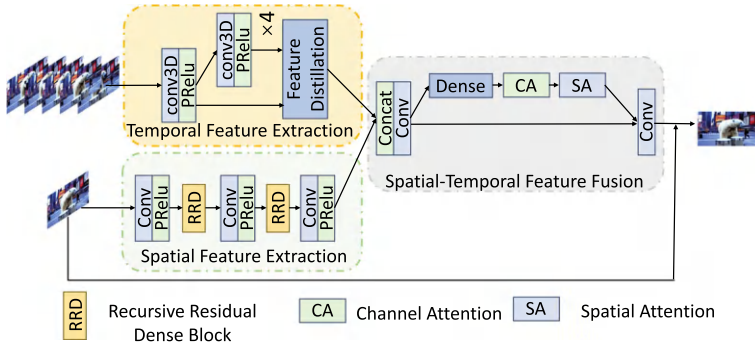


Fig. 8.9 The framework of PSTQE. ©2021 IEEE Reprinted, with permission, from Ref. [172]

in the feature fusion enhancement process should be different. On the other hand, for fast-moving and slow-moving video sequences, the corresponding sub-block enhancement methods should be processed separately. Some experimental phenomena show that spatial information is more efficient in enhancing fast-moving sub-blocks, while temporal information is generally used for slow-moving sub-blocks. In other words, temporal and spatial information need to be processed adaptively.

Ding et al. [172] introduced a patch-based spatial-temporal quality enhancement (PSTQE) model that dynamically employs and boosts compressed sub-blocks via temporal and spatial data. The PSTQE model encompasses three sub-networks: a sub-network for partial feature extraction, a sub-network for temporal feature extraction, and a sub-network for spatiotemporal feature fusion. To be specific, the initial sub-network carries out spatial feature extraction of the intermediate target frame. The subsequent sub-network extracts temporal details from various frames and employs the feature distillation module of temporal attention and the spatial attention mechanism to adaptively leverage compressed sub-blocks in successive frames. The temporal attention sub-network's function is to accentuate the significant features of neighboring frames, whereas the spatial attention sub-network focuses on highlighting the efficient compressed sub-blocks. The final sub-network adaptively integrates the previously gathered spatiotemporal features to accomplish quality enhancement of compressed frames. The channel attention module selectively picks out effective channel features within the integrated features. The foundational structure of the PSTQE model is depicted in Fig. 8.9.

8.3.5 Other Methods

Most methods focus on improving the objective evaluation metric or the signal fidelity of compressed video. However, the enhanced results optimized by loss functions related to PSNR or SSIM are not always friendly to human perception. For image

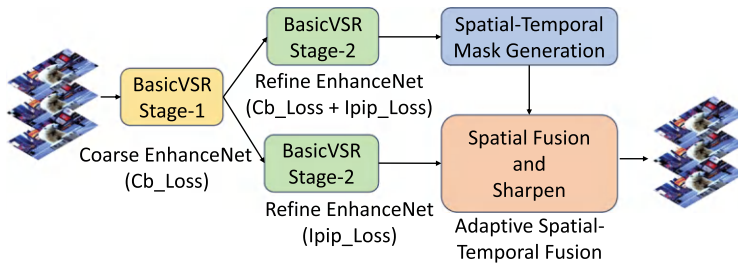


Fig. 8.10 The framework of ASTF. ©2021 IEEE. Reprinted, with permission, from Ref. [174]

data, Zhang et al. [173] proposed an evaluation metric for subjective image similarity, namely Learned Perceptual Image Patch Similarity (LPIPS). LPIPS captures the perceptual differences of images by utilizing a deep learning network (usually a pre-trained convolution neural network such as VGG), which is more in line with human visual perception than traditional metrics such as L2 and SSIM. Some experiments show that the deep model trained by PSNR and SSIM objective metrics tends to produce smooth restoration effects, while the model trained by LPIPS loss can produce detailed information with noisy points. Therefore, Zheng et al. [174] proposed the ASTF model to simultaneously enhance smooth areas and texture areas. The enhancement process of the ASTF model includes two stages. In the first part, the BasicVSR model trained with Charbonnier loss produces a rough high-fidelity image. In the second stage, EnhanceNet1 trained with LPIPS restores texture details that are consistent with human eye perception, while EnhanceNet2 trained with both LPIPS and Charbonnier loss is dedicated to smoothing information. Finally, the spatiotemporal mask module is used to fuse the texture information of EnhanceNet1 and the smoothing information of EnhanceNet2. The illustrations of ASTF is shown in Fig. 8.10.

Some studies have attempted to migrate video deblurring methods related to camera motion, target motion, and depth change to the compression artifact reduction task. Zhou et al. [175] proposed the STFAN to solve the artifacts caused by optical flow information in blurred videos. STFAN integrates the alignment process and deblurring process into the same framework. A new filter adaptive convolutional layer network is proposed to align the smooth features of the previous frame and remove the spatial blur features in the current frame. Shen et al. [176] proposed a video interpolation and enhancement model based on a pyramid cycle structure. The designed pyramid module has the ability to adjust the spatial receptive field range and the temporal range to control the model complexity and restoration degree.

Combining the encoding and decoding process or targeting specific types of compression distortion is another paradigm for enhancing the quality of compressed videos. The HLVC model proposed by Yang et al. [177] effectively improved the coding performance through hierarchical quality layers and recurrent enhancement

Table 8.1 Comparison of compressed video artifacts reduction methods

Method	Strengths	Limitations
Optical flow	• Accurate motion estimation, capturing pixel-level motion details • Reduces blocking artifacts, improving compressed video quality	• Highly dependent on the accuracy of optical flow estimation • Limited effectiveness for large motion displacements
Motion	• Utilizes high-quality information from adjacent frames to enhance low-quality frames • Avoids pixel-level estimation, reducing computational cost	• Overlooks high-quality detail areas within low-quality frames • Quality enhancement can be limited by the quality of adjacent frames
Deformable convolution	• Flexibly extracts spatial-temporal features • Capable of handling large displacements • Robust without relying on precise motion estimation	• High computational complexity with large model parameters • Limited effect on low-quality videos lacking sufficient features
Attention	• Adapts to crucial spatial-temporal features selectively • Effective on both slow and fast-moving video segments	• Dependent on the quality and similarity of selected reference blocks • High complexity, requiring more computational resources
Other methods	• Focuses on enhancing subjective visual quality (e.g., using LPIPS metric) • Some models address mixed distortions, like STFAN’s ability to handle blurred video	• Enhanced details may introduce noise artifacts • Optimizations for mixed distortions may not be as effective as specific artifact removal methods

Source Author

networks. Zhang et al. [178] designed a novel DCNN structure to address the distortion problem of the mouth area in video conference content. The structure analyzes the correlation between audio and video to remove compression artifacts in the facial area.

Videos transmitted over the mobile Internet usually exhibit mixed distortion, which includes downsampling distortion, blur distortion, noise, and compression distortion. At the same time, the different degrees of single distortion further aggravate the complexity of mixed distortion. Therefore, Zhang et al. [143–145] designed a series of methods for hybrid distortion restoration. These methods adopt the reinforcement learning paradigm to explore and utilize random tools of unknown distortion combinations to improve video quality more flexibly.

We present a comparative analysis of the methods listed in this section, as shown in Table 8.1.

8.4 Summary

In this chapter, some post-processing methods for removing compression artifacts from images and videos based on deep learning are briefly introduced. Based on the feature extraction method, the above methods are simply classified into several categories. At the same time, the selected methods generally represent the development of compression artifact removal methods. General image and video quality enhancement methods provide a certain processing paradigm for the compression distortion removal task. However, the latter requires additional consideration of the characteristics of the compression algorithm. Future research trends are roughly as follows: joint processing for human eye perception and machine perception, multi-modal processing combining text and speech, and synchronous development with the quality assessment system.

Exercises

1. Please briefly describe the common image compression standards and algorithms.
2. Please briefly describe the common video compression standards and algorithms.
3. Please outline the difference between image compression artifact removal and general image restoration and enhancement tasks, such as image deblurring and denoising.
4. Please specify the difference between video compression artifact removal and general image restoration and enhancement tasks, such as video deblurring and completion.
5. Please indicate how many categories image compression artifact removal methods with deep learning are divided into.
6. Please specify the number of categories into which existing deep learning-based video compression artifact removal methods are divided.
7. Please consider whether adding relevant captions to image content can further improve the quality of compressed images, such as by influencing quantization parameters or other compression settings.
8. Please identify the new challenges that inter-frame temporal features face in videos affected by compression distortion.
9. Please explain the guidance that deep learning-based image compression methods provide for post-processing distortion removal tasks.
10. Please describe how to combine diffusion-based image denoising methods with compression distortion removal tasks.

References

1. W. Gao, G. Li, H. Yuan, R. Hamzaoui, Z. Li, S. Liu, Apccpa'22: 1st international workshop on advances in point cloud compression, processing and analysis, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 7392–7393
2. W. Gao, G. Li, Point cloud compression, enhancement and applications: from 3d perception to large models, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 292–11 293
3. H. Sun, G. Li, S. Chen, L. Li, W. Gao, Special topic on 3d point cloud processing and applications. *ZTE Commun.* **21**(4), 1 (2023)
4. W. Gao, W. Gao, X. Mu, C. Peng, G. Li, Overview and comparison of avs point cloud compression standard. *APSIPA Trans. Signal Inf. Process.* (2024)
5. W. Gao, S. Wang, X. Zhang, S. Kwong, Editorial for special issue on deep learning-based data compression. *APSIPA Trans. Signal Inf. Process.* **13**(6) (2024)
6. W. Gao, H. Ye, G. Li, H. Zheng, Y. Wu, L. Xie, OpenPointCloud: an open-source algorithm library of deep learning based point cloud compression, in *ACM International Conference on Multimedia* (2022), pp. 7347–7350
7. Y. Zhang, W. Gao, G. Li, Openpointcloud-v2: a deep learning based open-source algorithm library of point cloud processing, in *Proceedings of the 1st International Workshop on Advances in Point Cloud Compression, Processing and Analysis* (2022), pp. 51–55
8. L. Xie, W. Gao, Learningpcc: a pytorch library for learning-based point cloud compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 234–11 238
9. L. Xie, W. Gao, Pchmvision: an open-source library of point cloud compression for human and machine vision, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 239–11 243
10. T. Qin, G. Li, W. Gao, S. Liu, Multi-grained point cloud geometry compression via dual-model prediction with extended octree. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(9), 1–30 (2024)
11. Y. Shao, W. Gao, S. Liu, G. Li, Advanced patch-based affine motion estimation for dynamic point cloud geometry compression. *Sensors* **24**(10), 3142 (2024)
12. Y. Shao, F. Song, W. Gao, S. Liu, G. Li, Texture-guided graph transform optimization for point cloud attribute compression. *Appl. Sci.* **14**(10), 4094 (2024)
13. Y. Shao, X. Yang, W. Gao, S. Liu, G. Li, 3d point cloud attribute compression using diffusion-based texture-aware intra prediction. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
14. J. Zhang, Y. Chen, G. Liu, W. Gao, G. Li, Efficient point cloud attribute compression framework using attribute-guided graph fourier transform, in *ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2024), pp. 8426–8430
15. W. Gao, H. Yuan, G. Li, Z. Li, H. Yuan, Low complexity coding unit decision for video-based point cloud compression. *IEEE Trans. Image Process.* **33**, 149–162 (2023)
16. Y. Shao, G. Li, Q. Zhang, W. Gao, S. Liu, Nonrigid registration-based progressive motion compensation for point cloud geometry compression. *IEEE Trans. Geosci. Remote Sens.* (2023)
17. F. Song, G. Li, X. Yang, W. Gao, S. Liu, Block-adaptive point cloud attribute coding with region-aware optimized transform. *IEEE Trans. Circuits Syst. Video Technol.* (2023)
18. Y. An, Y. Shao, G. Li, W. Gao, S. Liu, A fast motion estimation method with hamming distance for lidar point cloud compression, in *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2022), pp. 1–5
19. H. Yuan, W. Gao, G. Li, Z. Li, Rate-distortion-guided learning approach with cross-projection information for v-pcc fast cu decision, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 3085–3093
20. F. Song, G. Li, W. Gao, T.H. Li, Rate-distortion optimized graph for point cloud attribute coding. *IEEE Signal Process. Lett.* **29**, 922–926 (2022)

21. F. Song, G. Li, X. Yang, W. Gao, T.H. Li, Fine-grained correlation representation for graph-based point cloud attribute compression, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
22. F. Shen, W. Gao, A rate control algorithm for video-based point cloud compression, in *2021 International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2021), pp. 1–5
23. F. Song, Y. Shao, W. Gao, H. Wang, T. Li, Layer-wise geometry aggregation framework for lossless lidar point cloud compression. *IEEE Trans. Circuits Syst. Video Technol.* **31**(12), 4603–4616 (2021)
24. Z. Cai, W. Gao, G. Li, W. Gao, Distortion propagation model-based v-pcc rate control for 3d point cloud broadcasting. *IEEE Trans. Broadcast.* 1–13 (2024)
25. X. Mu, W. Gao, H. Yuan, S. Wang, G. Li, Adaptive lpu decision for dynamic point cloud compression. *IEEE Signal Process. Lett.* (2024)
26. C. Peng, W. Gao, Laplacian matrix learning for point cloud attribute compression with ternary search-based adaptive block partition, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 10 412–10 420
27. Y. Wang, W. Gao, X. Mu, H. Yuan, Rate control optimization for joint geometry and attribute coding of lidar point clouds, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
28. T. Guo, H. Yuan, T. Wang, W. Gao, Graph filter-based fast motion matching for inter frame coding of mpeg g-pcc, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 1151–1155
29. L. Xie, W. Gao, H. Zheng, G. Li, Spcgc: scalable point cloud geometry compression for machine vision, in *2024 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2024), pp. 17 272–17 278
30. L. Xie, W. Gao, H. Zheng, H. Ye, Semantic-aware visual decomposition for point cloud geometry compression, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 595–595
31. Z. Qi, W. Gao, Variable-rate point cloud geometry compression based on feature adjustment and interpolation, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 63–72
32. Z. Yu, W. Gao, When dynamic neural network meets point cloud compression: Computation-aware variable rate and checkerboard context, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 600–600
33. L. Xie, W. Gao, S. Fan, Z. Yao, Pdnet: parallel dual-branch network for point cloud geometry compression and analysis, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 596–596
34. L. Xie, W. Gao, H. Zheng, End-to-end point cloud geometry compression and analysis with sparse tensor, in *International Workshop on Advances in Point Cloud Compression, Processing and Analysis* (2022), pp. 27–32
35. C. Fu, G. Li, R. Song, W. Gao, S. Liu, Octattention: Octree-based large-scale contexts model for point cloud compression. *Proc. AAAI Conf. Artif. Intell.* **36**(1), 625–633 (2022)
36. C. Zhang, W. Gao, Adapcc: adaptive rate control and rate-distortion-complexity optimization for dynamic point cloud compression, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
37. K. Wang, W. Gao, Unipcgc: Towards practical point cloud geometry compression via an efficient unified approach, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
38. K. Zheng, W. Gao, H. Zheng, Octree-based learned point cloud geometry compression: A lossy perspective, in *2025 Data Compression Conference (DCC)* (IEEE, 2025)
39. Z. Qi, H. Liu, W. Gao, Variable-rate point cloud compression with channel-wise entropy modeling, in *2025 Data Compression Conference (DCC)* (IEEE, 2025)
40. H. Zheng, W. Gao, Z. Yu, T. Zhao, G. Li, Viewpcgc: view-guided learned point cloud geometry compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 7152–7161

41. L. Xie, W. Gao, H. Zheng, G. Li, Roi-guided point cloud geometry compression towards human and machine vision, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 3741–3750
42. Y. Wu, L. Xie, S. Sun, W. Gao, Y. Yan, Adaptive intra period size for deep learning-based screen content video coding, in *IEEE International Conference on Multimedia and Expo Workshops*(2024), pp. 1–6
43. H. Zheng, W. Gao, End-to-end rgb-d image compression via exploiting channel-modality redundancy. *Proc. AAAI Conf. Art. Intell.* **38**(7), 7562–7570 (2024)
44. L. Tao, W. Gao, G. Li, C. Zhang, Adanic: towards practical neural image compression via dynamic transform routing, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 16 879–16 888
45. Y. Wu, W. Gao, End-to-end lossless compression of high precision depth maps guided by pseudo-residual (2022). [arXiv:2201.03195](https://arxiv.org/abs/2201.03195)
46. Y. Wu, Z. Qi, H. Zheng, L. Tao, W. Gao, Deep image compression with latent optimization and piece-wise quantization approximation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 1926–1930
47. W. Gao, L. Tao, L. Zhou, D. Yang, X. Zhang, Z. Guo, Low-rate image compression with super-resolution learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2020), pp. 154–155
48. W. Gao, S. Sun, H. Zheng, Y. Wu, H. Ye, Y. Zhang, Opendmc: an open-source library and performance evaluation for deep-learning-based multi-frame compression, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9685–9688
49. W. Gao, H. Zheng, C. Zhang, K. Zheng, Z. Yu, Y. Li, H. Ye, Y. Zhang, Opendic: An open-source library and performance evaluation for deep-learning-based image compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 202–11 205
50. Y. Li, W. Gao, G. Li, S. Ma, Saliency segmentation oriented deep image compression with novel bit allocation. *IEEE Trans. Image Process.* **34**, 16–29 (2025)
51. C. Zhang, W. Gao, Learned rate control for frame-level adaptive neural video compression via dynamic neural network, in *European Conference on Computer Vision* (Springer, 2024)
52. Z. Yang, W. Gao, G. Li, Y. Yan, Sur-driven video coding rate control for jointly optimizing perceptual quality and buffer control. *IEEE Transactions on Image Processing* (2023)
53. Y. Li, Z. Yu, X. Chen, S. Wang, H. Yuan, W. Gao, Zero-shot quantization for large-kernels via shape-based distribution and diversity self-distillation, in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2025)
54. X. Chen, Y. Wang, Y. Li, X. Ling, M. Li, R. Liu, M. Ouyang, K. Zhao, T. Guan, Y. He, Low bit-width zero-shot quantization with soft feature-infused hints for iot systems. *IEEE Internet Things J.* (2024)
55. F. Shen, Z. Cai, W. Gao, An efficient rate control algorithm for intra frame coding in avs3, in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2021), pp. 3164–3169
56. H. Yuan, W. Gao, J. Wang, Dynamic computational resource allocation for fast inter frame coding in video conferencing applications, in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6
57. W. Gao, Q. Jiang, R. Wang, S. Ma, G. Li, S. Kwong, Consistent quality oriented rate control in hevc via balancing intra and inter frame coding. *IEEE Trans. Ind. Inform.* **18**(3), 1594–1604 (2021)
58. H. Yuan, W. Gao, A new coding unit partitioning mode for screen content video coding, in *Proceedings of the 2021 5th International Conference on Digital Signal Processing* (2021), pp. 66–72
59. W. Gao, On the performance evaluation of state-of-the-art rate control algorithms for practical video coding and transmission systems, in *Proceedings of the 2020 4th International Conference on Video and Image Processing* (2020), pp. 179–185

60. W. Gao, S. Kwong, Q. Jiang, C.-K. Fong, P.H. Wong, W.Y. Yuen, Data-driven rate control for rate-distortion optimization in hevc based on simplified effective initial qp learning. *IEEE Trans. Broadcast.* **65**(1), 94–108 (2018)
61. W. Gao, A multi-objective optimization perspective for joint consideration of video coding quality, in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (IEEE, 2019), pp. 986–991
62. W. Gao, S. Kwong, Y. Jia, Joint machine learning and game theory for rate control in high efficiency video coding. *IEEE Trans. Image Process.* **26**(12), 6074–6089 (2017)
63. W. Gao, S. Kwong, Y. Zhou, H. Yuan, Ssim-based game theory approach for rate-distortion optimized intra frame ctu-level bit allocation. *IEEE Trans. Multimedia* **18**(6), 988–999 (2016)
64. W. Gao, S. Kwong, H. Yuan, X. Wang, Dct coefficient distribution modeling and quality dependency analysis based frame-level bit allocation for hevc. *IEEE Trans. Circuits Syst. Video Technol.* **26**(1), 139–153 (2015)
65. W. Gao, S. Kwong, Phase congruency based edge saliency detection and rate control for perceptual image and video coding, in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2016), pp. 000 264–000 269
66. H. Yuan, W. Gao, Openfastvc: an open source library for video coding fast algorithm implementation, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9660–9663
67. H. Yuan, W. Gao, S. Ma, Y. Yan, Divide-and-conquer-based rdo-free cu partitioning for 8k video compression. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(4), 1–20 (2024)
68. L. Tao, W. Gao, A hardware implementation of entropy encoder for 8k video coding, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
69. Y. Guo, W. Gao, S. Ma, G. Li, Accelerating transform algorithm implementation for efficient intra coding of 8k uhd videos. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* **18**(4), 1–20 (2022)
70. Z. Cai, W. Gao, Efficient fast algorithm and parallel hardware architecture for intra prediction of avs3, in *IEEE International Symposium on Circuits and Systems (ISCAS)* (IEEE, 2021), pp. 1–5
71. W. Gao, H. Yuan, Y. Guo, L. Tao, Z. Cai, G. Li, Openhardwarevc: an open source library for 8k uhd video coding hardware implementation, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 7339–7342
72. W. Gao, H. Yuan, G. Liao, Z. Guo, J. Chen, Pp8k: a new dataset for 8k uhd video compression and processing. *IEEE MultiMedia* (2023)
73. W. Liu, W. Gao, G. Li, S. Ma, T. Zhao, H. Yuan, Enlarged motion-aware and frequency-aware network for compressed video artifact reduction. *IEEE Trans. Circuits Syst. Video Technol.* **34**(10), 10 339–10 352 (2024)
74. L. Xie, X. Mu, G. Li, W. Gao et al., Pku-dpcc: A new dataset for dynamic point cloud compression. *APSIPA Trans. Signal Inf. Process.* **13**(6) (2024)
75. H. Yuan, W. Gao, W. Gao, Opensep: an open source subjective experiment platform, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 248–11 251
76. S. Fan, W. Gao, Screen-based 3d subjective experiment software, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9672–9675
77. J. Wang, W. Gao, G. Li, Zoom to perceive better: No-reference point cloud quality assessment via exploring effective multiscale feature. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
78. J. Wang, W. Gao, G. Li, Applying collaborative adversarial learning to blind point cloud quality measurement. *IEEE Trans. Instrum. Meas.* (2023)
79. H. Li, W. Gao, Information exploration of projected views for point cloud quality measurement. *IEEE Trans. Instrum. Meas.* (2024)
80. B. Qu, X. Liang, S. Sun, W. Gao, Exploring aigc video quality: a focus on visual harmony, video-text consistency and domain distribution gap, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2024)

81. B. Qu, H. Li, W. Gao, Bringing textual prompt to ai-generated image quality assessment, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024)
82. S. Sun, X. Liang, S. Fan, W. Gao, W. Gao, Ve-bench: subjective-aligned benchmark suite for text-driven video editing quality assessment, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
83. Z. Guo, W. Gao, H. Wang, J. Wang, S. Fan, No-reference deep quality assessment of compressed light field images, in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6
84. G. Li, G. Wei, W. Gao, *Point Cloud Compression: Technologies and Standardization* (Springer Nature, 2024)
85. G. Li, W. Gao, W. Gao, Introduction, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 1–28
86. G. Li, W. Gao, W. Gao, Background knowledge, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 29–51
87. G. Li, W. Gao, W. Gao, Predictive coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 53–70
88. G. Li, W. Gao, W. Gao, Transform coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 71–96
89. G. Li, W. Gao, W. Gao, Quantization techniques, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 97–112
90. G. Li, W. Gao, W. Gao, Entropy coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 113–133
91. G. Li, W. Gao, W. Gao, Mpeg geometry-based point cloud compression (g-pcc) standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 135–165
92. G. Li, W. Gao, W. Gao, Avs point cloud compression standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 167–197
93. G. Li, W. Gao, W. Gao, Mpeg video-based point cloud compression (v-pcc) standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 199–218
94. G. Li, W. Gao, W. Gao, Mpeg ai-based 3d graphics coding standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 219–241
95. G. Li, W. Gao, W. Gao, Future work, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 243–250
96. Z. Yang, W. Gao, X. Lu, Danet: density-adaptive network for geometry-based point cloud compression artifacts removal, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
97. X. Zang, W. Gao, G. Li, H. Fang, C. Ban, Z. He, H. Sun, A baseline investigation: Transformer-based cross-view baseline for text-based person search, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 7737–7746
98. G. Liao, W. Gao, G. Li, J. Wang, S. Kwong, Cross-collaborative fusion-encoder network for robust rgb-thermal salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **32**(11), 7646–7661 (2022)
99. W. Gao, G. Liao, S. Ma, G. Li, Y. Liang, W. Lin, Unified information fusion network for multi-modal rgb-d and rgb-t salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **32**(4), 2091–2106 (2021)
100. Y. Chen, S. Sun, G. Li, W. Gao, T.H. Li, Closing the gap between theory and practice during alternating optimization for gans. *IEEE Trans. Neural Netw. Learn. Syst.* (2023)
101. Y. Chen, C. Jin, G. Li, T.H. Li, W. Gao, Mitigating label noise in gans via enhanced spectral normalization. *IEEE Trans. Circuits Syst. Video Technol.* (2023)
102. X. Zang, G. Li, W. Gao, Multidirection and multiscale pyramid in transformer for video-based pedestrian retrieval. *IEEE Trans. Ind. Inform.* **18**(12), 8776–8785 (2022)
103. X. Zang, G. Li, W. Gao, X. Shu, Learning to disentangle scenes for person re-identification. *Image Vis. Comput.* **116**, 104330 (2021)
104. X. Zang, G. Li, W. Gao, X. Shu, Exploiting robust unsupervised video person re-identification. *IET Image Process.* **16**(3), 729–741 (2022)

105. Z. Yue, G. Li, W. Gao, Cross-level guided attention for human-object interaction detection, in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2023), pp. 284–289
106. S. Sun, J. Liu, H. Li, G. Liu, T.H. Li, W. Gao, Streamflow: streamlined multi-frame optical flow estimation for video sequences, in *Advances in Neural Information Processing Systems* (2024)
107. Z. Yao, W. Gao, Iterative saliency aggregation and assignment network for efficient salient object detection in optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* (2024)
108. G. Liao, W. Gao, Rethinking feature mining for light field salient object detection. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(10), 1–24 (2024)
109. B. Yang, W. Gao, G. Li, Focus and adjust: progressive refinement network for human object interaction detection, in *2022 26th International Conference on Pattern Recognition (ICPR)* (IEEE, 2022), pp. 2546–2552
110. G. Liao, W. Gao, Q. Jiang, R. Wang, G. Li, Mmnet: multi-stage and multi-scale fusion network for rgb-d salient object detection, in *Proceedings of the 28th ACM International Conference on Multimedia* (2020), pp. 2436–2444
111. W. Gao, S. Kwong, Y. Zhou, Y. Jia, J. Zhang, W. Wu, Multiscale phase congruency analysis for image edge visual saliency detection, in *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 1 (IEEE, 2016), pp. 75–80
112. W. Gao, S. Kwong, Y. Zhou, X. Wang, Phase congruency analysis of down-sampled and blurring images for foreground extraction, in *2015 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)* (IEEE, 2015), pp. 152–157
113. Y. Guo, W. Gao, G. Li, Interpretable task-inspired adaptive filter pruning for neural networks under multiple constraints. *Int. J. Comput. Vis.* 1–17 (2024)
114. W. Gao, Y. Guo, S. Ma, G. Li, S. Kwong, Efficient neural network compression inspired by compressive sensing. *IEEE Trans. Neural Netw. Learn. Syst.* (2022)
115. Y. Guo, W. Gao, Semantic-driven automatic filter pruning for neural networks, in *IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
116. L. Tao, W. Gao, Efficient channel pruning based on architecture alignment and probability model bypassing, in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2021), pp. 3232–3237
117. S. Fan, W. Gao, G. Li, Salient object detection for point clouds, in *European conference on computer vision* (Springer, 2022), pp. 1–19
118. S. Luo, W. Gao, A general framework for rotation invariant point cloud analysis, in *ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2024), pp. 3665–3669
119. X. Lu, W. Gao, Attentivenet: detecting small objects for lidar point clouds by attending to important points, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
120. Z. Pan, N. Zhang, W. Gao, S. Liu, G. Li, Less is more: label recommendation for weakly supervised point cloud semantic segmentation. *Proc. AAAI Conf. Artif. Intell.* **38**(5), 4397–4405 (2024)
121. Z. Pan, G. Liu, W. Gao, T.H. Li, Epcontrast: effective point-level contrastive learning for large-scale point cloud understanding, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024), pp. 1–6
122. N. Zhang, Z. Pan, T.H. Li, W. Gao, G. Li, Improving graph representation for point cloud segmentation via attentive filtering, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 1244–1254
123. K. Wen, N. Zhang, G. Li, W. Gao, Mpvnn: multi-resolution point-voxel non-parametric network for 3d point cloud processing, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024), pp. 1–6
124. Z. Pan, N. Zhang, W. Gao, S. Liu, G. Li, Point cloud semantic segmentation with sparse and inhomogeneous annotations, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)

125. R. Zhang, G. Li, W. Gao, S. Liu, A quantum-inspired framework in master-servant mode for large-scale multi-modal place recognition. *IEEE Trans. Intell. Transp. Syst.* (2024)
126. D. Yang, W. Gao, Pointchd: a point cloud benchmark for congenital heart disease classification and segmentation. *IEEE J. Biomed. Health Inform.* (2024)
127. S. Fan, W. Gao, G. Li, Point-mpp: point cloud self-supervised learning from masked position prediction. *IEEE Trans. Neural Netw. Learn. Syst.* (2024)
128. Z. Pan, W. Gao, G. Li, Distribution guidance network for weakly supervised point cloud semantic segmentation, in *Advances in Neural Information Processing Systems* (2024)
129. R. Zhang, G. Li, W. Gao, T.H. Li, Compoint: can complex-valued representation benefit point cloud place recognition? *IEEE Trans. Intell. Transp. Syst.* (2024)
130. W. Liu, W. Gao, X. Mu, Fast inter-frame motion prediction for compressed dynamic point cloud attribute enhancement. *Proc. AAAI Conf. Artif. Intell.* **38**(4), 3720–3728 (2024)
131. X. Fan, G. Li, D. Li, Y. Ren, W. Gao, T.H. Li, Deep geometry post-processing for decompressed point clouds, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
132. X. Zhang, G. Liao, W. Gao, G. Li, Tdrnet: transformer-based dual-branch restoration network for geometry based point cloud compression artifacts, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
133. Z. Li, G. Li, T.H. Li, S. Liu, W. Gao, Semantic point cloud upsampling. *IEEE Trans. Multimedia* (2022)
134. Z. Li, S. Liu, W. Gao, G. Li, G. Li, S4r: rethinking point cloud sampling via guiding upsampling-aware perception. *IEEE Trans. Multimedia* (2025)
135. R. Zhang, W. Gao, G. Li, T.H. Li, Qinet: decision surface learning and adversarial enhancement for quasi-immune completion of diverse corrupted point clouds. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–14 (2022)
136. R. Bao, Y. Ren, G. Li, W. Gao, S. Liu, Flow-based point cloud completion network with adversarial refinement, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2559–2563
137. J. Chen, G. Li, R. Zhang, T.H. Li, W. Gao, Pointvae: invertible variational autoencoder framework for 3d point cloud generation, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 3216–3220
138. R. Zhang, J. Chen, W. Gao, G. Li, T.H. Li, Pointot: interpretable geometry-inspired point cloud generative model via optimal transport. *IEEE Trans. Circuits Syst. Video Technol.* **32**(10), 6792–6806 (2022)
139. Z. Li, G. Li, T. Li, S. Liu, W. Gao, Information-growth attention network for image super-resolution, in *Proceedings of the 29th ACM International Conference on Multimedia* (2021), pp. 544–552
140. L. Zhou, W. Gao, G. Li, H. Yuan, T. Zhao, G. Yue, Disentangled feature distillation for light field super-resolution with degradations, in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2023), pp. 116–121
141. L. Zhou, W. Gao, G. Li, End-to-end spatial-angular light field super-resolution using parallax structure preservation strategy, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 3396–3400
142. W. Gao, L. Zhou, L. Tao, A fast view synthesis implementation method for light field applications. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* **17**(4), 1–20 (2021)
143. X. Zhang, W. Gao, G. Li, Q. Jiang, R. Cong, Image quality assessment-driven reinforcement learning for mixed distorted image restoration. *ACM Trans. Multimedia Comput. Commun. Appl.* **19**(1s), 1–23 (2023)
144. X. Zhang, W. Gao, H. Yuan, G. Li, Je 2 net: joint exploitation and exploration in reinforcement learning based image restoration, in *IEEE International Conference on Acoustics, Speech and Signal Processing* (2022), pp. 2090–2094
145. X. Zhang, W. Gao, Hirl: hybrid image restoration based on hierarchical deep reinforcement learning via two-step analysis, in *IEEE International Conference on Acoustics, Speech and Signal Processing* (2022), pp. 2445–2449

146. S. Wang, Y. Lu, W. Xia, P. Xia, Z. Wang, W. Gao, Light field angular super-resolution by view-specific queries. *Visual Comput.* 1–16 (2024)
147. Y. Sun, Z. Li, S. Wang, W. Gao, Depth-assisted calibration on learning-based factorization for a compressive light field display. *Opt. Express* **31**(4), 5399–5413 (2023)
148. Y. Sun, Z. Li, L. Li, S. Wang, W. Gao, Optimization of compressive light field display in dual-guided learning, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2075–2079
149. W. Gao, S. Fan, G. Li, W. Lin, A thorough benchmark and a new model for light field saliency detection. *IEEE Trans. Pattern Anal. Mach. Intell.* (2023)
150. W. Gao, G. Li, *Deep Learning for 3D Point Clouds* (Springer Nature, 2024)
151. W. Gao, G. Li, Introduction to 3d point clouds: Datasets and perception, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 1–27
152. W. Gao, G. Li, Learning basics for 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 29–70
153. W. Gao, G. Li, Deep-learning-based point cloud enhancement i, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 71–97
154. W. Gao, G. Li, Deep-learning-based point cloud enhancement ii, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 99–130
155. W. Gao, G. Li, Deep-learning-based point cloud analysis i, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 131–162
156. W. Gao, G. Li, Deep-learning-based point cloud analysis ii, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 163–193
157. W. Gao, G. Li, Point cloud pre-trained models and large models, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 195–225
158. W. Gao, G. Li, Point cloud-language multi-modal learning, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 227–254
159. W. Gao, G. Li, Open-source projects for 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 255–272
160. W. Gao, G. Li, Typical engineering applications of 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2025), pp. 273–299
161. W. Gao, G. Li, Future work on deep learning-based point cloud technologies, in *Deep Learning for 3D Point Clouds* (Springer, 2025), pp. 301–315
162. C. Dong, Y. Deng, C.C. Loy, X. Tang, Compression artifacts reduction by a deep convolutional network (2015). [arXiv:1504.06993](https://arxiv.org/abs/1504.06993)
163. K. Zhang, W. Zuo, Y. Chen, D. Meng, L. Zhang, Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising (2016). [arXiv:1608.03981](https://arxiv.org/abs/1608.03981)
164. T. Xue, B. Chen, J. Wu, D. Wei, W.T. Freeman, Video enhancement with task-oriented flow (2017). [arXiv:1711.09078](https://arxiv.org/abs/1711.09078)
165. A. Ranjan, M.J. Black, Optical flow estimation using a spatial pyramid network, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 4161–4170
166. L. Xie, M. Hu, X. Bai, Towards hardware-friendly and robust facial landmark detection method, in *International Conference on Neural Information Processing* (Springer, 2022), pp. 432–444
167. L. Xie, M. Hu, X. Bai, online improved vehicle tracking accuracy via unsupervised route generation, in *IEEE 34th International Conference on Tools with Artificial Intelligence* (IEEE, 2022), pp. 788–792
168. Q. Xing, Z. Guan, M. Xu, R. Yang, T. Liu, Z. Wang, Mfqc 2.0: a new approach for multi-frame quality enhancement on compressed video (2019). [arXiv:1902.09707](https://arxiv.org/abs/1902.09707)
169. Y. Xu, L. Gao, K. Tian, S. Zhou, H. Sun, Non-local convlstm for video compression artifact reduction (2019). [arXiv:1910.12286](https://arxiv.org/abs/1910.12286)
170. X. Wang, K. C. Chan, K. Yu, C. Dong, C.C. Loy, Edvr: video restoration with enhanced deformable convolutional networks (2019). [arXiv:1905.02716](https://arxiv.org/abs/1905.02716)

171. J. Deng, L. Wang, S. Pu, C. Zhuo, Spatio-temporal deformable convolution for compressed video quality enhancement. *Proc. AAAI Conf. Artif. Intelli.* **34**(07), 10 696–10 703 (2020)
172. Q. Ding, L. Shen, L. Yu, H. Yang, M. Xu, Patch-wise spatial-temporal quality enhancement for hevc compressed video. *IEEE Trans. Image Process.* **30**, 6459–6472 (2021)
173. R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 586–595
174. H. Zheng, X. Li, F. Liu, L. Jiang, Q. Zhang, F. Li, Q. Dang, D. He, Adaptive spatial-temporal fusion of multi-objective networks for compressed video perceptual enhancement, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 268–275
175. S. Zhou, J. Zhang, J. Pan, H. Xie, W. Zuo, J. Ren, Spatio-temporal filter adaptive network for video deblurring, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 2482–2491
176. W. Shen, W. Bao, G. Zhai, L. Chen, X. Min, Z. Gao, Video frame interpolation and enhancement via pyramid recurrent framework. *IEEE Trans. Image Process.* **30**, 277–292 (2020)
177. R. Yang, F. Mentzer, L.V. Gool, R. Timofte, Learning for video compression with hierarchical quality and recurrent enhancement, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 6628–6637
178. X. Zhang, X. Wu, X. Zhai, X. Ben, C. Tu, Davd-net: deep audio-aided video decompression of talking heads, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 12 335–12 344

Chapter 9

Standards for AI-based Image and Video Coding



9.1 Introduction

As a powerful computing tool, deep learning can handle well not only the machine analysis tasks [1–34], but also the image and video data compression tasks [34–45]. AI-based image and video coding utilizes artificial intelligence (AI) technologies for compressing images and videos [35–39, 41–58]. The objective of this approach is to reduce storage requirements and transmission bandwidth while preserving the quality of the visual content [59–97]. Nevertheless, AI-based image and video coding encounters several challenges [98–106]. Firstly, the training of effective compression models necessitates significant computational resources and extensive training data. Secondly, as these models are optimized for specific tasks or datasets, their performance may not generalize well to new scenarios or datasets. Finally, since these models often operate as black-box models, interpreting and comprehending their compression decisions can be challenging. Indeed, recognizing the importance of AI-based image and video coding, several organizations have embarked on formulating standards in this domain. The primary objective of these standards is to establish uniform and interoperable compression methods that enhance the efficiency and quality of compressed image and video data [107–116, 116–137].

By defining common standards, these organizations aim to address the challenges associated with AI-based compression [138–149], such as compatibility issues and the lack of consistent performance evaluation metrics. These standards facilitate seamless integration and interoperability among different platforms and systems, enabling efficient compression techniques to be adopted across various applications and industries. These standards play a crucial role in driving the widespread adoption and deployment of AI-based image and video compression.

9.2 Overview of International Standards

Institute of Electrical and Electronics Engineers (IEEE) develops standards in many areas of electronics and electrical engineering, including communications, networking, and image processing. In the field of AI-based Image and Video Compression, the Future Video Coding Study Group (FVC-SG) developed a reference image codec, namely Neural Image Coding (NIC).

JPEG (Joint Photographic Experts Group) is an international organization known for its image compression standards. The JPEG standard developed by this group has been widely adopted for image and video compression. In recent years, the JPEG organization has shifted its focus towards AI-based Image and Video Compression. They have initiated a new project called JPEG AI, which aims to research and develop new standards for AI-based image compression.

MPEG VCM (Video Coding for Machines) is a new project recently launched by MPEG to research and develop new AI-based video compression standards. Different from traditional video compression standards targeting human viewing, VCM aims to provide a more efficient video compression solution for machine such as drones and robots. The main challenge of VCM is how to retain information useful for machine tasks while maintaining a high compression rate. To achieve this goal, VCM plans to leverage AI technology to optimize the video encoding and decoding process.

MPAI is a new international standardization organization focusing on moving picture, audio and data coding. AI-based End-to-End Video Coding (MPAI-EEV) is a standard initiated by MPAI with the aim of reducing the bit rate needed to represent video through the utilization of data coding technologies. Unlike traditional video coding methods, MPAI-EEV explores new approaches that are not limited by the conventional constraints of video coding.

9.3 IEEE AI-based Image and Video Coding Standard

9.3.1 *Introduction of IEEE AI Standards*

Recently, deep coding have been considered for image, video and 3d point cloud, which demonstrates great potential. FVC-SG has been studying AI-based image/video coding for more than five years. FVC-SG developed a reference image codec, namely Neural Image Coding (NIC). So far, the development of AI standards has been discussed in 18 meetings.

9.3.2 The Development of IEEE AI Standards

This section follow the development diagram provided above and introduce each proposal in chronological order, starting from the thirtieth meeting.

The 30th meeting: During the 30th meeting, the working group discussed and confirmed the reference software NIC as the initial milestone in the development of AI standards.

From the 31st meeting to the 37th meeting: Over the course of 7 meetings, the working group deliberated on several proposals, leading to the advancement of the reference software NIC. These proposals aimed to enhance its functionalities and address any identified limitations. Through collaborative discussions and improvements, the reference software NIC evolved to version 0.4. The development process is illustrated in Fig. 9.1.

The 38th meeting: During the thirty-eighth meeting, the working group conducted a solicitation for neural network compression systems. A total of 8 proposals were received from various participants.

From the 39th meeting to now: Starting from the thirty-ninth meeting, the focus shifted towards evaluating and optimizing the proposals received for the neural network compression systems. The selected reference software, namely NIC, BEE, and iWare, formed the basis for these optimization efforts. The goal was to merge the strengths of each proposal and create an improved and comprehensive compression system. The development process is illustrated in Fig. 9.2.

The 30th FVC-SG. After conducting subjective and objective evaluations, recurrent evaluations, discussions, and subsequent developments, the reference software NIC-0.1 was released, as shown in Fig. 9.3. This software implements the AI-based image compression method proposed by Nanjing University [150], which serves as a key standard in the field of AI image compression. The team's efforts resulted in three significant contributions, as outlined in their study [151].

Non-Local Attention Module: The team introduced a Non-Local Attention Module that incorporates global correlation into the compression process.

Attention-based Bit Allocation: The attention mechanism is employed to combine the non-local operators, enabling the generation of an importance mask. This mask is instrumental in achieving efficient bit allocation, thereby optimizing the compression process. Refer to Fig. 9.4 for a visual representation. The calculation process of module NLN is as follows:

$$\begin{aligned}
 \mathbf{Y}_i &= \frac{1}{C(\mathbf{X})} \sum_{\forall j} f(\mathbf{X}_i, \mathbf{X}_j) g(\mathbf{X}_j), \\
 f(\mathbf{X}_i, \mathbf{X}_j) &= e^{\phi(\mathbf{X}_i)^T \phi(\mathbf{X}_j)}, \\
 \mathbf{Y} &= \text{softmax}(\mathbf{X}^T W_\theta^T W_\phi \mathbf{X}) g(\mathbf{X}), \\
 \mathbf{Z} &= W_z \mathbf{Y} + \mathbf{X},
 \end{aligned} \tag{9.1}$$

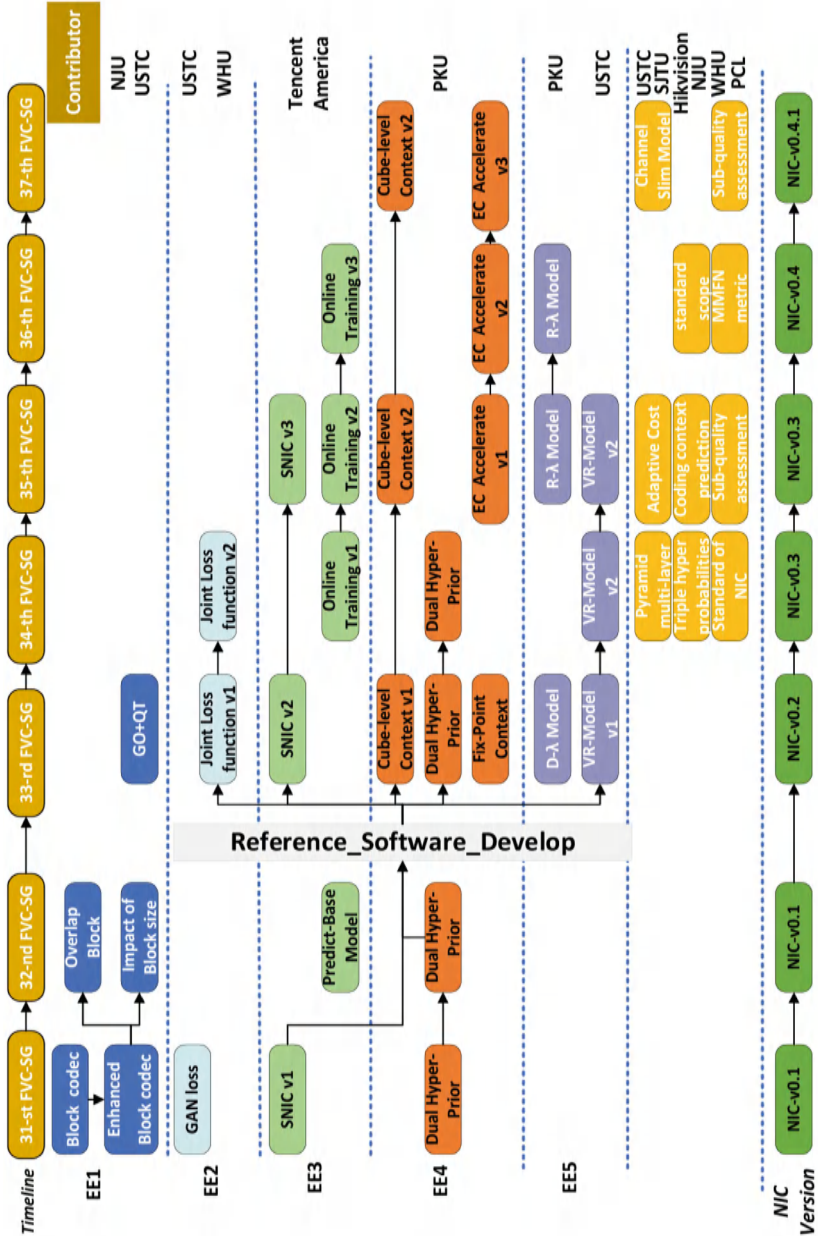


Fig. 9.1 The development process of NIC in IEEE AI standards, Source Author

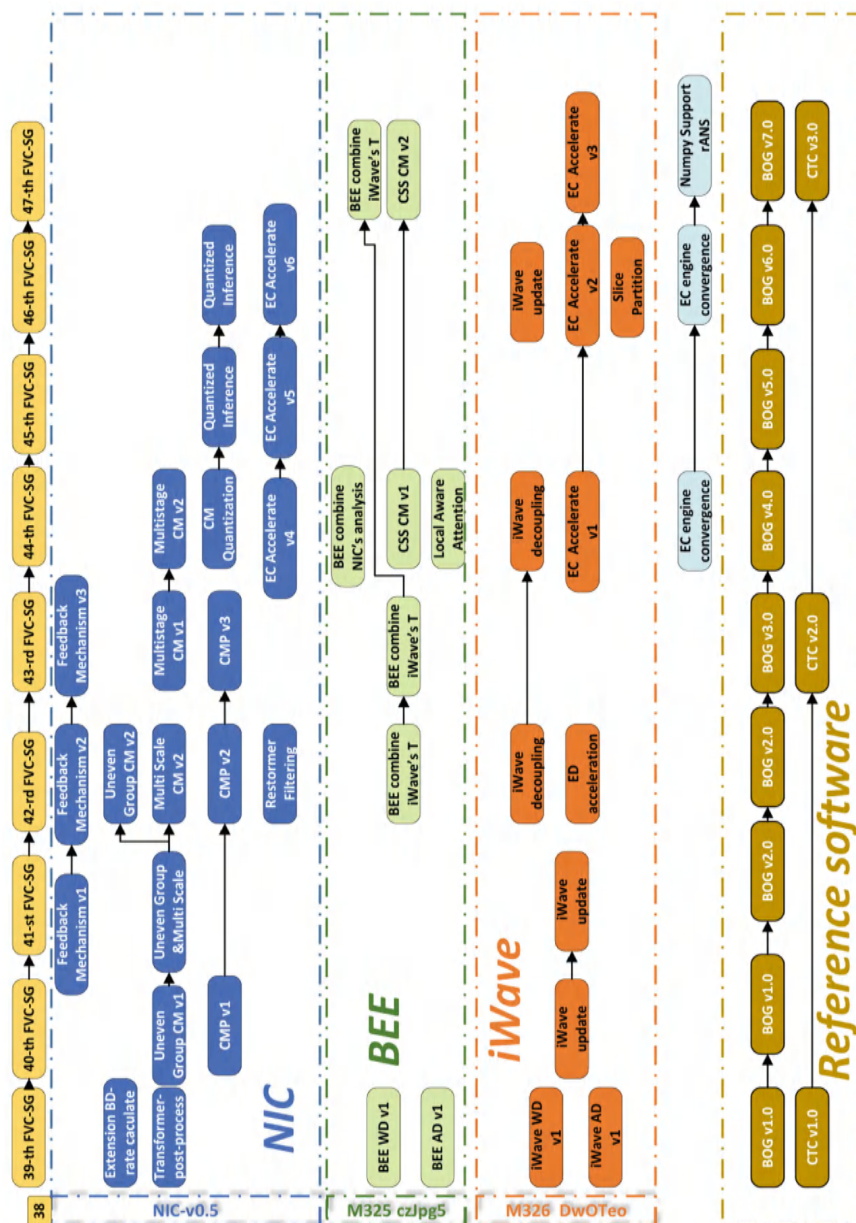


Fig. 9.2 The development process of BOG. *Source* Author

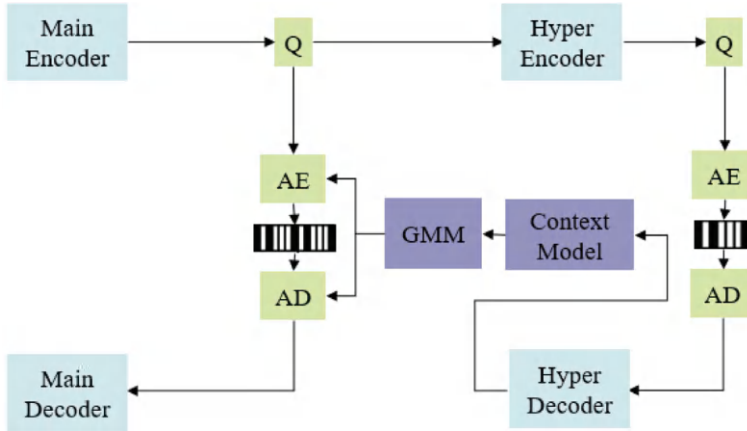


Fig. 9.3 Overall structure of NIC-0.1 network. The image shown is introduced with IEEE open access (OA) work under CC BY Licence (Copyright © 1988–2024, www.ieee.org) [150]

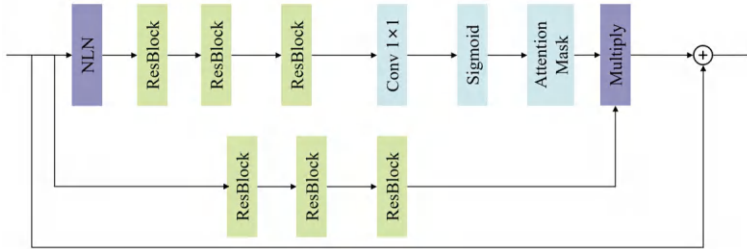


Fig. 9.4 Detailed diagram of NLAM module. The image shown is introduced with IEEE open access (OA) work under CC BY Licence (Copyright © 1988–2024, www.ieee.org) [150]

where $\mathbf{Y}_i$ is the output feature for the i th input, $C(\mathbf{X})$ is a normalization factor that ensures the outputs are properly scaled, $f(\mathbf{X}_i, \mathbf{X}_j)$ is a function that computes the similarity between the i th and j th input features based on their embeddings $\phi(\mathbf{X}_i)$ and $\phi(\mathbf{X}_j)$, which are transformed representations of the inputs, and $g(\mathbf{X}_j)$ is a function applied to the j th input feature. In the third equation, $\mathbf{Y}$ represents the attention weights computed using the softmax function applied to the quadratic form of the input features, where W_θ and W_ϕ are learnable weight matrices. Finally, $\mathbf{Z}$ is the output of the module, which is computed by combining the transformed features $\mathbf{Y}$ and the original input $\mathbf{X}$ using the weight matrix W_z .

Masked 3D CNN: It uses a single-layer masked 3D CNN for digging exploit the spatial-channel correlations, in order to improve the entropy coding.

The 31st FVC-SG. In the meeting, five proposals proposed improvements to NIC-0.1, as shown in the Table 9.1.

M265 [152] divided the image into 128×128 blocks and performs block-level parallel encoding and decoding. Subsequently, the block boundaries undergo post-

Table 9.1 The proposal in 31st FVC-SG

Proposal	Description	Organization
M265	Block-based neural codec	NJU
M266	Block-based neural codec with block enhancement	NJU
M267	Substitutional neural image compression	Tencent
M268	Perceptual quality oriented end-to-end compression	USTC
M269	Fast decoding end-to-end learned image coding framework for NIC-0.1	PKU

Source Author

processing. Building upon the advancements of M265, M266 [153] introduced additional post-processing techniques for the block center area. This addresses the issue of boundary discontinuity. However, it does not fully resolve the problem of boundary color difference.

M267 [154] presented a Substitutional Neural Image Compression (SNIC) for end-to-end image compression. The key idea is that, for each input image to be compressed, it find an optimal substitute and compress this substitute instead of the original image, and this substitute can achieve better compression performance in a way defined by the user. However, this improvement at the cost of increased complexity in the encoding process.

M268 [155] introduced a novel training method that enables existing image compression networks to reconstruct images with enhanced perceptual quality while achieving improved bit allocation. The key concept behind this approach is to design a new loss function using Generative Adversarial Networks (GANs) and fine-tune the network based on this trained model. By implementing this scheme, the subjective quality of the reconstructed images is noticeably enhanced.

M269 [156] introduced a significant improvement in decoding speed by removing the context model and implementing a multi-layer super-prior structure. This advancement results in a remarkable 30-fold increase in the decoding speed compared to previous methods. The dual-layer hyper-prior model is shown in Fig. 9.5.

The 32nd FVC-SG. In the meeting, four proposals proposed improvements to NIC-0.1, as shown in the Table 9.2.

M272 reports on the development of NIC reference software and establishes the following EE.

EE1—Block Coding. The block prediction technique combines residual coding with adaptive block size. This approach enhances the compression efficiency by efficiently representing image data in blocks.

EE2—Loss Function and Subjective Quality Optimization. The test dataset is divided into two parts, with 16 images having a resolution lower than 1080p and 8 images exceeding 1080p. The subjective testing process aligns with the NIC CfE (Common framework for experimentation). This EE aims to optimize the loss function and subjective quality, ensuring that the compressed images maintain high visual fidelity.

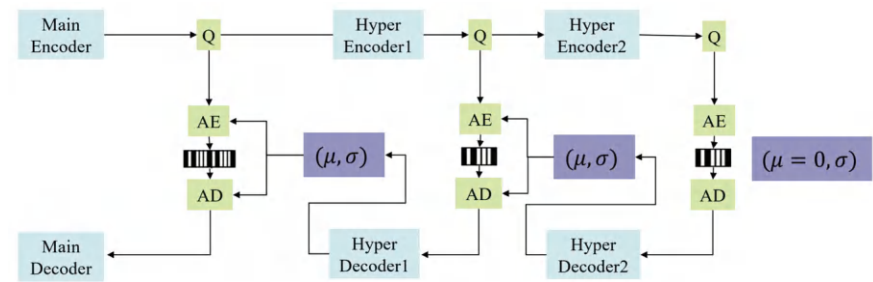


Fig. 9.5 Network structure of dual-hyper model. The image shown is introduced with IEEE open access (OA) work under CC BY Licence (Copyright © 1988–2024, www.ieee.org) [156]

Table 9.2 The proposal in 32nd FVC-SG

Proposal	Description	Organization
M272	NIC software development report	Software coordinator
M273	Codec using overlapping blocks	NJU
M274	The impact of block size on codec performance	NJU
M275	Efficient coarse-to-fine hyperprior model for NIC	PKU
M276	A block-wise prediction-based method with post filtering	Tencent America

Source Author

EE3—Encoding Preprocessing. Encoding preprocessing involves the preprocessing of input images. Post-processing methods are suggested to be studied within each EE independently. These methods can include block-based or frame-based processing techniques.

EE4—Complexity Optimization. This EE focuses on optimizing the context model and implementing a coarse-grained context model. The decoder part combines the information of hyper transform to reduce the performance loss caused by complexity optimization.

EE5—Variable Code Rate Model. The objective is to utilize a minimal number of models to cover a wide range of code rates while maintaining coding performance. Alternatively, establishing a corresponding relationship between the parameter λ and the target code rate R can be explored.

In M273 [157], a solution was proposed to address the color difference problem in M266 during block coding. The approach involves including the upper-left pixel in the block coding process. After that, entropy coding is performed on the feature map crop to further mitigate the color difference issue. In M274 [158], various block sizes were experimented with for block encoding. Among these different block sizes, it was observed that a block size of 512 exhibited the best performance in terms of

compression efficiency. However, it is important to note that utilizing a block size of 512 also required the highest amount of computational resources and memory.

In continuation of the research based on M269, M275 [159] further explored improvements in encoding performance. While there is a slight enhancement in encoding performance for Class D, it is important to note that Class B experiences a loss of nearly 10%.

The 33-rd FVC-SG. Based on the current integrated technologies, the meeting concluded to remove the QT block division and proceeded with the release of NIC version 0.2. Additionally, the decision was made to adopt and integrate the M285 technology into NIC.

In terms of PSNR, the high bit rate performance of NIC-0.2 (MSE) surpasses that of VVC, while at low bit rates, NIC-0.2 slightly lags behind VVC. However, when considering the MS-SSIM metric, NIC-0.2 demonstrates a significant improvement over VVC.

In the meeting, eight proposals proposed improvements to NIC, as shown in the Table 9.3.

M279 [160] is a continuation of M267. Using this method requires two parameters: learning rate and number of iterations. M280 [161] introduced variable bitrate model adopted into NIC. M281 [162] proposed a hybrid loss of L1 and MS-SSIM. The loss function formulation as follows:

$$L_{\text{distortion}} = w \cdot L_2 + (1 - w) \cdot L_{\text{MS-SSIM}}, \quad (9.2)$$

where w is used to adjust the weights of the two losses. M282 [163] proposed an encoding optimization method using geometric operation. However, there is almost no improvement in QT operation, and therefore it is recommended to remove it. M283 [164] proposed a D- λ model for neural image compression (NIC) to control the target distortion based on the proposed Graded exponential model (GEM).

Table 9.3 The proposal in 33rd FVC-SG

Proposal	Description	Organization
M279	Substitutional neural image compression	Tencent America
M280	Algorithm description of variable-rate NIC model	USTC
M281	Subjective quality assessment test results of EE2 and new NIC baseline	WHU
M282	Using Geometry Operations	USTC
M283	D- λ model for distortion control	PKU
M284	Cube-level context model For NIC acceleration	PKU
M285	Dual hyper-prior model for NIC	PKU
M286	Fix-point context model	PKU

Source Author

Table 9.4 The proposal in 34th FVC-SG

Proposal	Description	Organization
M290	Algorithm description of NIC-0.2 using 3 variable-rate NIC Models	USTC
M291	Pyramid multi-layer structure	SJTU
M292	Triple hyper probabilities	SJTU
M293	Online training-based post filtering	Tencent
M294	Exploration of joint loss function	WHU
M295	Dual hyper-prior model	PKU
M296	Towards the standardization of NIC	NJU and Hikvision

Source Author

M284 [165] focused on developing a cube-level context model to accelerate the NIC process. The team aims to study the trade-off between acceleration ratio and performance loss for M284. M285 [166] is a continuation of M275 research. Its encoding and decoding time is almost unchanged and the performance gain is about 4%.

The 34th FVC-SG. In the meeting, seven proposals proposed improvements to NIC, as shown in the Table 9.4.

M290 enhances NIC-0.2 by introducing a model specifically designed for low bit rates. M291 [167] modified the encoding and decoding network structure and improves the compression performance by 1%. M292 [168] modified the hyper encoding and decoding structure. The original N channels are split into two groups of 3 and N-3 channels. M293 [169] introduced a simple post-processing filter (a layer of 3×3). The filter parameters are learned and passed online (16bit floating point number) at the 512×512 (1024×1024) level encoder, and RDO decides whether to turn on the filter. Based on M281, M294 [170] further proposed the following loss function 9.3:

$$L_{\text{distortion}} = w \cdot L_{\text{MSE}} + (1 - w) \cdot \alpha \cdot L_{\text{MS-SSIM}}. \quad (9.3)$$

Similar to the previous proposal, this equation combines the MSE loss, L_{MSE} , and the MS-SSIM loss, $L_{\text{MS-SSIM}}$. M295 [171] is a further development of M285.

35th FVC-SG. In the meeting, eight proposals proposed improvements to NIC, as shown in the Table 9.5. Two proposals, M301 and M304, are adopted for further development and integration into the reference software.

M299 [172], proposed by PCL&HIT-SZ, describes an algorithm for image compression with self-adaptive computational cost. The proposal leverages the strong correlation between local channel convolution and all channel convolution to achieve acceleration. The algorithm employs a learnable gate to obtain a decision map. This decision map is then used to trim most channels, resulting in computational savings. The final compressed image is obtained by adding the convolution sum of the remaining channel parts. The loss function associated with this proposal is defined as:

Table 9.5 The proposal in 35th FVC-SG

Proposal	Description	Organization
M299	Self-adaptive computational cost	PCL& HIT-SZ
M300	An entropy coding accelerating algorithm	PKU
M301	Time complexity analysis of preprocessing	Tencent America
M302	Online training-based post filtering	Tencent America
M303	Subjective quality assessment of NIC-0.2	WHU & USTC
M304	Coding context prediction using hyper prior	SJTU
M305	R-lambda model based rate control method	USTC
M306	Cube-level context model	PKU

Source Author

$$L = R + \lambda D + \gamma \sum_{i=1}^n (th_i - gth), \quad (9.4)$$

where L is the total loss, R is the rate of the compressed image, D is the distortion loss, λ balances rate and distortion, γ is a regularization factor, n is the number of channels, th_i is the output for the i th channel, and gth is the target output for the image. The summation penalizes deviations from the target output.

M300 [173] introduced a channel-adaptive codebook construction method based on the training data. This codebook allows for the determination of the all-zero channels within a macro block. This information is then utilized to skip the entropy encoding process for those specific channels within the macro block. M301 [174] is a further study of M279. This proposal reduces the number of pre-processing steps and is adopted into the reference software. M302 [175] is an improvement of M293. When online training is optimized for MS-SSIM, MSE is first used as the loss function, and then replaced with MS-SSIM. M304 [176] introduced the use of a hyper prior to predict the main features in coding context. Subsequently, quantization and entropy coding are applied to the prediction residual, as shown in Fig. 9.6. M306 [177] improved M284, changing the $4 \times 4 \times 4$ cube into $2 \times 4 \times 4$.

After the meeting, NIC version 0.4 is released. A comparison with the previous version reveals that the encoding time has been accelerated by approximately 12 times, while the decoding time remains largely unchanged. However, there is a slight loss in compression performance, amounting to around 0.5%.

The 36th FVC-SG. In the meeting, eight proposals proposed improvements to NIC, as shown in the Table 9.6. Additionally, two proposals, M311 and M316, were discussed, and their outcomes were considered for integration into NIC version 0.4.

In order to pursue subjective consistency, M309 [178] provided an evaluation metric, MMFN, for learning-based image coding. For each network model, M310 [179] determined that the priors for certain channels are all 0 based on 800 training images. These channels are skipped directly during encoding and decoding. M311 [180] is an improvement of M302. It trains and transmits post-processing network

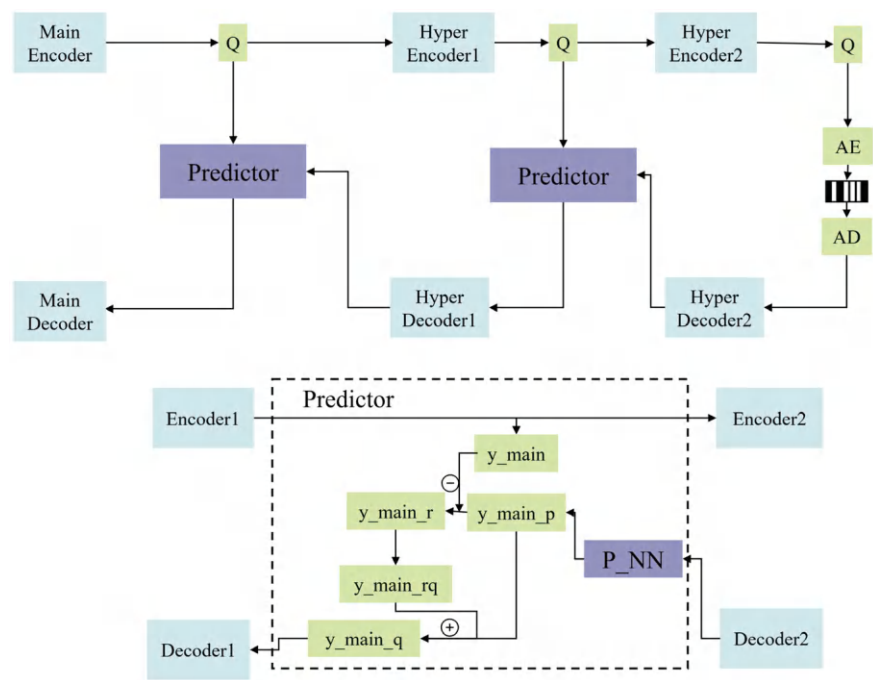


Fig. 9.6 Places of predictors in NIC model and the structure of the predictor. The image shown is introduced with IEEE open access (OA) work under CC BY Licence (Copyright © 1988–2024, www.ieee.org) [176]

Table 9.6 The proposal in 36th FVC-SG

Proposal	Description	Organization
M309	MMFN metric for FVC	USTC
M310	An improved entropy coding algorithm	PKU
M311	Online training-based post enhancement	Tencent America
M312	The standardization scope	USTC
M313	VR-models for NIC-0.3 and NIC-0.4	USTC
M314	Rate control for NIC-0.4 through R- λ model	USTC
M315	Modify the configuration of the anchor NIC	USTC
M316	Randomness in online-training methods	Tencent America

Source Author

parameters in units of 2048×2048 image slices. M312 [181] recommended that only the entropy decoding part can be normalized in the upcoming standard. M313 [182] has been performance verified on NIC-0.3 and 0.4 versions respectively, and the performance has been improved in most cases. M314 [183] provided a code rate- λ model and provides block-level λ writing for code rate optimization. In order to

Table 9.7 The proposal in 37th FVC-SG

Proposal	Description	Organization
M319	Channel slimming of main codec for real-time inference acceleration	PCL
M320	An improved entropy coding algorithm	PKU
M321	Draft of the subjective quality assessment test	WHU
M322	Cube-level context model	PKU

Source Author

reduce the fluctuation of each experimental result, M316 [184] proposed to add the setting seed = 0 to the NIC CTC.

The 37th FVC-SG. In the meeting, NIC-0.4.1 version has been released with almost no change in performance compared to NIC-0.4 and four proposals proposed improvements to NIC, as shown in the Table 9.7. Two additional proposals, M320 and M321, were discussed and their outcomes were considered for integration into NIC.

M319 [185], proposed by PCL, focuses on channel slimming of the main codec for real-time inference acceleration. This proposal introduces a binary mask to the coding/decoding network, allowing for the removal of certain channels. By retraining the MSE-200 model using this approach, the encoding and decoding time can be reduced to 60–70% of the original, and the code rate can be increased by more than 5%. While there is a slight degradation in PSNR by 0.12dB, the RD (rate-distortion) performance is significantly improved.

M320 [186] is a further study of M310. The difference is that this solution does not use a prior all-zero information. M321 suggests a subjective quality assessment scheme for evaluating the perceived quality of NIC-encoded images. M322 is a continuation of M306 research.

The 38th FVC-SG. The IEEE 1857.11 Call for Proposals (CfP) started on December 21, 2021, and 7 valid proposals (M325 to M331) were received by May 16, 2022, including 12 technical solutions. The following highlights two proposals that were adopted as reference software.

M325 [187] presented an end-to-end neural network-based image coding solution known as BEE platform in response to IEEE 1857.11 Call for Proposals. There are two variants of the proposed platform: one is for objective quality and the other is for subjective quality. As a response to the CfP, it propose an end-to-end neural network-based image coding solution known as Bytedance end-to-end (BEE) platform. Two variants of the proposed platform, one for objective quality and the other for subjective quality, are presented and evaluated. The architecture of the BEE platform is illustrated in Fig. 9.7. First, bitstream parsing is decoupled from the sample prediction. Second, the proposed contribution features parallel prediction of $\hat{y}$. Third, to further improve the performance without retraining, a series of additional operations can optionally be applied in the latent domain, such as the quantized residual $\hat{w}$. Another important improvement proposed in the contribution is that the

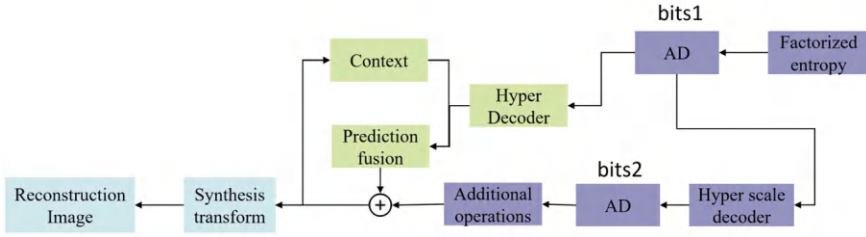


Fig. 9.7 The architecture of the BEE platform. The image shown is introduced with IEEE open access (OA) work under CC BY Licence (Copyright © 1988–2024, www.ieee.org) [187]

prediction of the quantized latent sample is not processed sample-by-sample, but column-by-column. The fully parallel prediction allows full utilization of the GPU, as the samples of $\hat{y}$ can be processed in parallel. The usage of decoupled parsing and prediction together with the parallel processing accelerates the decoding process while retaining the coding gain provided by the context model-based autoregressive network. In the latent domain, some of the latent representations, such as σ , can be processed by additional operations. The typical procedure typically includes selection of the samples, followed by a scaling process. An additional benefit achieved by latent domain additional operations is adaptive quantization.

One key feature of proposal M326 [188] is that the used iWave transform is invertible, thus their method could support both lossy compression and lossless compression. In the proposal, they provide three different solutions for lossy compression, namely MSE-driven lossy solution, MSE-driven plus online training lossy solution, and perception-driven lossy solution. The structure of the iWave is illustrated in Fig. 9.8.

Generally speaking, a 2D image X is transformed into four sub-bands after one iWave vertical transformation and two iWave horizontal transformations. The four obtained subbands are named LL, HL, LH and HH. LL can be further transformed to form a pyramid structure. In the proposal, they use four levels of iWave transformation. As shown in Fig. 9.9, two kinds of iWave transforms are used in this proposal. Among the used 13 models, 10 low bit rate models use affine iWave, and 3 high bit

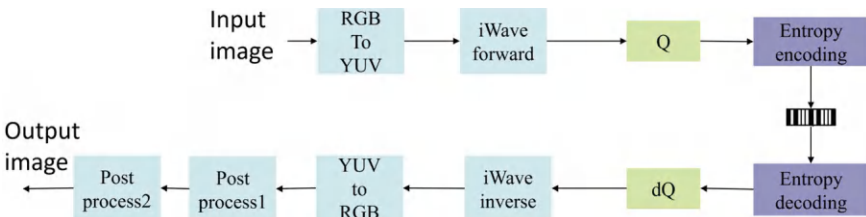


Fig. 9.8 Overall structure of the iWare. The image shown is introduced with IEEE open access (OA) work under CC BY Licence (Copyright © 1988–2024, www.ieee.org) [188]



Fig. 9.9 One level of 2D iWave transform. The image shown is introduced with IEEE open access (OA) work under CC BY Licence (Copyright © 1988–2024, www.ieee.org) [188]

rate models use additive iWave. To compress the wavelet coefficients into a bitstream, the authors use a context model, a probability model, and an arithmetic codec. The probability model employed is a Gaussian Mixture Model (GMM) composed of three Gaussians. In this proposal, for each wavelet coefficient to be encoded, the context model generates the entropy parameters of the probability model, specifically the means, standard deviations, and weights for the three Gaussians. Each subband is processed sequentially, and within a subband, each wavelet coefficient is processed one by one, from top to bottom and left to right. For lossless solution, both the post-process-1 and 2 are not used. To achieve lossless reconstruction, the color space change method is different from that used in lossy compression. In lossless solution, the transform module is directly implemented with CDF 5/3 wavelet.

On the CFP test dataset, compared to BPG, the MSE-driven lossy solution achieves 33.54% BD-rate reduction when measured with RGB-PSNR. The MSE-driven plus online training lossy solution achieves 36.80% BD-rate reduction when measured with RGB-PSNR. The lossless solution achieves 39.72% bit rate saving than BPG.

The 39th FVC-SG. In the 38th FVC-SG, it is decided to use NIC, BEE, and iWare as reference software at the same time, and try to merge them. Figure 9.10 illustrates the proposed flowchart of the decoder in the reference software that called BOG (version 1.0). At the same time, there are two proposals at this meeting to improve the reference software, as shown in Table 9.8.

BD-Rate is an objective metric used in video/image compression for performance evaluation. Existing BD-Rate function can handle most cases, but still fail to provide reasonable results for some extreme cases when the PSNR ranges of the two curves no longer overlap. M354 [190] proposed to extrapolate the curve to calculate the BD-rate for these cases. Furthermore, they suggest to use BD-PSNR as another reference metric when the PSNR gap is too large.

M355 [191] proposed a post-processing module for NIC (Neural Image Codec). The proposed model is based on a transformer and is trained offline. Meanwhile, a simple block-level RDO scheme is applied when encoding an image, and the flag information is written into the bitstream.

The 40th FVC-SG. In the meeting, four proposals proposed improvements to NIC, as shown in the Table 9.9.

The Hyper Encoder and Decoder network in NIC splits the feature map along the channel direction, thus reducing the number of parameters in the model. In NIC, the channels are evenly grouped, and each group contains the same number of channels.

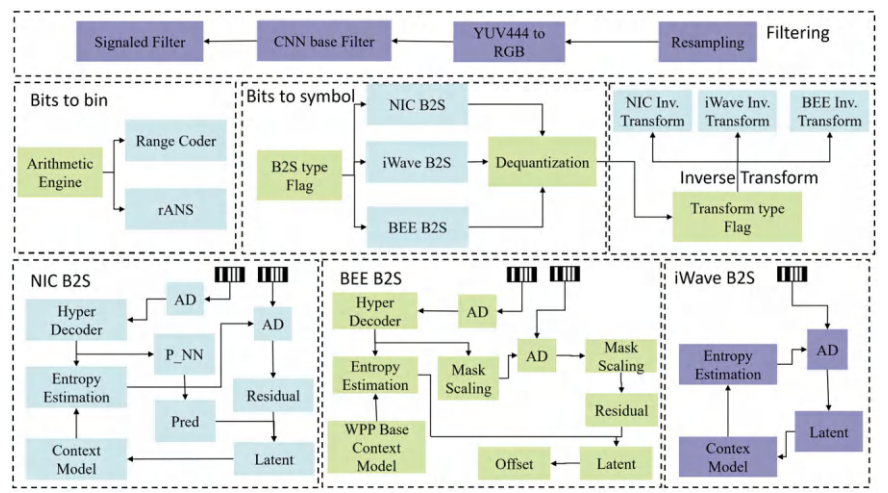


Fig. 9.10 The suggested flowchart of the decoder. The image shown is introduced with IEEE open access (OA) work under CC BY Licence (Copyright © 1988–2024, www.ieee.org) [189]

Table 9.8 The proposal in the 39th FVC-SG

Proposal	Description	Organization
M354	Extension of BD-Rate calculation function for NIC evaluation	PKU
M355	Post-processing for NIC	PKU

Source Author

Table 9.9 The proposal in 40th FVC-SG

Proposal	Description	Organization
M358	Uneven group convolution for NIC	Dahua technology
M359	Add padding bits in the picture header of iWave	USTC
M360	Context model based on prediction for NIC	Dahua technology
M361	Conditional filtering or post-processing	PKU

Source Author

However, in learned image compression, energy is not evenly distributed in channels. The front channel contains more energy, while the later channel has less energy. So, M358 [192] proposed an uneven group convolution method NIC. The uneven group convolution method is shown in Fig. 9.11.

M360 [193] described an improved context model based on prediction of latent representation for the NIC reference software. The proposed context model can combine not only the processed elements, but also the unprocessed ones with output of hyper decoder for accurate conditional statistics. As shown in Fig. 9.12.

The 41st FVC-SG. In the meeting, three proposals proposed improvements to reference software, as shown in the Table 9.10.

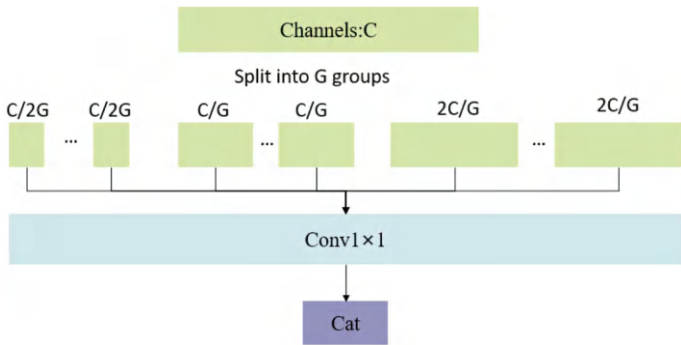


Fig. 9.11 Schematic diagram of uneven grouping method. The image shown is introduced with IEEE open access (OA) work under CC BY Licence (Copyright © 1988–2024, www.ieee.org) [192]

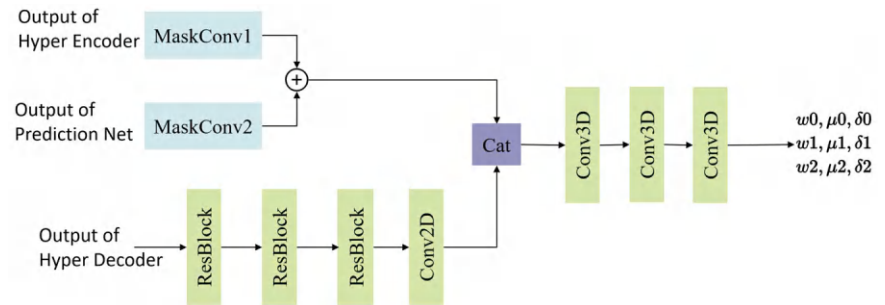


Fig. 9.12 The improved context model based on prediction and GMM. The image shown is introduced with IEEE open access (OA) work under CC BY Licence (Copyright © 1988–2024, www.ieee.org) [193]

Table 9.10 The proposal in 41st FVC-SG

Proposal	Description	Organization
M362	Add byte_alignment function into 1857.11	USTC and Bytedance
M363	Enhanced block based on feedback mechanism	Dahua technology
M364	Unevenly channel group and multi-scale mask convolution	Dahua technology

Source Author

The proposal M362 [194], presented by USTC and Bytedance, is a follow-up to M359, which focused on adding padding bits in the picture header of iWave for 1857.11.

M363 [195] described an enhanced block based on the error feedback mechanism, which can be used in any stage in encoder or decoder architecture for better performance. Specifically, features could be improved by residual information between them and their output after up-down sampling transformation. The input is pre-

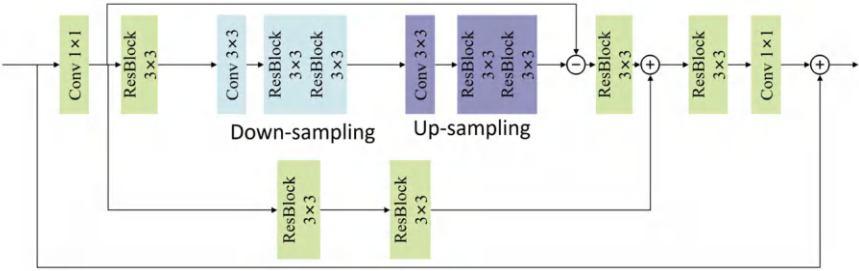


Fig. 9.13 Detailed structure diagram of the Feedback mechanism. The image shown is introduced with IEEE open access (OA) work under CC BY Licence (Copyright © 1988–2024, www.ieee.org) [195]

Table 9.11 The proposal in the 42nd FVC-SG

Proposal	Description	Organization
M367	Decoupling parsing and reconstruction process in iWave	USTC
M368	Entropy decoding acceleration for iWave	USTC
M369	Combination of iWave’s wavelet-like transformation with BEE	Bytedance
M370	Context model based on prediction for NIC	Dahua technology
M371	Unevenly channel group for context model	Dahua technology
M372	Multi-scale mask convolution for context model	Dahua technology
M373	Enhanced block based on feedback mechanism	Dahua technology
M374	Filtering technologies	PKU

Source Author

sampled first, and then up-sampled to obtain the pre-reconstructed feature with the same dimension as the input. The residual is calculated between the feature and the input. The difference information obtained is adjusted by the residual block and fed back to the input by adding. Then a jump connection is introduced from the input to form a residual block. As shown in Fig. 9.13. M364 [196] is a continuation of M358, which uses Unevenly Channel Group and Multi-scale Mask Convolution for context models.

The 42nd FVC-SG. In the meeting, eight proposals proposed improvements to reference software, as shown in the Table 9.11.

To ensure the sequential structure of different modules, the parsing and reconstruction process in iWave is decoupled in WD. However, in iWave part of reference software, entropy coding and image reconstruction are coupled. To match the reference software with WD, M367 [197] retrained iWave model following WD and CTC document. M368 [198] proposed to simplify the context model to accelerate iWave, which does pixel-level auto-regression on the smallest LL sub-band and cross sub-bands auto-regression on the other sub-bands. As shown in Fig. 9.14. The proposal

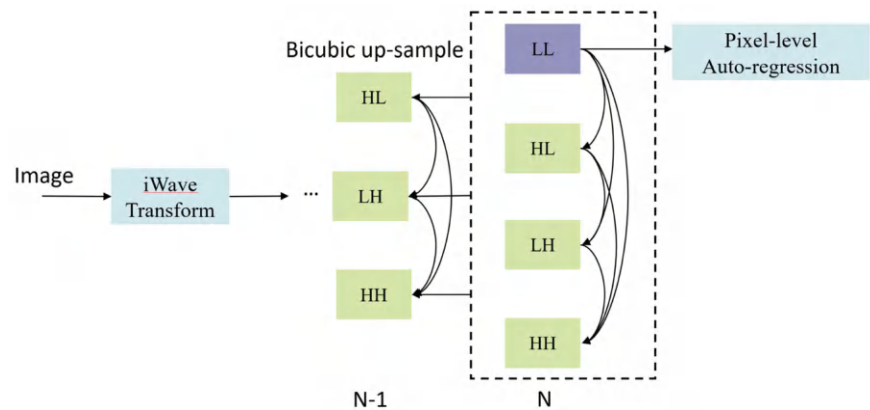


Fig. 9.14 Simplified auto-regressive entropy model in iWave. The image shown is introduced with IEEE open access (OA) work under CC BY Licence (Copyright © 1988–2024, www.ieee.org) [198]

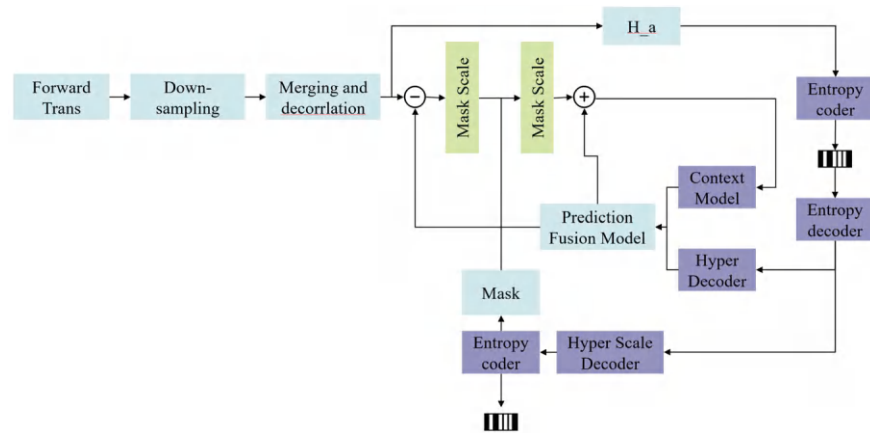


Fig. 9.15 Structure of the encoding process. The image shown is introduced with IEEE open access (OA) work under CC BY Licence (Copyright © 1988–2024, www.ieee.org) [199]

tests the model on Class C in lossless mode. The results show that it is 32.6% faster than the original iWave when decoding with 5.88% rate increasing.

M369 [199] proposed a method that combines iWave’s wavelet-like analysis and synthesis transformation with BEE’s entropy coding network. Specifically, on encoder side, the sub-band obtained by iWave’s forward transformation will be processed by designed network and fed to BEE’s entropy encoder. As shown in Fig. 9.15. And in decoder side, obtained latent feature will be resized as the input of iWave’s inverse transformation to reconstruct the image.

Several research extensions build upon previous work in video coding:

Table 9.12 The proposal in 43rd FVC-SG

Proposal	Description	Organization
M378	Multistage context model and entropy engine optimization	NJU
M379	Unevenly channel group for context model	Dahua technology
M380	Multi-scale mask convolution for context model	Dahua technology
M381	Enhanced block based on feedback mechanism	Dahua technology
M382	Time consumption of the combination model of BEE and iWave	ByteDance and USTC

Source Author

- M370 [200] extended M360 by proposing conventional convolution to replace masked convolution, though with limited effectiveness
- M371 [201] and M372 [202] both developed complexity reduction approaches based on M364
- M373 [203] advanced M363's work through strategic repositioning of inserted modules to reduce complexity
- M374 [204] introduced a Restormer-based post-processing module that demonstrates significant coding performance improvements.

The 43-rd FVC-SG. In the meeting, five proposals proposed improvements to reference software, as shown in the Table 9.12. Meanwhile, the meeting released WD v3.0, reference software v3.0, CTC v2.0.

Current NIC uses 3D autoregressive context model to predict the distribution of latent features. However, its lack of parallelism makes decoding an image take an unacceptable amount of time. So M378 [205] proposed to use multistage context model in replace of the 3D autoregressive context model. In addition, the proposal modifies the entropy engine in NIC to further improve the efficiency.

M379 [206] is a continuation of M370 and M371. The M380 combines the M370, M371 and M372, and its R-D performance has been improved. M381 [207] is a continuation of M373, a proposal that reverses the positions of downsampling and upsampling modules in the main transform. M382 [208] is a continuation of M369.

The 44th FVC-SG. In the meeting, ten proposals proposed improvements to reference software, as shown in the Table 9.13.

M385 [209] presented the solution by integrating NIC analysis and synthesis transforms into BEE, which is the preliminary step towards the complete combination. M386 [210] proposed a method to accelerate the decoding process in NIC base M378. The proposed method can improve the decoding speed by 30 times. M388 [211] transformed the floating-point operations of context model to INT8 operations based on the proposed scheme. The test results indicate that this method is capable of ensuring the cross-platform consistent decoding capability with negligible quantization loss. M389 [212] proposed to skip the entropy coding procedure of part of the zero elements to accelerate entropy coding. For the match of encoding and

Table 9.13 The proposal in 44th FVC-SG

Proposal	Description	Organization
M385	Combination of NIC analysis and synthesis transforms with BEE	Bytedance
M386	Improved multistage context model	NJU and Hikvision
M387	Simplification of main decoder	NJU and Hikvision
M388	A Context model quantization scheme for NIC	PKU
M389	An improved entropy coding algorithm for NIC	PKU
M390	Parallel context model for BEE	AttrSense
M391	Local aware attention for BEE	AttrSense
M392	Entropy codec engine convergence	USTC
M393	Decoupling parsing and reconstruction process in iWave	USTC
M394	Entropy coding acceleration for iWave	USTC

Source Author

Table 9.14 The proposal in 45th FVC-SG

Proposal	Description	Organization
M400	Quantized inference for NIC	PKU
M401	An improved entropy coding algorithm for NIC	PKU

Source Author

decoding, it proposes to utilize the Gaussian parameters, which are shared by both the encoder side and the decoder side, to predict the sparsity of the latent representation. M390 [213] proposed a parallel context model called Channel Space Shuffle Context Model (CSSC) for BEE reference software. And it helps model achieve better RD performance and accelerates the encoding/decoding procedure. M391 [214] proposed Local Aware Attention for BEE. M392 [215] compared the performance of rANS and range coder when applied to iWave and NIC respectively. M394 [216] proposed to simplify the context model to accelerate iWave, which does pixel-level auto-regression on the smallest LL sub-band and 4-step checkerboard auto-regression on the other sub-bands.

The 45th FVC-SG. In the meeting, two proposals proposed improvements to reference software, as shown in the Table 9.14. M400 [217] presented a post-training quantization scheme for all layers in the NIC model that are shared between the encoder and the decoder. It achieve this by replacing the unsafe floating-point arithmetic in the network with corresponding integer arithmetic. This approach eliminates the root cause of any potential floating-point errors that could result in disastrous decoding errors. M401 [218] is a continuation of M389.

The 46th FVC-SG. In the meeting, seven proposals proposed improvements to reference software, as shown in the Table 9.15.

Table 9.15 The proposal in 46th FVC-SG

Proposal	Description	Organization
M407	Entropy codec engine convergence	USTC
M408	Non uniform quantization between different subbands in iWave	USTC
M409	Quantized inference for NIC	PKU
M410	Sparsity-bypassing based entropy coding optimization for NIC	PKU
M411	Entropy coding acceleration for iWave on lossy mode	USTC&PKU
M412	Further entropy coding acceleration for iWave on lossless mode	USTC&PKU
M413	New design codec process with slice partition for iWave	PKU

Source Author

M407 [219] is a continuation of M392. M408 [220] recommended using different quantization scales for different subbands. M409 [221] is a continuation of M400, a proposal for cross-platform encoding/decoding with significant BD rate loss observed. M410 [222] is a continuation of M401. Expert group recommends extending the technology to BEE and iWave. M411 [223] is a continuation of M394, a proposal that made changes to the entropy model and entropy engine implementation. M412 [224] proposed to further simplify the context model based on M394 to accelerate iWave, which still performs pixel-level auto-regression on the smallest LL sub-band, while selecting multi-step checkerboard auto-regression on the other subbands based on their resolution. M413 [225] proposed a new design of iWave’s codec process with slice partition to mitigate the performance penalty of using small slices. It designs a compensating method of inter-slice information based on the properties of wavelet-like transform, and replaces the pixel range information in each slices with pixel range information of the whole image.

The 47th FVC-SG. In the meeting, five proposals proposed improvements to reference software, as shown in the Table 9.16.

The 48th FVC-SG. In the meeting, five proposals proposed improvements to reference software, as shown in the Table 9.17.

9.3.3 Performance Comparison of Latest Reference Software

As of November 2023, iWare software has achieved a significant BD-rate gain of 33.23% on the RGB PSNR indicator, outperforming the other two reference software options. The NIC reference software, optimized using MSE, achieved BD-rate gains of 25.77% and 20.47% on the RGB PSNR and YUV PSNR indicators, respectively. Furthermore, the NIC reference software, optimized with MS-SSIM, demonstrated

Table 9.16 The proposal in 47th FVC-SG

Proposal	Description	Organization
M422 [226]	Numpy support for rans	PKU
M423 [227]	Efficient parallel context model for BEE	attr sense
M424 [228]	Entropy coding acceleration for iWave on lossless mode	PKU&USTC
M427 [229]	Combination of iWave's wavelet-like transformation with BEE	ByteDance
M428 [230]	Alignment for latent representation quantization	Dahua technology

Source Author

Table 9.17 The proposal in 48th FVC-SG

Proposal	Description	Organization
M433 [231]	Lightweight optimization for iWave filtering technology	PKU
M434 [232]	Optimization of efficient parallel context model for BEE	attr sense
M435 [233]	Entropy coding acceleration for iWave on lossy mode	PKU
M436 [234]	Overlap tile for removing boundary effect	Dahua technology

Source Author

an impressive BD-rate gain of 53.96% on the RGB MS-SSIM indicator. On the other hand, the BEE reference software performed exceptionally well on the YUV MS-SSIM indicator, achieving a remarkable BD-rate gain of 48.8%. These improvements are all in comparison to the BPG standard. The reference software will be continuously updated in the future to combine the advantages of each solution [105].

9.4 JPEG AI-based Image and Video Coding Standard

This section will briefly introduce the latest JPEG AI standard. The biggest difference between the JEPG AI standard and the IEEE AI standard introduced in detail earlier is that when it converts images into YUV space for processing, the networks used for the Y and UV components are different.

9.4.1 Introduction of JPEG AI Standards

The covered range of JPEG AI is as follows: As a learning-based image coding standard, JPEG AI is required to provide a single code stream as a compact represen-

tation for the input image, which can be decoded into a human-viewable image or can be used directly as input for different visual information processing tasks without decoding specific pixel values.

JPEG AI is a standardization effort by the Joint Photographic Experts Group (JPEG) that focuses on the application of artificial intelligence (AI) techniques in visual data coding. The primary goals of JPEG AI include improving the compression efficiency, enhancing image quality, and enabling new AI-based functionalities in image-related applications.

The process in the figure can be explained as: The input image is transformed through the neural network. The neural network here is generally a convolutional network that contains multiple convolutional layers and spatial domain downsampling operations. The filters of each convolutional layer are learnable. The transformed image is quantized or simply rounded to obtain a latent representation. The code stream is obtained after the latent representation is further eliminated by entropy coding to eliminate redundancy. At the decoding end, the code stream is entropy decoded and restored to the latent representation. Hidden representations can be used for a variety of tasks. For traditional human-viewable images, the latent representation undergoes a standard reconstruction operation to obtain a standard decoded image. For image processing tasks and computer vision tasks based on neural networks and deep learning [235–246], latent representations can be directly used as input to these models, without the need to completely decode the latent representation into pixels and then input them into the model.

9.4.2 The Development of JPEG AI Standards

During the 96th meeting, JPEG selected proposals from Team14 and Team24 for integration into the standard. The integrated model resulting from this collaboration is known as the Verification Model (VM). The VM represents a consolidated effort to incorporate innovative techniques and approaches from the selected proposals.

Subsequently, in the meetings following the 96th meeting, the JPEG community has made significant progress in refining and updating the VM. The effectiveness of various methods has been rigorously evaluated through a Conformance Evaluation (CE) process. This continuous verification process has led to updates and improvements in the VM with each subsequent meeting.

As of November 2023, the VM has undergone several updates, and the latest version is 4.1. These updates represent the collective efforts of the JPEG community to push the boundaries of AI in image coding and processing, with a focus on achieving higher compression efficiency, superior image quality, and enabling novel AI-driven functionalities [247, 248].

9.4.3 Quality Metrics in JPEG AI

Only reasonably well correlated with visual quality metrics have been selected to be included into JPEG AI CTTC (Common Training and Test Conditions). The metrics used in this standard include: FSIM [249], IW-SSIM [250], MS-SSIM [251], NLPD [252], PSNR-HVS [253], VIFp [254] and VMAF [255].

9.4.4 Performance Comparison of the Latest VM

The hop model in version 4.1 has demonstrated notable improvements in various indicators compared to version 3.4. Specifically, it has achieved BD-rate gains of 40, 24, 37, 24, and 15% in the MSSSIM, NLPD, IW-SSIM, VMAF, and PSNR indicators, respectively, surpassing the performance of v3.4. However, on the VIF and FSIM indicators, it achieved slightly lower BD-rate gains of 14 and 25% compared to v3.4. Indeed, the BD-rate gain achieved by the bop model is comparatively smaller.

It's important to note that all these gains are relative to the VVC standard. The HOP model is a complexity reduction model, while the BOP model is a performance improvement model.

9.4.5 High Level Encoder Diagram and Decoder Diagram

The architecture of the encoder supports combinations of output picture format and scaling factors, and ultimately it outputs four bit streams, including Prediction and Entropy Information and Residual Information of Y and UV.

The decoder architecture receives Prediction and Entropy Information and Residual Information of Y and UV and restores the final image.

In such architecture, the YUV components are separated and processed separately. This has the advantage of reducing peak memory usage. Moreover, unlike conventional hyperprior methods, this architecture utilizes prediction/residual coding. This choice is made due to the presence of two-dimensional (2D) location and scale parameters when modeling the cumulative distribution function (CDF) using a Gaussian distribution. By employing residual coding, the location parameter becomes 0, resulting in a reduction of the parameter space to just one dimension (1D).

9.5 MPEG Video Coding for Machines Standard

9.5.1 *Introduction of MPEG VCM Standards*

The MPEG Video Coding for Machines (VCM) Standard, developed by the MPEG working group, are a set of international standards specifically designed for video coding and compression targeted at machine vision and machine-based video processing applications. These standards aim to optimize video coding techniques for efficient representation and transmission of video content in machine-driven environments.

The MPEG VCM standards are specifically designed to cater to the distinctive needs and features of machine vision systems. These standards prioritize key aspects like low-latency, high-throughput, and resilience to different types of video content. Their objective is to deliver cutting-edge compression algorithms that efficiently handle video data generated by machines. Such data often varies from conventional video contents aimed at human viewers in terms of complexity, characteristics, and intended usage [51, 103].

The scope of VCM encompasses the definition of a bitstream derived from compressed video or extracted features. This specialized bitstream is intended to be utilized in various machine-related applications, ensuring both high compression efficiency and optimal performance in machine-based intelligent tasks. By tailoring the bitstream to meet the requirements of machine vision systems, VCM strives to enable efficient and effective utilization of video data for a wide range of machine tasks.

By harnessing the capabilities of VCM, applications can efficiently leverage a single compressed video bitstream or extracted features to perform multiple machine vision tasks. This approach optimizes the utilization of storage and transmission resources, while achieving high compression efficiency. It empowers machine-based intelligent tasks in various domains, including video surveillance, smart traffic management, smart cities, smart industries, smart content delivery, and consumer electronics. By adopting VCM, these applications can benefit from enhanced efficiency, improved resource allocation, and seamless integration of machine vision technologies into diverse use cases [35, 106].

This section will highlight some of the experiments currently being explored by the latest MPEG-VCM framework.

9.5.2 *VCM Core Experiment 1*

The focus of this experiment is on Region of Interest (ROI) based video coding for machines. The experiment is scheduled to take place between the 12th and 13th meetings of WG 04.

m61460 [256]: The proposed approach involves dividing the input into a High Attention Region (HAR) and a Low Attention Region (LAR). A 64×64 grid structure is utilized to apply different scaling to the HAR and LAR, which remains constant for an entire Group of Pictures (GOP). The frames within a GOP are analyzed, and a common HAR structure is defined. The LAR is downsampled and encoded as a separate part of the same frame, resulting in a larger frame compared to the original.

Objects within the HAR can also be downsampled. The encoder has the option to code the LAR for each frame or only for the first frame within a GOP. If the latter option is chosen, the LAR is reused for all frames. This decision can be configured, and in the proposal, the LAR is only coded for the key frame. The VVC bitstream header is expanded to include necessary information for the decoder to decode frames containing both HAR and LAR.

m61519 [257]: It is based on the ISO/IEC SC29 MPEG Versatile Video Coding (VVC) standard and can be decoded using any compliant VVC decoder. EMVC incorporates an encoder-only module that determines a Quantization Parameter (QP) offset on a Coding Tree Unit (CTU) basis, taking into account the presence and size of objects. Furthermore, the Random Access configuration is modified to utilize a higher quantization parameter for frames on a high temporal layer. These enhancements are designed to optimize coding efficiency and improve the quality of machine vision applications.

m61520 [258]: The main difference from the previous proposals lies in the inclusion of an inpainting step, which involves removing and replacing small background objects with background color or content. The decision regarding which objects to remove is based on their object class and size.

9.5.3 VCM Core Experiment 2

As a potential technology in Video Coding for Machines (VCM), Neural Network-based Inner Coding (NNIC) is employed to encode and decode intra and inter frames in a video, leading to notable performance improvements.

m60753 [259]: Regarding the proposal, it suggested the utilization of a progressive probability model to estimate the probability distribution of elements in the quantized latent representation, also referred to as the latent tensor.

m60741 [260]: The proposed MIKEnc (Machine Intelligence oriented Key frame Encoder) system consists of three main components: PreP (preprocessing), which handles picture resizing, padding, and normalization; NN-FE (neural network for feature extraction), which extracts relevant features from the input data; and NN-FC (neural network for feature encoding), which encodes the extracted features.

In contrast, the MIKDec (Machine Intelligence oriented Key frame Decoder) system comprises three components: NN-FR (decoding neural network for feature reconstruction), responsible for reconstructing the encoded features; NN-IR (image reconstruction neural network), which generates the final reconstructed image; and

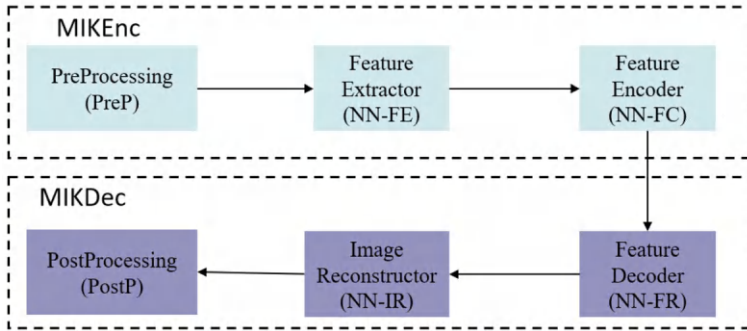


Fig. 9.16 Machine intelligence oriented key frame encoder and decoder. The image shown is introduced with MPEG open access (OA) work under CC BY Licence (Copyright © 1988–2024, www.mpeg.org) [260]

PostP (postprocessing), which performs denormalization and resizing operations on the reconstructed image.

For a visual representation of the MIKEnc and MIKDec architecture, please refer to Fig. 9.16.

The key components of MIKEnc and MIKDec are comprised of an end-to-end picture compression neural network, which is a modified version of Cheng2020. In order to streamline the codec, attention modules have been omitted in this design, and a single Gaussian likelihood model is derived for each latent code element. This revised version of Cheng2020 is divided into four segments: NN-FE, NN-FC, NN-FR, and NN-IR. NN-FE and NN-FC serve as the core components of MIKEnc, while NN-FR and NN-IR form the primary elements of MIKDec.

NN-FE functions as a feature extractor, incorporating three ResBlocks and two downsampling steps, enabling the extraction of essential features. On the other hand, NN-IR acts as a picture reconstructor, utilizing the decoded features and consisting of two ResBlocks along with two upsampling steps to reconstruct the image. The feature codec is established through the combined efforts of NN-FC and NN-FR.

m60775 [261]: The end-to-end image compression pipeline for machine vision object detection. This framework is specifically designed to cater to downstream machine vision tasks and is optimized jointly through the loss functions of its components.

Given that the feature distribution of infrared images differs from that of RGB images, m60775 employs adaptive threshold segmentation as a means to enhance the compression performance specifically for the FLIR dataset.

9.5.4 VCM Core Experiment 3

The objective of CE3 is to conduct an evaluation of spatial resampling technologies for VCM at the group level, providing valuable information and recommendations for the development of VCM standardization.

CE3.1. m61913: In m61913, the pre-analysis module has been updated and its details are described as follows. The motivation behind this update stems from the diversity of video content. The spatial resampling process is now adaptively performed based on the resolution of the input video and its spatial complexity, as determined by the pre-analysis stage. Specifically, if the resolution of the input video exceeds 2K (2560×1600), spatial resampling is applied. This is because high-resolution video content typically contains a higher degree of spatial redundancy. On the other hand, if the resolution of the input video is smaller than 1920×1080 , spatial resampling is not performed. For input videos with a resolution of 1920×1080 , spatial resampling is adaptively performed based on the spatial complexity, which is determined through pre-analysis. Object detection is conducted on the input video to extract the areas occupied by objects. The performance improvement achieved through spatial resampling is directly proportional to the size of the object areas. In other words, larger object areas yield greater performance gains with spatial resampling.

m60741 [260]: In the case of m60741, the input picture undergoes scaling operations to ensure consistent dimensions. Specifically, the picture is scaled proportionally so that the length of the shorter side becomes 800 pixels. Should the length of the longer side exceed 1333 pixels after scaling, it will be further reduced to 1333 pixels. The shorter side will then be adjusted proportionally to preserve the original aspect ratio of the image. This scaling procedure helps maintain a standardized size for the input picture, ensuring compatibility and facilitating subsequent processing or analysis.

CE3.2. m61455: This test focuses on studying the resampling method for input with a resolution higher than 1920×1080 . Given that the LIC, IHA, and IMA models are trained using images of resolution lower than 1920×1080 , any input video/image with a higher resolution undergoes downsampling by a factor of 3/4.

CE3.3. m64124 [262]: The proposed adaptive spatial resampling method, based on MI-RPR, consists of two stages. The first stage involves a pre-analysis step to determine the optimal scale factor for encoding. This is accomplished by utilizing Object Occupancy Distribution (OOD) information, which is obtained by analyzing the area occupied by valid objects within a single frame. By leveraging the derived adaptive scale factor, an appropriate downscaling ratio is applied to reduce the spatial resolution of the input frame prior to video encoding. This step ensures that machine vision tasks are not adversely affected by the spatial reduction. The second stage of the method involves a compression process using VTM-12.0 incorporated with MI-RPR. This compression process generates a bitstream that is specifically designed for machine vision tasks. Please refer to Fig. 9.17 for a visual representation of the MI-RPR method.

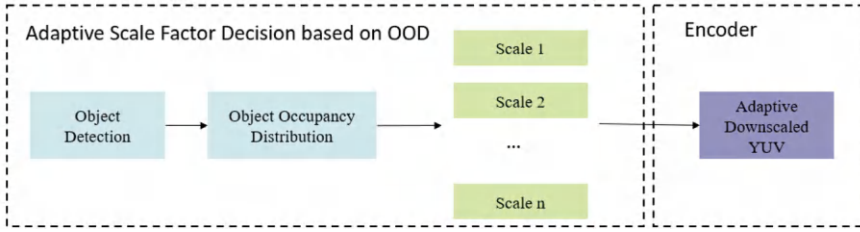


Fig. 9.17 Detailed flow chart of the MI-RPR method. The image shown is introduced with MPEG open access (OA) work under CC BY Licence (Copyright © 1988–2024, www.mpeg.org) [262]

CE3.4. m60741 [260]: In this proposal, the input picture undergoes proportional scaling to ensure that the length of the shorter side is set to 800 pixels. If, after scaling, the length of the longer side exceeds 1333 pixels, it is further scaled down to 1333 pixels while maintaining the proportional scaling of the shorter side.

This approach ensures that the image maintains its original aspect ratio while being resized to a specific target size. By setting the shorter side to 800 pixels, the proposal aims to standardize the dimensions of the input pictures. Additionally, by limiting the longer side to 1333 pixels, the proposal ensures that the images are not excessively large, which can help optimize processing and storage resources.

9.5.5 VCM Core Experiment 4

The group will proceed with the evaluation of temporal-related technologies, specifically focusing on two tests related to temporal resampling technology in this core experiment.

WG04N0336 [263]: This test examines the temporal resampling method described in WG04N0336. In this method, the original video undergoes temporal downsampling at the encoder side, meaning that some frames are dropped during the encoding process. These dropped frames are later recovered using an interpolation network at the decoder side.

m61465: This test experiment focuses on the extraction, encoding, and reconstruction of video frames. The process begins by uniformly sampling the frames of the video sequence at a fixed step size. This sampling results in a lower frame rate video, where some frames are removed or omitted.

The encoded video is then generated based on this sampled video sequence. The step size, which determines the frequency of removed frames, is represented by the first byte of the output stream. This encoded video contains the necessary information to reconstruct the low-frame rate video. During the decoding process, the decoder reconstructs the low-frame rate video using a combination of a neural network and the number of extracted frames. The frames from the low-frame rate video serve as reference frames during this reconstruction process. By leveraging the information

from the neural network and the available frames, the decoder aims to recreate the original video as closely as possible.

In this experiment, it is prohibited to remove the last frame of the original video, particularly when the total number of frames in the original video is an even number (e.g., m is an even number and k is odd). While this restriction may result in the reconstructed video having more frames than the original video, it is done to avoid any significant loss of task accuracy in the continuous content video.

9.5.6 VCM Core Experiment 5

This subsection will introduce 2 tests related to the post-resampling technique studied in this core experiment.

m61937: In this test, the focus is on studying a post-filtering method proposed in m61452. This method utilizes a neural network-based approach for post-filtering the decoded video frames. The post-filter takes the decoded video frame and the corresponding QP (Quantization Parameter) map as input and generates a filtered video frame as output. This architecture provides a visual representation of the structure and design of the network used for the post-filtering process. The network is trained to enhance the quality of the decoded video frames by leveraging the information from both the frame itself and the associated QP map.

m60753 [259]: In the research described in m61455, the Inter Machine Adapter (IMA) and Fallback IMA were proposed as post-filters for video frames. These filters utilize a network architecture that is described as follows: The Fallback IMA (F-IMA) takes the frame's quantization parameter (QP) and input resolution as additional inputs. These inputs are injected into multiple layers within an autoencoder. On the other hand, the IMA does not require any extra inputs. It operates solely based on the video frames without considering additional information such as QP or resolution. The filter mode used by both IMA and F-IMA is selected based on the value of the quantization parameter. This means that the specific filtering technique or parameters employed by the filters may vary depending on the quantization level of the frames.

9.6 MPAI End-to-End Video Coding Standard

AI-based End-to-End Video Coding (MPAI-EEV) is a standard project initiated by the Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) with the aim of reducing the bit rate needed to represent video through the utilization of AI-based visual data coding technologies. Unlike traditional video coding methods, MPAI-EEV explores new approaches that are not limited by the conventional constraints of video coding. As of now, MPAI-EEV has advanced to the 4th version of the reference model. This section will provide a brief introduction to this standard. MPAI conducted comprehensive encoding experiments to evaluate the rate-distortion

efficiency of UAV video. Additionally, MPAI employed the MPAI End-to-end Video (EEV) reference model specifically designed for UAV video compression. By comparing the baseline coding results from these three codecs, MPAI gained insights into their performance. The codecs are visually represented in a schematic diagram, where the classical hybrid codec is shown in the left panel, while the learned codecs, OpenDVC and EEV, are depicted in the remaining panels.

Train-and-test data consistency is a crucial factor to consider when evaluating learned codecs. In the machine learning community, it is widely acknowledged that the training and testing data should be independent and identically distributed. Nevertheless, in the case of OpenDVC and EEV codecs, both are trained using the Vimeo-90k dataset, which consists of natural videos, with mean square error (MSE) as the distortion metric. This approach ensures that the evaluation is conducted in a general setting, maintaining consistency with the training data. By using the pre-trained weights, the performance of the codecs can be objectively assessed and compared on the drone video dataset. MPAI-EEV demonstrates a significant BD-rate gain of 19.84% compared to DVC and a BD-rate gain of 15.23% compared to HEVC.

Since its official formation in November 2021, the MPAI-EEV project has successfully developed three learning-based methods as the corresponding reference models. Through these iterative developments, MPAI-EEV is continuously pushing the boundaries of end-to-end video coding and striving to become a valuable resource for AI-based video coding techniques. By driving innovation and advancements in this field, the MPAI-EEV project aims to contribute significantly to the progress of AI-based video coding technologies.

9.7 Summary

AI-based image and video compression has evolved rapidly, prompting several organizations to establish standards for its development and integration. This section compares the main AI-based compression standards, focusing on IEEE, JPEG AI, MPEG VCM, and MPAI-EEV.

Compression Efficiency and Quality

The AI-based image and video compression standards approach compression efficiency and quality in different ways. IEEE 1857.11, also known as Neural Image Coding (NIC), excels in compression efficiency. JPEG AI, in contrast, focuses on incremental improvements to traditional JPEG methods, emphasizing compatibility with existing workflows while offering moderate enhancements in compression efficiency by integrating AI-based techniques. MPEG VCM takes a different approach, focusing on maintaining the quality required for machine vision analysis tasks, including object detection and scene understanding, while ensuring that compression remains efficient for machine analysis. Finally, MPAI-EEV (AI-based End-to-End Video Coding) focuses on breaking away from traditional video coding constraints by adopting an end-to-end approach, which aims to provide highly efficient bit-rate

representation along with enhanced visual quality. These varying approaches reflect the unique priorities of each standard, whether it is backward compatibility, efficiency for machine-based analysis, or pushing the boundaries of compression technology through end-to-end AI integration.

Target Applications

IEEE NIC and JPEG AI primarily target general purpose image compression, with NIC optimising compression efficiency and JPEG AI considering both human visual quality and machine coding tasks. MPEG VCM, in contrast, is specifically designed for machine vision applications, where the priority is to retain critical information for machine tasks such as robotics, autonomous vehicles, and AI analysis. Finally, MPAI-EEV emphasizes AI-first approaches to video compression, applicable to both machine vision and human viewing scenarios, with a focus on flexible and highly efficient coding for different contexts.

Adoption and Compatibility

The objective of JPEG AI is to facilitate a seamless transition for industries that rely on traditional JPEG by ensuring backward compatibility. Although the IEEE NIC does not prioritize backward compatibility, its objective is to achieve optimal efficiency through neural coding. MPEG VCM is oriented towards emerging technologies such as autonomous vehicles and robotics, where compatibility is less of a priority than the capacity to analyse machine-readable compressed content. In contrast, MPAI-EEV provides a more comprehensive basis for innovation, encompassing end-to-end AI strategies that enhance both machine and human viewing scenarios.

This chapter offers a comprehensive overview of the development of IEEE AI standards, highlighting their significance in the field of AI-based image and video compression. A brief overview of recent developments in the JPEG AI, MPEG VCM, and MPAI-EEV standards is also provided, offering valuable insights into the latest standardization efforts in AI-based compression. By exploring these standards, readers can gain a deeper understanding of the current state of AI-based image and video compression and stay informed about the latest advances in this rapidly evolving field.

Exercises

1. Please outline the current AI-based image and video compression standards.
2. Please explain what reference software generally refers to in compression standards.
3. Please list the reference software in IEEE's current AI-based image compression standards.
4. Please describe the advantages and disadvantages of the reference software mentioned above.

5. Please specify the future development plans for the IEEE AI-based image compression standard.
6. Please describe the scope of JPEG AI.
7. Please specify the quality metrics used in JPEG AI.
8. Please list some of the multi-tasking capabilities of MPEG VCM.
9. What are the scope of MPEG VCM?
10. Please describe the scope of MPEG VCM.

References

1. S. Fan, W. Gao, G. Li, Salient object detection for point clouds, in *European Conference on Computer Vision* (Springer, 2022), pp. 1–19
2. S. Luo, W. Gao, A general framework for rotation invariant point cloud analysis, in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2024), pp. 3665–3669
3. X. Lu, W. Gao, Attentivenet: Detecting small objects for lidar point clouds by attending to important points, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
4. Z. Pan, N. Zhang, W. Gao, S. Liu, G. Li, Less is more: label recommendation for weakly supervised point cloud semantic segmentation. *Proc. AAAI Conf. Artif. Intell.* **38**(5), 4397–4405 (2024)
5. Z. Pan, G. Liu, W. Gao, T.H. Li, Epcontrast: effective point-level contrastive learning for large-scale point cloud understanding, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024), pp. 1–6
6. N. Zhang, Z. Pan, T.H. Li, W. Gao, G. Li, Improving graph representation for point cloud segmentation via attentive filtering, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 1244–1254
7. K. Wen, N. Zhang, G. Li, W. Gao, Mpvnn: multi-resolution point-voxel non-parametric network for 3d point cloud processing, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024), pp. 1–6
8. Z. Pan, N. Zhang, W. Gao, S. Liu, G. Li, Point cloud semantic segmentation with sparse and inhomogeneous annotations, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
9. R. Zhang, G. Li, W. Gao, S. Liu, A quantum-inspired framework in master-servant mode for large-scale multi-modal place recognition, *IEEE Transactions on Intelligent Transportation Systems* (2024)
10. D. Yang, W. Gao, Pointchd: a point cloud benchmark for congenital heart disease classification and segmentation. *IEEE J. Biomed. Health Inform.* (2024)
11. S. Fan, W. Gao, G. Li, Point-mpp: Point cloud self-supervised learning from masked position prediction. *IEEE Trans. Neural Netw. Learn. Syst.* (2024)
12. Z. Pan, W. Gao, G. Li, Distribution guidance network for weakly supervised point cloud semantic segmentation, in *Advances in Neural Information Processing Systems* (2024)
13. R. Zhang, G. Li, W. Gao, T.H. Li, Compoinet: can complex-valued representation benefit point cloud place recognition? *IEEE Trans. Intell. Transp. Syst.* (2024)
14. X. Zang, W. Gao, G. Li, H. Fang, C. Ban, Z. He, H. Sun, A baseline investigation: Transformer-based cross-view baseline for text-based person search, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 7737–7746
15. G. Liao, W. Gao, G. Li, J. Wang, S. Kwong, Cross-collaborative fusion-encoder network for robust rgb-thermal salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **32**(11), 7646–7661 (2022)

16. W. Gao, G. Liao, S. Ma, G. Li, Y. Liang, W. Lin, Unified information fusion network for multi-modal rgb-d and rgb-t salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **32**(4), 2091–2106 (2021)
17. Y. Chen, S. Sun, G. Li, W. Gao, T.H. Li, Closing the gap between theory and practice during alternating optimization for gans. *IEEE Trans. Neural Netw. Learn. Syst.* (2023)
18. Y. Chen, C. Jin, G. Li, T.H. Li, W. Gao, Mitigating label noise in gans via enhanced spectral normalization. *IEEE Trans. Circuits Syst. Video Technol.* (2023)
19. X. Zang, G. Li, W. Gao, Multidirection and multiscale pyramid in transformer for video-based pedestrian retrieval. *IEEE Trans. Ind. Inform.* **18**(12), 8776–8785 (2022)
20. X. Zang, G. Li, W. Gao, X. Shu, Learning to disentangle scenes for person re-identification. *Image Vis. Comput.* **116**, 104330 (2021)
21. X. Zang, G. Li, W. Gao, X. Shu, Exploiting robust unsupervised video person re-identification. *IET Image Process.* **16**(3), 729–741 (2022)
22. Z. Yue, G. Li, W. Gao, Cross-level guided attention for human-object interaction detection, in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2023) pp. 284–289
23. S. Sun, J. Liu, H. Li, G. Liu, T. H. Li, W. Gao, Streamflow: streamlined multi-frame optical flow estimation for video sequences, in *Advances in Neural Information Processing Systems* (2024)
24. Z. Yao, W. Gao, Iterative saliency aggregation and assignment network for efficient salient object detection in optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* (2024)
25. G. Liao, W. Gao, Rethinking feature mining for light field salient object detection. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(10), 1–24 (2024)
26. B. Yang, W. Gao, G. Li, Focus and adjust: progressive refinement network for human object interaction detection, in *2022 26th International Conference on Pattern Recognition (ICPR)* (IEEE, 2022), pp. 2546–2552
27. G. Liao, W. Gao, Q. Jiang, R. Wang, G. Li, Mmnet: multi-stage and multi-scale fusion network for rgb-d salient object detection, in *Proceedings of the 28th ACM International Conference on Multimedia* (2020), pp. 2436–2444
28. W. Gao, S. Kwong, Y. Zhou, Y. Jia, J. Zhang, W. Wu, Multiscale phase congruency analysis for image edge visual saliency detection, in *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 1 (IEEE, 2016), pp. 75–80
29. W. Gao, S. Kwong, Y. Zhou, X. Wang, Phase congruency analysis of down-sampled and blurring images for foreground extraction, in *2015 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)* (IEEE, 2015), pp. 152–157
30. Y. Guo, W. Gao, G. Li, Interpretable task-inspired adaptive filter pruning for neural networks under multiple constraints. *Int. J. Comput. Vis.* 1–17 (2024)
31. W. Gao, Y. Guo, S. Ma, G. Li, S. Kwong, Efficient neural network compression inspired by compressive sensing. *IEEE Trans. Neural Netw. Learn. Syst.* (2022)
32. Y. Guo, W. Gao, Semantic-driven automatic filter pruning for neural networks, in *IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
33. L. Tao, W. Gao, Efficient channel pruning based on architecture alignment and probability model bypassing, in *IEEE international conference on systems, man, and cybernetics (SMC)* (IEEE, 2021), pp. 3232–3237
34. Y. Li, Z. Yu, X. Chen, S. Wang, H. Yuan, W. Gao, Zero-shot quantization for large-kernels via shape-based distribution and diversity self-distillation, in *ICASSP 2025–2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2025)
35. Y. Wu, L. Xie, S. Sun, W. Gao, Y. Yan, Adaptive intra period size for deep learning-based screen content video coding, in *IEEE International Conference on Multimedia and Expo Workshops* (2024), pp. 1–6
36. H. Zheng, W. Gao, End-to-end rgb-d image compression via exploiting channel-modality redundancy. *Proc. AAAI Conf. Artif. Intell.* **38**(7), 7562–7570 (2024)
37. L. Tao, W. Gao, G. Li, C. Zhang, Adanic: towards practical neural image compression via dynamic transform routing, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 16 879–16 888

38. Y. Wu, W. Gao, End-to-end lossless compression of high precision depth maps guided by pseudo-residual (2022). [arXiv:2201.03195](https://arxiv.org/abs/2201.03195)
39. Y. Wu, Z. Qi, H. Zheng, L. Tao, W. Gao, Deep image compression with latent optimization and piece-wise quantization approximation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 1926–1930
40. X. Chen, Y. Wang, Y. Li, X. Ling, M. Li, R. Liu, M. Ouyang, K. Zhao, T. Guan, Y. He, Low bit-width zero-shot quantization with soft feature-infused hints for iot systems. *IEEE Internet Things J.* (2024)
41. W. Gao, L. Tao, L. Zhou, D. Yang, X. Zhang, Z. Guo, Low-rate image compression with super-resolution learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2020), pp. 154–155
42. W. Gao, S. Sun, H. Zheng, Y. Wu, H. Ye, Y. Zhang, Opendmc: an open-source library and performance evaluation for deep-learning-based multi-frame compression, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9685–9688
43. W. Gao, H. Zheng, C. Zhang, K. Zheng, Z. Yu, Y. Li, H. Ye, Y. Zhang, Opendic: An open-source library and performance evaluation for deep-learning-based image compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 202–11 205
44. Y. Li, W. Gao, G. Li, S. Ma, Saliency segmentation oriented deep image compression with novel bit allocation. *IEEE Trans. Image Process.* **34**, 16–29 (2025)
45. C. Zhang, W. Gao, Learned rate control for frame-level adaptive neural video compression via dynamic neural network, in *European Conference on Computer Vision* (Springer, 2024)
46. L. Xie, W. Gao, H. Zheng, G. Li, Spcgc: scalable point cloud geometry compression for machine vision, in *2024 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2024), pp. 17 272–17 278
47. L. Xie, W. Gao, H. Zheng, H. Ye, Semantic-aware visual decomposition for point cloud geometry compression, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 595–595
48. Z. Qi, W. Gao, Variable-rate point cloud geometry compression based on feature adjustment and interpolation, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 63–72
49. Z. Yu, W. Gao, When dynamic neural network meets point cloud compression: Computation-aware variable rate and checkerboard context, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 600–600
50. L. Xie, W. Gao, S. Fan, Z. Yao, Pdnet: parallel dual-branch network for point cloud geometry compression and analysis, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 596–596
51. L. Xie, W. Gao, H. Zheng, End-to-end point cloud geometry compression and analysis with sparse tensor, in *International Workshop on Advances in Point Cloud Compression, Processing and Analysis* (2022), pp. 27–32
52. C. Fu, G. Li, R. Song, W. Gao, S. Liu, Octattention: octree-based large-scale contexts model for point cloud compression. *Proc. AAAI Conf. Artif. Intell.* **36**(1), 625–633 (2022)
53. C. Zhang, W. Gao, Adadpcc: adaptive rate control and rate-distortion-complexity optimization for dynamic point cloud compression, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
54. K. Wang, W. Gao, Unipcgc: towards practical point cloud geometry compression via an efficient unified approach, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
55. K. Zheng, W. Gao, H. Zheng, Octree-based learned point cloud geometry compression: a lossy perspective, in *2025 Data Compression Conference (DCC)* (IEEE, 2025)
56. Z. Qi, H. Liu, W. Gao, Variable-rate point cloud compression with channel-wise entropy modeling, in *2025 Data Compression Conference (DCC)* (IEEE, 2025)
57. H. Zheng, W. Gao, Z. Yu, T. Zhao, G. Li, Viewpcgc: view-guided learned point cloud geometry compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 7152–7161

58. L. Xie, W. Gao, H. Zheng, G. Li, Roi-guided point cloud geometry compression towards human and machine vision, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 3741–3750
59. T. Qin, G. Li, W. Gao, S. Liu, Multi-grained point cloud geometry compression via dual-model prediction with extended octree. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(9), 1–30 (2024)
60. Y. Shao, W. Gao, S. Liu, G. Li, Advanced patch-based affine motion estimation for dynamic point cloud geometry compression. *Sensors* **24**(10), 3142 (2024)
61. Y. Shao, F. Song, W. Gao, S. Liu, G. Li, Texture-guided graph transform optimization for point cloud attribute compression. *Appl. Sci.* **14**(10), 4094 (2024)
62. Y. Shao, X. Yang, W. Gao, S. Liu, G. Li, 3d point cloud attribute compression using diffusion-based texture-aware intra prediction. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
63. J. Zhang, Y. Chen, G. Liu, W. Gao, G. Li, Efficient point cloud attribute compression framework using attribute-guided graph fourier transform, in *ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2024), pp. 8426–8430
64. W. Gao, H. Yuan, G. Li, Z. Li, H. Yuan, Low complexity coding unit decision for video-based point cloud compression. *IEEE Trans. Image Process.* **33**, 149–162 (2023)
65. Y. Shao, G. Li, Q. Zhang, W. Gao, S. Liu, Nonrigid registration-based progressive motion compensation for point cloud geometry compression. *IEEE Trans. Geosci. Remote Sens.* (2023)
66. F. Song, G. Li, X. Yang, W. Gao, S. Liu, Block-adaptive point cloud attribute coding with region-aware optimized transform. *IEEE Trans. Circuits Syst. Video Technol.* (2023)
67. Y. An, Y. Shao, G. Li, W. Gao, S. Liu, A fast motion estimation method with hamming distance for lidar point cloud compression, in *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2022), pp. 1–5
68. H. Yuan, W. Gao, G. Li, Z. Li, Rate-distortion-guided learning approach with cross-projection information for v-pcc fast cu decision, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 3085–3093
69. F. Song, G. Li, W. Gao, T.H. Li, Rate-distortion optimized graph for point cloud attribute coding. *IEEE Signal Process. Lett.* **29**, 922–926 (2022)
70. F. Song, G. Li, X. Yang, W. Gao, T.H. Li, Fine-grained correlation representation for graph-based point cloud attribute compression, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
71. F. Shen, W. Gao, A rate control algorithm for video-based point cloud compression, in *2021 International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2021), pp. 1–5
72. F. Song, Y. Shao, W. Gao, H. Wang, T. Li, Layer-wise geometry aggregation framework for lossless lidar point cloud compression. *IEEE Trans. Circuits Syst. Video Technol.* **31**(12), 4603–4616 (2021)
73. Z. Cai, W. Gao, G. Li, W. Gao, Distortion propagation model-based v-pcc rate control for 3d point cloud broadcasting. *IEEE Trans. Broadcast.* 1–13 (2024)
74. X. Mu, W. Gao, H. Yuan, S. Wang, G. Li, Adaptive lpu decision for dynamic point cloud compression. *IEEE Signal Process. Lett.* (2024)
75. C. Peng, W. Gao, Laplacian matrix learning for point cloud attribute compression with ternary search-based adaptive block partition, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 10 412–10 420
76. Y. Wang, W. Gao, X. Mu, H. Yuan, Rate control optimization for joint geometry and attribute coding of lidar point clouds, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
77. T. Guo, H. Yuan, T. Wang, W. Gao, Graph filter-based fast motion matching for inter frame coding of mpeg g-pcc, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 1151–1155

78. Z. Yang, W. Gao, G. Li, Y. Yan, Sur-driven video coding rate control for jointly optimizing perceptual quality and buffer control. *IEEE Trans. Image Process.* (2023)
79. F. Shen, Z. Cai, W. Gao, An efficient rate control algorithm for intra frame coding in avs3, in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2021), pp. 3164–3169
80. H. Yuan, W. Gao, J. Wang, Dynamic computational resource allocation for fast inter frame coding in video conferencing applications, in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6
81. W. Gao, Q. Jiang, R. Wang, S. Ma, G. Li, S. Kwong, Consistent quality oriented rate control in hevc via balancing intra and inter frame coding. *IEEE Trans. Ind. Inform.* **18**(3), 1594–1604 (2021)
82. H. Yuan, W. Gao, A new coding unit partitioning mode for screen content video coding, in *Proceedings of the 2021 5th International Conference on Digital Signal Processing* (2021), pp. 66–72
83. W. Gao, On the performance evaluation of state-of-the-art rate control algorithms for practical video coding and transmission systems, in *Proceedings of the 2020 4th International Conference on Video and Image Processing* (2020), pp. 179–185
84. W. Gao, S. Kwong, Q. Jiang, C.-K. Fong, P.H. Wong, W.Y. Yuen, Data-driven rate control for rate-distortion optimization in hevc based on simplified effective initial qp learning. *IEEE Trans. Broadcast.* **65**(1), 94–108 (2018)
85. W. Gao, A multi-objective optimization perspective for joint consideration of video coding quality, in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (IEEE, 2019), pp. 986–991
86. W. Gao, S. Kwong, Y. Jia, Joint machine learning and game theory for rate control in high efficiency video coding. *IEEE Trans. Image Process.* **26**(12), 6074–6089 (2017)
87. W. Gao, S. Kwong, Y. Zhou, H. Yuan, Ssim-based game theory approach for rate-distortion optimized intra frame ctu-level bit allocation. *IEEE Trans. Multimedia* **18**(6), 988–999 (2016)
88. W. Gao, S. Kwong, H. Yuan, X. Wang, Dct coefficient distribution modeling and quality dependency analysis based frame-level bit allocation for hevc. *IEEE Trans. Circuits Syst. Video Technol.* **26**(1), 139–153 (2015)
89. W. Gao, S. Kwong, Phase congruency based edge saliency detection and rate control for perceptual image and video coding, in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2016), pp. 000 264–000 269
90. H. Yuan, W. Gao, Openfastvc: an open source library for video coding fast algorithm implementation, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9660–9663
91. H. Yuan, W. Gao, S. Ma, Y. Yan, Divide-and-conquer-based rdo-free cu partitioning for 8k video compression. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(4), 1–20 (2024)
92. L. Tao, W. Gao, A hardware implementation of entropy encoder for 8k video coding, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
93. Y. Guo, W. Gao, S. Ma, G. Li, Accelerating transform algorithm implementation for efficient intra coding of 8k uhd videos. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* **18**(4), 1–20 (2022)
94. Z. Cai, W. Gao, Efficient fast algorithm and parallel hardware architecture for intra prediction of avs3, in *IEEE International Symposium on Circuits and Systems (ISCAS)* (IEEE, 2021), pp. 1–5 (2021)
95. W. Gao, H. Yuan, Y. Guo, L. Tao, Z. Cai, G. Li, Openhardwarevc: an open source library for 8k uhd video coding hardware implementation, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 7339–7342
96. W. Gao, H. Yuan, G. Liao, Z. Guo, J. Chen, Pp8k: a new dataset for 8k uhd video compression and processing. *IEEE MultiMedia* (2023)
97. W. Liu, W. Gao, G. Li, S. Ma, T. Zhao, H. Yuan, Enlarged motion-aware and frequency-aware network for compressed video artifact reduction. *IEEE Trans. Circuits Syst. Video Technol.* (2024)

98. W. Gao, G. Li, H. Yuan, R. Hamzaoui, Z. Li, S. Liu, Apccpa'22: 1st international workshop on advances in point cloud compression, processing and analysis, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 7392–7393
99. W. Gao, G. Li, Point cloud compression, enhancement and applications: from 3d perception to large models, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 292–11 293
100. H. Sun, G. Li, S. Chen, L. Li, W. Gao, Special topic on 3d point cloud processing and applications. *ZTE Commun.* **21**(4), 1 (2023)
101. W. Gao, W. Gao, X. Mu, C. Peng, G. Li, Overview and comparison of avs point cloud compression standard. *APSIPA Trans. Signal Inf. Process.* (2024)
102. W. Gao, S. Wang, X. Zhang, S. Kwong, Editorial for special issue on deep learning-based data compression. *APSIPA Trans. Signal Inf. Process.* **13**(6) (2024)
103. W. Gao, H. Ye, G. Li, H. Zheng, Y. Wu, L. Xie, OpenPointCloud: an open-source algorithm library of deep learning based point cloud compression, in *ACM International Conference on Multimedia* (2022), pp. 7347–7350
104. Y. Zhang, W. Gao, G. Li, Openpointcloud-v2: a deep learning based open-source algorithm library of point cloud processing, in *Proceedings of the 1st International Workshop on Advances in Point Cloud Compression, Processing and Analysis* (2022), pp. 51–55
105. L. Xie, W. Gao, Learningpcc: a pytorch library for learning-based point cloud compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 234–11 238
106. L. Xie, W. Gao, Pchmvision: an open-source library of point cloud compression for human and machine vision, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 239–11 243
107. L. Xie, X. Mu, G. Li, W. Gao et al., Pku-dpcc: a new dataset for dynamic point cloud compression. *APSIPA Trans. Signal Inf. Process.* **13**(6) (2024)
108. H. Yuan, W. Gao, W. Gao, Opensep: an open source subjective experiment platform, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 248–11 251
109. S. Fan, W. Gao, Screen-based 3d subjective experiment software, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9672–9675
110. J. Wang, W. Gao, G. Li, Zoom to perceive better: No-reference point cloud quality assessment via exploring effective multiscale feature. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
111. J. Wang, W. Gao, G. Li, Applying collaborative adversarial learning to blind point cloud quality measurement. *IEEE Trans. Instrum. Meas.* (2023)
112. H. Li, W. Gao, Information exploration of projected views for point cloud quality measurement. *IEEE Trans. Instrum. Meas.* (2024)
113. B. Qu, X. Liang, S. Sun, W. Gao, Exploring aigc video quality: a focus on visual harmony, video-text consistency and domain distribution gap, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2024)
114. B. Qu, H. Li, W. Gao, Bringing textual prompt to ai-generated image quality assessment, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024)
115. S. Sun, X. Liang, S. Fan, W. Gao, W. Gao, Ve-bench: Subjective-aligned benchmark suite for text-driven video editing quality assessment, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
116. Z. Guo, W. Gao, H. Wang, J. Wang, S. Fan, No-reference deep quality assessment of compressed light field images, in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6
117. W. Liu, W. Gao, X. Mu, Fast inter-frame motion prediction for compressed dynamic point cloud attribute enhancement. *Proceedings of the AAAI Conference on Artificial Intelligence* **38**(4), 3720–3728 (2024)
118. Z. Yang, W. Gao, X. Lu, Danet: density-adaptive network for geometry-based point cloud compression artifacts removal, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5

119. X. Fan, G. Li, D. Li, Y. Ren, W. Gao, T.H. Li, Deep geometry post-processing for decompressed point clouds, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
120. X. Zhang, G. Liao, W. Gao, G. Li, Tdnet: transformer-based dual-branch restoration network for geometry based point cloud compression artifacts, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
121. Z. Li, G. Li, T.H. Li, S. Liu, W. Gao, Semantic point cloud upsampling. *IEEE Trans. Multimedia* (2022)
122. Z. Li, S. Liu, W. Gao, G. Li, G. Li, S4r: rethinking point cloud sampling via guiding upsampling-aware perception. *IEEE Trans. Multimedia* (2025)
123. R. Zhang, W. Gao, G. Li, T.H. Li, Qinet: decision surface learning and adversarial enhancement for quasi-immune completion of diverse corrupted point clouds. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–14 (2022)
124. R. Bao, Y. Ren, G. Li, W. Gao, S. Liu, Flow-based point cloud completion network with adversarial refinement, in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2559–2563
125. J. Chen, G. Li, R. Zhang, T.H. Li, W. Gao, Pointvae: invertible variational autoencoder framework for 3d point cloud generation, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 3216–3220
126. R. Zhang, J. Chen, W. Gao, G. Li, T.H. Li, Pointot: interpretable geometry-inspired point cloud generative model via optimal transport. *IEEE Trans. Circuits Syst. Video Technol.* **32**(10), 6792–6806 (2022)
127. Z. Li, G. Li, T. Li, S. Liu, W. Gao, Information-growth attention network for image super-resolution, in *Proceedings of the 29th ACM International Conference on Multimedia* (2021), pp. 544–552
128. L. Zhou, W. Gao, G. Li, H. Yuan, T. Zhao, G. Yue, Disentangled feature distillation for light field super-resolution with degradations, in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2023), pp. 116–121
129. L. Zhou, W. Gao, G. Li, End-to-end spatial-angular light field super-resolution using parallax structure preservation strategy, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 3396–3400
130. W. Gao, L. Zhou, L. Tao, A fast view synthesis implementation method for light field applications. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* **17**(4), 1–20 (2021)
131. X. Zhang, W. Gao, G. Li, Q. Jiang, R. Cong, Image quality assessment-driven reinforcement learning for mixed distorted image restoration. *ACM Trans. Multimedia Comput. Commun. Appl.* **19**(1s), 1–23 (2023)
132. X. Zhang, W. Gao, H. Yuan, G. Li, Je 2 net: joint exploitation and exploration in reinforcement learning based image restoration, in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2090–2094
133. X. Zhang, W. Gao, Hirl: hybrid image restoration based on hierarchical deep reinforcement learning via two-step analysis, in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2445–2449
134. S. Wang, Y. Lu, W. Xia, P. Xia, Z. Wang, W. Gao, Light field angular super-resolution by view-specific queries. *Visual Comput.* 1–16 (2024)
135. Y. Sun, Z. Li, S. Wang, W. Gao, Depth-assisted calibration on learning-based factorization for a compressive light field display. *Opt. Express* **31**(4), 5399–5413 (2023)
136. Y. Sun, Z. Li, L. Li, S. Wang, W. Gao, Optimization of compressive light field display in dual-guided learning, in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2075–2079
137. W. Gao, S. Fan, G. Li, W. Lin, A thorough benchmark and a new model for light field saliency detection. *IEEE Trans. Pattern Anal. Mach. Intell.* (2023)
138. G. Li, G. Wei, W. Gao, *Point Cloud Compression: Technologies and Standardization* (Springer Nature, 2024)

139. G. Li, W. Gao, W. Gao, Introduction, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 1–28
140. G. Li, W. Gao, W. Gao, Background knowledge, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 29–51
141. G. Li, W. Gao, W. Gao, Predictive coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 53–70
142. G. Li, W. Gao, W. Gao, Transform coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 71–96
143. G. Li, W. Gao, W. Gao, Quantization techniques, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 97–112
144. G. Li, W. Gao, W. Gao, Entropy coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 113–133
145. G. Li, W. Gao, W. Gao, Mpeg geometry-based point cloud compression (g-pcc) standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 135–165
146. G. Li, W. Gao, W. Gao, Avs point cloud compression standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 167–197
147. G. Li, W. Gao, W. Gao, Mpeg video-based point cloud compression (v-pcc) standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 199–218
148. G. Li, W. Gao, W. Gao, Mpeg ai-based 3d graphics coding standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 219–241
149. G. Li, W. Gao, W. Gao, Future work, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 243–250
150. T. Chen, H. Liu, M. Lu, Z. Ma, Description of deep learning-based image compression technology proposal by nanjing university, Technical Report. VC-30-M256 (2020)
151. T. Chen, H. Liu, Z. Ma, Q. Shen, X. Cao, Y. Wang, End-to-end learnt image compression via non-local attention optimization and improved context modeling. *IEEE Trans. Image Process.* **30**, 3179–3191 (2021)
152. X. Wang, T. Chen, Z. Ma, D. Wang, Z. Xie, F. Wang, Block-based neural codec, Technical Report VC-31-M265 (2020)
153. X. Wang, T. Chen, Z. Ma, D. Wang, Z. Xie, F. Wang, Block-based neural codec with block enhancement, Technical Report VC-30-M266 (2020)
154. D. Ding, W. Jiang, W. Wang, S. Liu, Substitutional neural image compression, Technical Report VC-30-M267 (2020)
155. R. Yang, D. Liu, F. Wu, Perceptual quality oriented end-to-end compression, Technical Report VC-30-M268 (2020)
156. Y. Hu, D. Wang, J. Liu, Fast decoding end-to-end learned image coding framework for nic-0.1, Technical Report VC-30-M269 (2020)
157. X. Wang, T. Chen, Z. Ma, D. Wang, Z. Xie, F. Wang, Codec using overlapping blocks, Technical Report VC-32-M273 (2020)
158. X. Wang, T. Chen, Z. Ma, D. Wang, Z. Xie, F. Wang, The impact of block size on codec performance, Technical Report VC-32-M274 (2020)
159. D. Wang, Y. Hu, J. Liu, Efficient coarse-to-fine hyperprior model for nic, Technical Report VC-32-M275 (2020)
160. D. Ding, W. Jiang, S. Lin, W. Wang, X. Xu, S. Liu, Substitutional neural image compression, Technical Report VC-33-M279 (2021)
161. J. Lin, H. Zhang, D. Liu, Algorithm discription of variable-rate nic model, Technical Report VC-33-M280 (2021)
162. F. Liu, Y. Zhang, A. Feng, G. Ren, Z. Chen, Subective quality assessment test results of ee2 and new nic baseline, Technical Report VC-33-M281 (2021)
163. A. Feng, Y. Wang, D. Liu, X. Wang, Y. Chen, Geometric operation and adaptive quad-tree partition of the input image for nic, Technical Report VC-33-M282 (2021)
164. Z. Ge, C. Jia, S. Wang, S. Ma, A D- λ model for distortion control in NIC, Technical Report VC-33-M283 (2021)

165. D. Wang, Y. Hu, J. Liu, Cube-level context model for nic acceleration, Technical Report VC-33-M284 (2021)
166. D. Wang, Y. Hu, J. Liu, Dual hyper-prior model for nic, Technical Report VC-33-M285 (2021)
167. Y. Chen, W. Lin, Pyramid multi-layer structure for nic, Technical Report VC-34-M291 (2021)
168. Y. Chen, W. Lin, Triple hyper probabilities for nic, Technical Report VC-34-M292 (2021)
169. D. Ding, W. Wang, S. Liu, Online training-based post filtering, Technical Report VC-34-M293 (2021)
170. H. Zhu, H. Wang, F. Liu, Z. Chen, Exploration of joint loss function, Technical Report VC-34-M294 (2021)
171. D. Wang, Y. Hu, J. Liu, Multi hyper-prior model for nic, Technical Report VC-34-M295 (2021)
172. S. Yin, Y. Bao, F. Meng, Y. Liang, W. Liu, Algorithm discription for image compression with self-adaptive computational cost, Technical Report VC-35-M299 (2021)
173. Z. Ge, C. Jia, S. Wang, S. Ma, An entropy coding accelerating algorithm for nic, Technical Report VC-35-M300 (2021)
174. D. Ding, W. Wang, S. Liu, Time complexity analysis of preprocessing, Technical Report VC-35-M301 (2021)
175. D. Ding, W. Wang, S. Liu, Online training-based post filtering, Technical Report VC-35-M302 (2021)
176. Y. Chen, W. Lin, Coding context prediction using hyper prior, Technical Report VC-35-M304 (2021)
177. D. Wang, Y. Hu, J. Liu, Cube-level context model for nic acceleration, Technical Report VC-35-M306 (2021)
178. Y. Peng, Z. Chen, Analysis on fvc subject dataset and proposal of mmfn metric for fvc, Technical Report VC-36-M309 (2021)
179. Z. Ge, C. Jia, S. Wang, S. Ma, An improved entropy coding algorithm for nic, Technical Report VC-36-M310 (2021)
180. D. Ding, W. Wang, S. Liu, Online training-based post enhancement, Technical Report VC-36-M311 (2021)
181. H. Ma, H. Zhang, L. Li, D. Liu, Suggestions on the standardization scope of neural network-based image coding standard, Technical Report VC-36-M312 (2021)
182. H. Zhang, J. Lin, D. Liu, Algorithm description of vr-models for nic-0.3 and nic-0.4, Technical Report VC-36-M313 (2021)
183. H. Zhang, J. Lin, L. Li, D. Liu, Rate control for nic-0.4 through r-lambda model, Technical Report VC-36-M314 (2021)
184. D. Ding, X. Xu, S. Liu, Discussion of randomness in online-training methods, Technical Report VC-36-M316 (2021)
185. S. Yin, C. Li, Y. Liang, F. Meng, W. Liu, Channel slimming of main codec for real-time inference acceleration, Technical Report VC-37-M319 (2022)
186. Z. Ge, C. Jia, Q. Wang, L. Li, S. Wang, S. Ma, An improved entropy coding algorithm for nic, Technical Report VC-37-M320 (2022)
187. S. Esenlik, Y. Li, Y. Wu, K. Zhang, L. Zhang, Z. Zhang, Response to the call for proposals on neural network-based image coding by team czjpg5, Technical Report VC-38-M325 (2022)
188. H. Ma, C. Dong, H. Wang, H. Zhang, Z. Zhang, H. Zhu, D. Ding, L. Li, X. Xu, S. Liu, Z. Chen, D. Liu, Response to the call for proposals on neural network-based image coding by team dwoteo, Technical Report VC-38-M326 (2022)
189. I.W. Subgroup, Notes of the 4th IEEE 1857.11 meeting, Technical Report (2022)
190. S. Li, D. Wang, C. Jia, J. Liu, Extension of bd-rate calculation function for nic evaluation, Technical Report VC-39-M354 (2022)
191. S. Li, D. Wang, C. Jia, J. Liu, Post-processing for nic, Technical Report VC-39-M355 (2022)
192. L. Dai, J. Lin, D. Jiang, C. Nian, X. Shi, Uneven group convolution for nic, Technical Report VC-40-M358 (2022)
193. X. Shi, J. Lin, D. Jiang, C. Nian, L. Dai, Context model based on prediction for nic, Technical Report VC-40-M360 (2022)

194. Q. Li, L. Li, Add byte alignment function into 1857.11, Technical Report VC-41-M362 (2023)
195. X. Shi, J. Lin, D. Jiang, C. Nian, L. Dai, X. Shi, Enhanced block based on feedback mechanism, Technical Report VC-41-M363 (2023)
196. L. Dai, J. Lin, D. Jiang, C. Nian, X. Shi, H. Jin, Unevenly channel group and multi-scale mask convolution for context model, Technical Report VC-41-M364 (2023)
197. B. Kang, H. Zhang, C. Dong, L. Li, D. Liu, Match the reference software with wd: decoupling parsing and reconstruction process in iwave, Technical Report VC-42-M367 (2023)
198. Y. Jiang, H. Zhang, C. Dong, L. Li, D. Liu, Entropy decoding acceleration for iwave, Technical Report VC-42-M368 (2023)
199. K. Ma, Y. Wu, S. Esenlik, Z. Zhang, L. Zhang, K. Zhang, Combination of iwave's wavelet-like transformation with bee, Technical Report VC-42-M369 (2023)
200. X. Shi, J. Lin, D. Jiang, C. Nian, L. Dai, X. Shi, Context model based on prediction for nic, Technical Report VC-42-M370 (2023)
201. L. Dai, J. Lin, D. Jiang, C. Nian, X. Shi, H. Jin, Unevenly channel group for context model, Technical Report VC-42-M371 (2023)
202. L. Dai, J. Lin, D. Jiang, C. Nian, X. Shi, H. Jin, Multi-scale mask convolution for context model, Technical Report VC-42-M372 (2023)
203. X. Shi, J. Lin, D. Jiang, C. Nian, L. Dai, X. Shi, Enhanced block based on feedback mechanism, Technical Report VC-42-M373 (2023)
204. S. Li, Z. Fan, D. Wang, C. Jia, J. Liu, Filtering technologies, Technical Report VC-42-M374 (2023)
205. X. Wang, T. Chen, M. Lu, Z. Ma, F. Chen, Context model acceleration for nic: Multistage context model and entropy engine optimization, Technical Report VC-43-M378 (2023)
206. L. Dai, C. Nian, D. Jiang, J. Lin, K. Fu, P. Zhang, X. Zhang, D. Cheng, W. Huang, F. Ye, Unevenly channel group for context mode, Technical Report VC-43-M379 (2023)
207. X. Shi, C. Nian, D. Jiang, J. Lin, R. Wu, W. Huang, Y. Mao, X. Shi, Enhanced block based on feedback mechanism, Technical Report VC-43-M381 (2023)
208. K. Ma, Y. Wu, S. Esenlik, Z. Zhang, L. Zhang, K. Zhang, Time consumption of the combination model of bee and iwave, Technical Report VC-43-M382 (2023)
209. X. Zhang, Z. Zhang, Y. Wu, S. Esenlik, K. Zhang, L. Zhang, Combination of nic analysis and synthesis transforms with bee, Technical Report VC-44-M385 (2023)
210. Y. Kong, X. Wang, T. Chen, M. Lu, Z. Ma, F. Chen, Improved multistage context model, Technical Report VC-44-M386 (2023)
211. X. Hang, W. Duan, C. Jia, S. Ma, A context model quantization scheme for nic, Technical Report VC-44-M388 (2023)
212. Z. Ge, C. Jia, S. Ma, W. Gao, An improved entropy coding algorithm for nic, Technical Report VC-44-M389 (2023)
213. X. Wu, Z. Zhang, Y. Chen, Q. Yuan, Y. Feng, T. Fan, Parallel context model for bee, Technical Report VC-44-M390 (2023)
214. Y. Chen, Q. Yuan, Z. Zhang, X. Wu, Y. Feng, Local aware attention for bee, Technical Report VC-44-M391 (2023)
215. H. Zhang, F. Mei, L. Li, D. Liu, Entropy codec engine convergence, Technical Report VC-44-M392 (2023)
216. Y. Jiang, H. Zhang, C. Dong, L. Li, D. Liu, Entropy coding acceleration for iwave, Technical Report VC-44-M394 (2023)
217. X. Hang, C. Jia, S. Ma, W. Gao, Quantized inference for nic, Technical Report VC-45-M400 (2023)
218. Z. Ge, C. Jia, S. Ma, W. Gao, An improved entropy coding algorithm for nic, Technical Report VC-45-M401 (2023)
219. T. Zhang, Z. Zuo, H. Zhang, J. Liao, F. Mei, L. Li, D. Liu, Entropy codec engine convergence, Technical Report VC-46-M407 (2023)
220. Y. Xie, B. Kang, Q. Chen, C. Gao, L. Li, D. Liu, Non uniform quantization between different subbands in iwave, Technical Report VC-46-M408 (2023)
221. X. Hang, C. Jia, S. Ma, Quantized inference for nic, Technical Report VC-46-M409 (2023)

222. Z. Ge, C. Jia, S. Ma, Sparsity-bypassing based entropy coding optimization for nic, Technical Report VC-46-M410 (2023)
223. X. Liang, H. Kuang, Y. Jiang, L. Li, J. Liu, D. Liu, Entropy coding acceleration for iwave on lossy mode, Technical Report VC-46-M411 (2023)
224. H. Kuang, Y. Jiang, X. Liang, L. Li, J. Liu, D. Liu, Further entropy coding acceleration for iwave on lossless mode, Technical Report VC-46-M412 (2023)
225. H. Kuang, Z. Fan, Y. Ma, J. Zhang, C. Jia, J. Liu, New design codec process with slice partition for iwave, Technical Report VC-46-M413 (2023)
226. X. Hang, C. Jia, S. Ma, Numpy support for rans engine, Technical Report VC-47-M422 (2023)
227. X. Wu, Y. Chen, Q. Yuan, Z. Zhang, Y. Feng, T. Fan, Efficient parallel context model for bee, Technical Report VC-47-M423 (2023)
228. H. Kuang, Y. Jiang, X. Liang, L. Li, J. Liu, D. Liu, Entropy coding acceleration for iwave on lossless mode, Technical Report VC-47-M424 (2023)
229. K. Ma, Y. Wu, S. Esenlik, Z. Zhang, L. Zhang, K. Zhang, Combination of iwave's wavelet-like transformation with bee, Technical Report VC-47-M427 (2023)
230. X. Shi, C. Nian, D. Jiang, J. Lin, Alignment for latent representation quantization in the training and test process, Technical Report VC-47-M428 (2023)
231. Z. Fan, H. Kuang, Y. Ma, J. Zhang, C. Jia, J. Liu, Lightweight optimization for iwave filtering technology, Technical Report VC-48-M433 (2023)
232. X. Wu, Y. Chen, Q. Yuan, Z. Zhang, Y. Feng, T. Fan, Optimization of efficient parallel context model for bee, Technical Report VC-48-M434 (2023)
233. H. Kuang, X. Liang, Y. Jiang, L. Li, D. Liu, J. Liu, Entropy coding acceleration for iwave on lossy mode, Technical Report VC-48-M435 (2023)
234. X. Shi, C. Nian, D. Jiang, J. Lin, Overlap tile for removing boundary effect, Technical Report VC-48-M436 (2023)
235. W. Gao, G. Li, *Deep Learning for 3D Point Clouds* (Springer Nature, 2024)
236. W. Gao, G. Li, Introduction to 3d point clouds: Datasets and perception, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 1–27
237. W. Gao, G. Li, Learning basics for 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 29–70
238. W. Gao, G. Li, Deep-learning-based point cloud enhancement i, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 71–97
239. W. Gao, G. Li, Deep-learning-based point cloud enhancement ii, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 99–130
240. W. Gao, G. Li, Deep-learning-based point cloud analysis i, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 131–162
241. W. Gao, G. Li, Deep-learning-based point cloud analysis ii, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 163–193
242. W. Gao, G. Li, Point cloud pre-trained models and large models, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 195–225
243. W. Gao, G. Li, Point cloud-language multi-modal learning, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 227–254
244. W. Gao, G. Li, Open-source projects for 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 255–272
245. W. Gao, G. Li, Typical engineering applications of 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2025), pp. 273–299
246. W. Gao, G. Li, Future work on deep learning-based point cloud technologies, in *Deep Learning for 3D Point Clouds* (Springer, 2025), pp. 301–315
247. L. Xie, M. Hu, X. Bai, Towards hardware-friendly and robust facial landmark detection method, in *International Conference on Neural Information Processing* (Springer, 2022), pp. 432–444
248. L. Xie, M. Hu, X. Bai, Online Improved Vehicle Tracking Accuracy via Unsupervised Route Generation, in *IEEE 34th International Conference on Tools with Artificial Intelligence* (IEEE, 2022), pp. 788–792

249. L. Zhang, L. Zhang, X. Mou, D. Zhang, Fsim: a feature similarity index for image quality assessment. *IEEE Trans. Image Process.* **20**(8), 2378–2386 (2011)
250. Z. Wang, Q. Li, Information content weighting for perceptual image quality assessment. *IEEE Trans. Image Process.* **20**(5), 1185–1198 (2011)
251. Z. Wang, E. Simoncelli, A. Bovik, Multiscale structural similarity for image quality assessment, in *The Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, vol. 2 (2003), pp. 1398–1402
252. V. Laparra, J. Ballé, A. Berardino, E.P. Simoncelii, Perceptual image quality assessment using a normalized laplacian pyramid, in *Human Vision and Electronic Imaging, HVEI 2016. Society for Imaging Science and Technology*, pp. 43–48 (2016)
253. N.N. Ponomarenko, F. Silvestri, K.O. Egiazarian, M. Carli, J. Astola, V.V. Lukin, On between-coefficient contrast masking of dct basis functions (2007). [Online]. Available: <https://api.semanticscholar.org/CorpusID:7057277>
254. H. Sheikh, A. Bovik, Image information and visual quality. *IEEE Trans. Image Process.* **15**(2), 430–444 (2006)
255. S. Winkler, Perceptual video quality metrics—a review. *Dig. Video Image Quality Perceptual Coding* 155–180 (2017)
256. Y. Lee, Response to vcm cfp: Video coding with machine-attention, Technical Report M61460 (2023)
257. S. Rózek, Poznań university of technology proposals a and b in response to cfp on video coding for machines, Technical Report M61519 (2023)
258. S. Rózek, Poznań university of technology proposal c in response to cfp on video coding for machines, Technical Report M61520 (2023)
259. H. Zhang, N. Le, J. Ahonen, A. Hallapuro, A. Aminlou, L. Xu, R.G. Youvalari, F. Cricri, H.R. Tavakoli, E. Aksu, M. Hannuksela, Response to the cfp of the vcm by nokia, Technical Report M60753 (2023)
260. K. Jia, L. Yu, H. Yu, A. Liu, J. Liu, Response to the cfp on video coding for machine from zhejiang university, Technical Report M60741 (2023)
261. H. Wang, J. Xue, Y. Zhang, W. Ji, J. Liu, Response to cfp on video coding for machines from china telecom, Technical Report M60775 (2023)
262. A. Kim, E.-V. An, K. Deok Seo, S. Heung Jung, W.-S. Cheong, J. Y. Lee, H.-G. Choo, Adaptive spatial resampling based on mi-rpr for vcm, Technical Report M64124 (2023)
263. MPEG, Technologies under consideration for video coding for machines, Technical Report ISO/IEC JTC 1/SC 29/WG 4 WG04N0336 (2023)

Chapter 10

Implementations for AI-based Image and Video Coding



10.1 Introduction

As artificial intelligence rapidly advances, neural networks have shown strong capabilities in areas like image recognition and natural language processing [1–46]. However, as the scale and complexity of models continue to increase, so does the demand for computational resources, posing challenges to deployment and execution. In the field of image coding, learning-based neural network algorithms also face deployment difficulties [47–91]. In many practical application scenarios, such as edge devices, mobile devices, and Internet of Things (IoT) devices, resource-constrained environments limit the application of large-scale neural networks [21, 92–143].

Neural network light weight design has become a crucial means to address this challenge. The core objective of light weight design is to reduce the model's parameter count, computation complexity, and memory footprint while maintaining performance. This not only facilitates efficient deployment of models on resource-constrained devices but also improves the model's response speed in real-time applications. Light weight design techniques mainly include pruning, quantization, distillation, and Neural Architecture Search (NAS). Pruning streamlines the structure of neural networks by removing connections or neurons that contribute minimally to the final predictions, thereby reducing the model's parameter size. Quantization reduces the number of bits used to represent neural network weights and activation values, converting them from floating-point numbers to lower bit-width integer representations. By lowering the precision of numerical representations, it significantly decreases the model's storage requirements and computational overhead, thus improving real-time performance on embedded systems or mobile devices. Distillation achieves light weight design by transferring knowledge from a large, complex model to a smaller, simplified one. NAS, meanwhile, automates the design of optimal neural network architectures by searching for models that are efficient under specific constraints, such as computational power or memory limits. This technique has proven effective in discovering lightweight models that retain high performance while

using fewer resources. Recently, researchers have developed various lightweight methods tailored for learning-based image compression networks, which will be introduced below, to preserve visual quality [91, 144–152] while reducing model size and computation time.

10.2 Basics of Neural Network Compression

10.2.1 Network Pruning

Network pruning utilizes both the redundancy of neural network parameters and structures for simplifying the deep learning model [153], and the most basic idea is to prune the least important parts [154]. This method results in a model with fewer parameters and a simpler structure, without compromising task performance. Due to the smaller computational resource and memory requirements of the compressed model, it can meet a wider range of application requirements compared to the original model. Deep learning network models contain a significant number of redundant parameters, from the convolutional layers to the fully connected layers, with many neuron activation values approaching zero. By removing these neurons, the model can still maintain the same representation capability. This process can be named as over-parameterization [155], and the involved method is dubbed as deep learning model pruning. The basic idea of network pruning is to trim the least important parts. There are many benefits to simplifying the model, including but not limited to:

- The most important and direct benefit is that it reduces the computation complexity and parameter quantity of the model, reduces computational time and power consumption.
- Due to the reduction in parameter count, many large networks can run on end-to-end embedded devices. This method can no longer rely on cloud computing and network transmission of results, and can actually display the results [141].
- Expanding the application scenarios of neural network models can enable them to be applied in various computationally limited environments, such as autonomous driving and handheld devices [156].

Due to the varying degrees of model redundancy at different granularity levels, many different pruning methods have emerged.

Types of Network Pruning. Pruning methods can be classified into four types based on granularity, including fine-grained, vector-level, kernel-level, and filter-level [157, 158]. The fine-grained, vector-level, and kernel-level pruning methods strike a balance between the number of parameters and the model performance, but the topology of the network itself has changed, requiring specialized algorithm design to support this sparse operation. These three methods are also known as unstructured pruning. Filter-level pruning only reduces the number of filters and feature channels in the network, allowing the pruned model to run without requiring

specialized algorithm design, a technique known as structured pruning. There are also methods that prune entire network layers, which can be viewed as a generalized form of filter-level pruning that removes all filters.

- **Vector-level Sparsity.** This method, with a larger particle size than fine-grained pruning, focuses on pruning within the convolution kernel.
- **Kernel-level Sparsity.** This method removes an entire convolutional kernel, which eliminates the corresponding response to the computational channel in the input.
- **Filter-level Sparsity.** This method prunes an entire group of convolutional kernels, potentially changing the number of output feature channels during inference.

Methods. Model pruning involves multiple research aspects, such as how to determine the unimportance of model parameters, the pruning rate, and how to restore the performance of the pruned model. This part will focus on exploring the research work of these parts below.

A straightforward approach to pruning involves determining the importance of model weights and subsequently ranking them based on this importance. Wen et al. [159] used the absolute value of weights as a measure of importance. They added regularization terms to the loss function during training to make the weights sparse after training. Liu et al. [160] introduced channel-level scaling factors into the batch-normalization model for neural networks equipped with batch-normalization layers and incorporated regularization terms into the loss function. This method ensures the sparsity of weight parameters, and then they prune the corresponding partial weights with small scale factor values. Other methods determine the importance of pruned parameters based on their impact on the final performance. Based on the Taylor expansion of the neural network algorithm, Molchanov et al. [153] used the absolute value of the first-order term in the expansion of the activation value as the evaluation criterion for pruning. This method avoids calculating the Hessian matrix and is faster.

Another research focus is how to determine the proportion of pruning. The proportion of pruning can be manually set. The manually set pruning ratio is difficult to achieve a balance between performance and computation complexity, therefore a method for neural networks to automatically determine the pruning ratio has emerged. Tan et al. [161] considered the resolution of the model, the depth of the model, and the number of channels for outputting features together. They use α , β , and γ to composite constraints on three different parameters, thereby reducing the computation complexity of the model. α , β , and γ are calculated manually. Han et al. [162] analyzed the sensitivity of different network layers. They found that the fully connected layer is more sensitive to pruning, so they used different predefined pruning ratios for different network layers. The method of automatically determining the pruning ratio requires the automatic determination of the pruning ratio, often using methods such as network structure search and reinforcement learning. Haakak et al. [163] used reinforcement learning to determine the pruning ratio of different network layers. Different pruning ratios are determined based on different purposes, such as ensuring accuracy or reducing computation complexity.

How to restore the accuracy of the pruned model is also a topic of concern for researchers. He et al. [154] directly minimized the difference in reconstructed features when performing filter pruning. For pruned models, most studies use fine-tuning to restore the performance of the pruned model. Zhu et al. [164] proposed a progressive pruning algorithm. This algorithm prunes faster at the beginning when there is more redundancy, gradually increasing from the initial pruning ratio to the target pruning ratio.

10.2.2 Quantization

Quantization is to approximate the weight or activation value represented by the high bit width (Float32) with the lower bit width (INT8, INT4, ...) [165]. Its numerical manifestation is to discretize continuous values. The quantization process can be represented by the following formula:

$$q = \text{round} \left(\frac{r}{S} + Z \right), \quad (10.1)$$

where q represents the quantized value and r represents the full precision value. S refers to the quantization scale range. Z represents the zero point of quantization. The process of inverse quantization of integer values into real values can be formalized as follows:

$$r = S(q - Z), \quad (10.2)$$

where Z and S can be formalized as the following formula:

$$S = \frac{r_{\max} - r_{\min}}{q_{\max} - q_{\min}}, \quad (10.3)$$

$$Z = \text{round} \left(q_{\max} - \frac{r_{\max}}{S} \right). \quad (10.4)$$

In this way, the floating-point calculation process has been successfully transformed into an integer calculation process. The calculations during the training process are all floating-point operations. Therefore, the training process can correct quantization errors. The calculation of the inference process is all integer operations, and hence the inference process can be accelerated. Compared with models with floating point weights, the quantized model has many advantages, as shown in the Table 10.1.

Types of Model Quantization. According to the uniformity of quantization, it can be divided into uniform quantization (linear quantization) and non-uniform quantization (nonlinear quantization) [157, 166, 167]. Uniform quantization refers to the use of a uniform width across different data ranges. Non-uniform quantization refers to the use of different widths across different data ranges.

Table 10.1 Comparison between the original model and the quantized model

Model	Parameters	Flops	Memory usage	Performance
Normal network	Numerous	Numerous	High	High
Quantized network	Low	Low	Low	Near high precision

Source Author

The quantization bit width can be any width, such as 8-bit, 4-bit, 2-bit, 1-bit. According to the granularity of quantization parameters, it can be divided into per channel quantization and per layer quantization [168]:

- Per channel quantization means that each channel of the weight uses a separate quantization parameter.
- Per layer quantization represents the use of independent quantization parameters for each layer in convolutional or fully connected layers.

Methods. There are also many research directions in the task of model quantization [167], including how to determine the mapping relationship between integers and float numbers, how to determine the quantization parameters, and so on. The discussion of these topics follows below:

- How to determine the mapping relationship between integers and float numbers, like rounding, unbiased estimation, clustering quantization, random sharing of quantization values based on neural networks, and so on.
- Ultra low precision quantization, like 1-bit quantization method, 2-bit quantization method [169].
- How to determine the bit width of each layer in mixed precision quantization, like reinforcement learning, network structure search, and so on.

If there are a small number of outliers in the quantified parameters, it will seriously affect the range of quantized data. A large number of parameters will be quantified to the quantization boundary, which will cause significant performance loss. To address this issue, Jacob et al. [170] designed a hyperparameter to adjust the pre-quantized data range. This method uses the truncation parameters of the activation function as training parameters to adaptively adjust and ensure the uniformity of the quantized data range. On the basis of this work, Esser et al. [171] also added scale factors to the training parameters of the model. In addition, they also indicate that the values of the quantization boundary will undergo sudden changes, so the corresponding partial derivatives should also undergo sudden changes in order to be more reasonable. In addition, they also added constraints and adjusted the gradient amplitude of the scale factor to be closer to the gradient amplitude of the weights, making the training process more stable. Bhalgat et al. [169] proposed using scale factors and zeros as training parameters in asymmetric quantization for automatic optimization. This method achieves good performance in asymmetric quantization.

10.2.3 Knowledge Distillation

Knowledge distillation is an efficient model compression technique that transfers knowledge from a large and complex model (also known as the teacher) to a smaller but simpler model (also known as the student) [172, 173]. The student model is trained to replicate the teacher's output, enabling it to learn complex representations with fewer parameters. The resultant smaller models not only require less computational resources, making them ideal for deployment in environments with limited capacity, but they also tend to generalize better due to the richer information contained in the soft probabilities of the teacher's output. Additionally, this technique is versatile and not tied to specific model architectures, offering flexibility in optimizing models for various operational constraints while maintaining high performance levels.

The knowledge distillation framework consists of two models [174], i.e., a large pre-trained teacher model and a smaller student model. Both models take the same input, and each processes it through several layers to produce logits (output predictions). The teacher model's logits serve as a reference for the student model through a distillation loss, which helps the student learn from the teacher's predictions. Additionally, the student model is trained using a classification loss, which compares its output to the ground truth labels. This dual loss approach—combining distillation loss and classification loss—enables the student model to mimic the teacher's performance while being more efficient in terms of computational resources.

Prediction Probabilities Between Teacher and Student. Hinton et al. [175] introduced a methodology wherein an expansive teacher model is initially trained, and its predicted soft labels are employed in lieu of hard labels for training the network. This approach guides the student model by using the Teacher's soft labels, with the option to elevate the softmax temperature during distillation, making the labels more diffuse and informative for training the Student model.

Matching Intermediate Weights. Romero et al. [176] introduced a novel approach for learning the weights of intermediate layers within neural networks. Deviating from the conventional usage of cross-entropy distillation loss, their method employed an L2 loss function to evaluate the discrepancies between the weights of the teacher and student models. Additionally, a linear transformation is applied to the student's intermediate layers to match the dimensionality of the teacher's corresponding layers.

Intermediate Features. Huang et al. [177] advocated for a knowledge distillation approach wherein alignment extends beyond the output probability distributions to include the feature distributions of intermediate layers between the teacher and student models. They utilized cosine similarity as a metric to ensure maximal alignment of the feature maps from both models. Subsequently, Kim et al. [178] furthered this line of research by employing an L2 loss function to quantify and minimize the differences in the latent features of the teacher and student models. In this methodology, the teacher's feature map is processed through an AutoEncoder structure, optimized to capture meaningful latent representations, which the student model then attempts to replicate using a Multi-Layer Perceptron (MLP).

Gradient. Zagoruyko et al. [179] introduced the concept of employing the gradients of CNN feature maps as a form of attention mechanism, positing that larger gradients signify a heightened sensitivity of the output to minor perturbations. The approach involves minimizing the L2 loss between the gradients of intermediate layer feature maps from the student and teacher models. This method postulates that feature map attentions from well-performing models exhibit high similarity, in contrast to those from models with inferior performance, which display significant divergence.

Sparsity Patterns. Heo et al. [180] advanced the hypothesis that the teacher and student models should demonstrate parallel sparsity after being subjected to the ReLU activation function. To achieve this, the study introduces a minimization objective for the optimization model, defined by the loss function $L(I) = |P(T(I)) - P(S(I))|$, where T symbolizes the teacher model, S the student model, and P the indicator function that quantifies sparsity.

10.2.4 Neural Architecture Search

NAS has evolved significantly since its inception, transforming from an experimental idea to a cornerstone in the development of efficient deep learning models. Traditional models like ResNet and MobileNet, which were manually designed by experts, marked the early stages of deep learning. In contrast, the advent of NAS signals a paradigm shift towards automated model design. NAS endeavors to automate the architectural discovery process, seeking to identify the most effective network structures for specific tasks. Its goals typically involve maximizing accuracy while optimizing computational efficiency, which is particularly crucial in model compression, where the challenge lies in reducing model size without sacrificing performance [47, 181, 182]. Elsken et al. [183] identified three core aspects of NAS:

- **Search Space:** This includes all possible architectures that NAS can explore. It's a comprehensive set of network designs, varying in depth, width, and types of layers.
- **Search Strategy:** The methodologies employed to traverse the search space. This could range from systematic approaches like grid search to more stochastic methods like random search.
- **Evaluation Criteria:** The metrics and benchmarks used to assess the performance of the architectures found. This often includes accuracy, efficiency, and latency.

10.3 Software and Hardware Platforms for Acceleration

10.3.1 *Software and Hardware Co-design*

Deep Neural Networks (DNNs) are a key component of machine learning, leading to major progress in fields like image and speech recognition, natural language processing, and autonomous systems. Their development has been closely tied to improvements in computing hardware and software. Initially, DNNs primarily operated on Central Processing Units (CPUs), the general-purpose workhorses of computing. However, as DNNs expanded in depth and width, the computational limitations of CPUs became apparent, especially in tasks requiring extensive parallel processing [184], a key characteristic of deep learning algorithms.

To address these challenges, the use of Graphics Processing Units (GPUs) became more prevalent. GPUs, with their superior parallel processing capabilities, are better suited for handling the large-scale matrix and vector operations common in deep learning [62, 185].

Despite the advantages offered by GPUs, the escalating requirements of DNNs, coupled with the slowing pace of Moore's Law, present new challenges. Energy efficiency and sustainability have become major concerns; as models increase in size, so does their energy consumption, leading to higher costs and environmental implications. Additionally, real-time processing and edge computing introduce further challenges and opportunities. Many deep learning applications, such as autonomous driving, robotics, and real-time 3D data processing, require rapid, on-the-spot data processing. This shift towards edge computing necessitates more efficient hardware support to minimize latency and bandwidth demands [186, 187].

The need to balance efficiency and flexibility has led to the development of specialized hardware, such as Application-Specific Integrated Circuits (ASICs) and Field-Programmable Gate Arrays (FPGAs), optimized for deep learning tasks. These specialized processors offer improved performance and energy efficiency for specific types of neural network computations.

Moreover, software optimizations have played a crucial role in complementing hardware advancements. Frameworks like TensorFlow [188] and PyTorch [189] have evolved to better leverage the capabilities of both general-purpose and specialized hardware, further enhancing the performance and scalability of deep learning models.

In summary, cost-effectiveness is a critical driver in this landscape. As deep learning applications become more widespread, reducing the costs associated with training and deploying models is increasingly important. This has spurred research into more economical hardware and software solutions, aiming to make deep learning technologies more accessible and feasible.

The computational intensity of DNNs necessitates the use of specialized acceleration platforms. This section delves into the current landscape of hardware and software platforms designed to accelerate DNNs.

10.3.2 Graphics Processing Units (GPUs)

Graphics Processing Units (GPUs), which is initially developed for computer graphics, have remarkably extended their utility into broader computing domains, particularly deep learning. Their architecture, based on a Single Instruction, Multiple Threads (SIMT) paradigm, efficiently orchestrates numerous computational cores, significantly enhancing data throughput with dedicated Video Random Access Memory (VRAM).

The core principle of GPU operation lies in its parallel processing architecture. Unlike CPUs, which have fewer cores optimized for sequential tasks, GPUs feature thousands of smaller cores designed for parallel processing. This design is ideal for both graphics rendering and deep learning computations. The SIMT approach enables multiple processing threads to execute the same instruction on different data points, excelling in parallelizable operations like matrix multiplications.

In terms of memory, GPUs are outfitted with high-bandwidth memory (HBM) or Graphics Double Data Rate (GDDR), facilitating faster data transfer rates that are essential for managing large datasets and complex algorithms. NVIDIA's CUDA platform and compute shaders across various GPU brands further harness this computational power, allowing for intricate programming and execution on the GPU.

The rise of deep learning frameworks including Caffe [190], PyTorch [189], and TensorFlow [188] has simplified the construction and execution of neural networks on GPUs. This ease of use has solidified the role of GPUs as a cornerstone in deep learning. As the demand for computational power in deep learning continues to grow, GPU manufacturers are increasingly tailoring their products towards neural network-specific enhancements.

A notable example is NVIDIA's introduction of the cuDNN library [191], a dedicated tool for deep neural network acceleration that unlocks the full parallel potential of GPUs. However, GPUs face challenges in handling tasks with irregular computation and memory access patterns, such as pruned neural networks. In these scenarios, the gap between their theoretical and actual performance becomes evident.

The NVIDIA V100 GPU [192], based on the Volta architecture, exemplifies this evolution. It features Tensor Cores, specialized hardware for performing tensor operations that are crucial in deep learning, significantly accelerating matrix operations at the core of neural network training and inference.

Power consumption remains a critical consideration. As advancements in chip fabrication slow down, the substantial power demands of GPUs, due to their extensive parallel processing capabilities, make them more suitable for environments like data centers and cloud infrastructures. High-end models, such as NVIDIA's V100 and A100 [193], illustrate this, with thermal power reaching up to 300 and 400 W, respectively. This aspect underscores the ongoing challenges of balancing computational power with energy efficiency in GPU design.

10.3.3 *Application-Specific Integrated Circuits (ASICs)*

Application-Specific Integrated Circuits (ASICs) have carved a niche in the acceleration of deep neural networks (DNNs), offering a unique blend of specialized efficiency and performance that distinguishes them from general-purpose processors like CPUs and GPUs. ASICs are custom-designed with a dedicated data flow architecture specifically tailored for neural network applications, enabling them to handle the repetitive and parallel nature of DNN tasks with remarkable efficiency.

The core advantage of ASICs lies in their ability to deliver significant performance and energy efficiency benefits. In certain scenarios, they can outperform general-purpose processors by several times, even achieving performance improvements of up to an order of magnitude. This makes ASICs particularly suitable for Convolutional Neural Network (CNN) hardware acceleration and a prime choice for large-scale commercial applications. Notable examples of ASICs designed for deep learning computation include the DianNao series [194–196], the Cambricon series [197–199], and Google’s Tensor Processing Unit (TPU) series [200, 201]. These processors exemplify the potential of ASICs in large-scale machine learning tasks.

Further advancements in ASIC technology have been achieved through specialized designs that incorporate techniques such as model bit-width quantization and sparsification. Innovations like Bit Fusion [202], Thinker-II [203], and SparTen [204] have pushed the boundaries of hardware performance and energy efficiency. These developments underscore the ongoing evolution and specialization of ASICs in the realm of deep learning.

However, the specialized nature of ASICs brings inherent limitations. Their design, tailored for specific tasks, means they lack the flexibility found in CPUs and GPUs. Once fabricated, their hardware circuits cannot be altered, limiting their adaptability to evolving neural networks or new algorithms. This inflexibility is compounded by the long design cycle (typically one to two years) and high costs associated with ASIC development. Designers often need to anticipate task requirements years in advance, a challenging prospect given the rapid pace of development in neural network applications.

In conclusion, ASICs play a crucial role in the field of DNN acceleration, providing task-specific efficiency and performance advantages. Their tailored architecture and energy-efficient design make them particularly suitable for specialized applications where power efficiency and high-throughput, low-latency processing are critical. However, their limitations regarding adaptability, design cycle length, and cost must be carefully considered. Due to the developing deep learning theories and methods, the importance of ASICs is likely to increase, but their role will be shaped by the balance between their specialized capabilities and the need for flexibility and adaptability in a rapidly advancing technological landscape.

10.3.4 *Field-Programmable Gate Arrays (FPGAs)*

Field-Programmable Gate Arrays (FPGAs) have emerged as a versatile and powerful tool in the acceleration of deep neural networks (DNNs), distinguished by their unique blend of adaptability and efficiency. As semi-custom circuits, FPGAs can be reprogrammed to alter their internal structure and functionality, setting them apart from the fixed architectures of ASICs and the general-purpose design of GPUs. This reconfigurability allows FPGAs to adapt to the diverse and rapidly evolving requirements of deep learning, making them particularly valuable in this field.

FPGAs integrate a variety of programmable logic units and interconnect resources, including programmable logic blocks, memory units, clock management modules, and embedded hard cores. Users can configure these internal circuits by loading bitstream files, granting a high degree of flexibility. Originally gaining traction as platforms for digital circuit design verification, FPGAs have since broadened their application scope, becoming increasingly prevalent in reconfigurable computing and deep learning acceleration.

One of the key strengths of FPGAs in DNN acceleration is their ability to offer circuit customization capabilities akin to ASICs. They can be tailored to create optimized circuits for various neural network computations, such as general convolution accelerators [205], quantized convolution accelerators [206], and sparse convolution accelerators [207]. These specialized configurations demonstrate FPGAs' significant performance and energy efficiency advantages, particularly in scenarios where they are optimized for specific tasks.

Despite these strengths, FPGAs traditionally faced limitations in computational, storage, and bandwidth resources, which initially confined their use primarily to embedded domains. However, the landscape has shifted with technological advancements in FPGAs. Manufacturers like Xilinx have introduced platforms such as the UltraScale+ series, designed for diverse applications ranging from edge to cloud scenarios. This expansion in capabilities has broadened the applicability of FPGAs, which makes them better suited for a broader range of deep learning applications.

Moreover, advancements in High-Level Synthesis (HLS) technology [208] have significantly lowered programming barriers, making FPGAs more accessible and reducing their development cycle. This progress is crucial in enabling FPGAs to keep pace with the rapid development of neural networks and other advanced applications.

In summary, FPGAs represent a dynamic and efficient solution for DNN acceleration, particularly in scenarios that demand high adaptability and power efficiency. Their ability to be reconfigured for specific deep learning tasks, coupled with recent advancements in technology and resources, makes them a valuable asset in the deep learning landscape. However, the complexities involved in their programming and the need for specialized knowledge to optimize their configurations are important considerations for their deployment. As deep learning continues to evolve, the role of FPGAs is likely to expand, especially in areas where their unique attributes of flexibility, efficiency, and adaptability are most beneficial.

10.3.5 AutoML Platforms

While hardware accelerators like GPUs, ASICs, and FPGAs have significantly boosted computational power in AI, the evolution of software, particularly Automatic Machine Learning (AutoML) platforms, marks a parallel advancement. The rise of AutoML platforms has marked a transformative shift in the field of artificial intelligence and machine learning. These platforms are ingeniously designed to automate critical aspects of the machine learning process, encompassing model selection, feature engineering, hyperparameter tuning, and even the optimization of neural network architectures. The primary objective of AutoML is to democratize access to advanced machine learning techniques. This makes these powerful tools more accessible to a wider audience.

AutoML platforms offer impressive efficiency and accessibility. They greatly simplify the model development process, cutting down on the time and expertise needed to create effective machine learning models. This democratization of AI technology not only accelerates the pace of innovation but also opens up new opportunities for applications across various sectors. Moreover, the optimization capabilities inherent in these platforms often result in models that are more efficient and performance than those manually tuned, thereby enhancing the overall effectiveness of machine learning solutions.

Among AutoML platforms, several have gained prominence due to their distinctive features and widespread adoption. Google AutoML, for example, is notable for its user-friendly interface and elegant integration with the Google Cloud services, offering specialized solutions for tasks in vision, language, and translation. Its appeal lies in enabling users to develop custom models with minimal technical intervention. H2O AutoML, an open-source platform, excels in automating the entire machine learning workflow, making it a robust choice for both rapid prototyping and complex model development. Amazon SageMaker [209] takes a holistic approach, offering a fully integrated service that simplifies the entire machine learning process, and is particularly effective in cloud-based environments.

In the open-source domain, Auto-Keras [210] and Auto-Sklearn [211] have made significant contributions. Auto-Keras, leveraging the Keras library, simplifies the neural network architecture search, a process traditionally requiring extensive expertise. Auto-Sklearn, built on Scikit-Learn, automates model selection and hyperparameter tuning, streamlining the machine learning pipeline. TPOT [212] introduces a unique approach using genetic algorithms to automate the design and optimization of machine learning pipelines, exploring a vast array of configurations to identify the most efficient solutions.

Kubeflow [213], a Kubernetes-native platform, is particularly adept at managing complex machine learning systems, making it a preferred solution for machine learning operations (MLOps) in distributed environments. Neural Network Intelligence (NNI) [214] by Microsoft offers a comprehensive toolkit for automating model tuning, architecture search, and model compression, known for its flexibility and wide range of supported optimization techniques.

In conclusion, AutoML platforms represent a significant leap forward in the field of machine learning. They offer a unique blend of efficiency, accessibility, and optimization, catering to a diverse range of needs and applications. From Google AutoML's intuitive interface to NNI's comprehensive optimization toolkit, these platforms are pivotal in empowering a broader spectrum of users to harness the power of machine learning. As the field continues to evolve, the role of AutoML platforms in facilitating and accelerating this growth becomes increasingly indispensable, heralding a new era of innovation and discovery in machine learning.

10.4 Lightweight Methods for Deep Compression Network

The previous methods of model compression were all aimed at high-level visual tasks, and compared to image compression networks, these types of tasks usually do not include information redundancy removal modules and feature reconstruction modules. The analysis encoder of image compression network removes redundant information of features and converts the image into a more compact latent representation. During this process, low-level feature information is retained to achieve better image reconstruction without extracting high-level features, which is different from networks used for advanced visual tasks. These types of features have a simple structure, high repeatability, and are more convenient for identifying ineffective filters in image compression networks through these features. However, these types of features are less helpful for tasks that require better reconstruction of image edges and texture details, so they usually do not appear in the results extracted by image compression networks, which poses difficulties in reducing the size of the network. In the process of image decoding [181, 182, 185, 215–220], the synthetic decoder network reconstructs the input image based on latent representations, which is an upsampling process that typically does not exist in advanced visual networks. Super prior networks can model the probability of potential features. In image compression networks, the optimization objective of analyzing the encoder network and the synthetic decoder network is to minimize image distortion, while the optimization objective of the hyper prior network is to minimize bit rate. The optimization objectives of the two parts are different, and the redundancy after joint optimization is also different. It is difficult to jointly compress the two parts of the network. Due to the above reasons, many studies have focused on designing lightweight network methods that meet the characteristics of image compression tasks. There are already many methods to reduce the parameter and computation complexity of image encoding networks, including pruning, quantization, and lightweight structures. This part will introduce them below.

10.4.1 Pruning for Deep Image Compression Network

Yang et al. [221] proposed a network structure for image compression with variable number of channels and variable bit rate. They reduced the consumption of computing resources at low bitrates by controlling the number of channels. Other variable bit rate image compression methods often do not change the original computation complexity, such as bottleneck scaling and feature modulation. However, this work reduces the number of parameters and computation complexity while changing the bit rate. The original autoencoder can be set to consist of multiple sub encoders. Each autoencoder can be represented as a parameter pair $\psi^{(k)} = (\theta^{(k)}, \phi^{(k)}) \in (\theta^{(1)}, \phi^{(1)}), \dots, (\theta^{(k)}, \phi^{(k)})$. Then they consider that $\theta^{(1)} \subset \dots \subset \theta^{(K)} = \theta$ and $\phi^{(1)} \subset \dots \subset \phi^{(K)} = \phi$, where θ and ϕ denote the parameters of encoder and decoder, respectively. During the training process, different parameter groups will undergo joint training to optimize different rate distortion performance. The loss function is defined in the following formula:

$$L(\psi) = \sum_{k=1}^K D(\psi^{(k)}) + \lambda^{(k)} R(\psi^{(k)}). \quad (10.5)$$

During the training process, joint optimization with different bit rates and computation complexity is carried out until optimal performance is achieved.

The joint control of rate and parameters of image compression network is a focus of many researchers. Guo et al. [222] proposed the CBAnet to simultaneously control the parameters and bit rate. However, unlike the above method, CBAnet does not interfere with the adjustment of parameters and bit rate. CBAnet proposed a bit adaptive module (BAM) to control the variation of bit rate. In the decoder network segment, a Complexity Adaptive Module (CAM) was constructed to achieve switching between different computational loads. CAM has multiple parallel branches, each containing different computational costs. In the reasoning process, different numbers of branches can be chosen to achieve various computational complexities.

BAM consists of lightweight depth-wise convolution and point-wise convolution modules. In the process of network inference, BAM does not increase computation complexity.

10.4.2 Dynamic Neural Network (DNN) for Image Compression

In recent years, Dynamic Neural Networks have been widely used in machine recognition tasks. Compared to conventional neural networks that keep structures and parameters fixed during inference, Dynamic Neural Networks can adaptively adjust networks' structure and parameters based on different inputs, allocating computational resources in a dynamic and reasonable way.

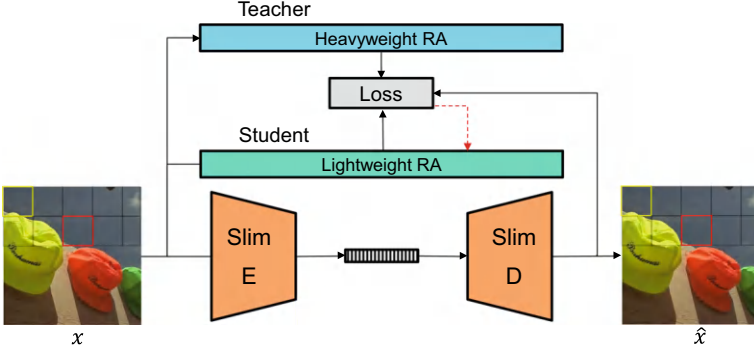


Fig. 10.1 Illustration of AdaNIC framework. ©2023 IEEE. Reprinted, with permission, from Ref. [49]

Tao et al. [49] introduced AdaNIC, an image compression network that uses spatially-adaptive convolution. This allows block-wise nonlinear transformations to allocate FLOPs unevenly across the input image. The maximum allowable rate-distortion penalty is controlled by a threshold ϵ :

$$L'_{max} = L'_{old} + \epsilon, \quad (10.6)$$

$$\epsilon = d\mu sec(\theta) = d\sqrt{1 + \mu^2}, \quad (10.7)$$

where L'_{max} is the maximum tolerance of coding loss, L'_{old} is the original tolerance of coding loss, μ is the balancing factor and θ is the angle of the R-D curve.

An lightweight yet effective Routing Agent is designed and serves as a widget to predict Transform Capacity Map (TCM) by approximating JUMC targets. The training of a dynamic routing agent is a two-stage learning pipeline. First, a heavyweight teacher model is trained to capture R-D responses of image compression model for various inputs. Then, a lightweight student model is participated to learn the behavior of the pre-trained teacher model, searching for the best capacity for different source image content. The whole architecture of AdaNIC is shown in Fig. 10.1.

10.4.3 Quantization for Deep Image Compression Network

Quantization is a method that is closely integrated with practical applications. In order to accelerate the quantified model, not only do all parameters of the model need to be integers, but also hardware is required to support acceleration of integer operations. Jia et al. [223] proposed a 4k neural network video coding system based on FPGA. In this work, they quantized the proposed image compression network using the quantization method proposed by Google [170] earlier, and then fine tuned the network to restore the performance. The quantized network is deployed on FPGA

and only performs integer operations during the inference process. Compared to the scenario deployed on NVIDIA 3080Ti, the quantized model deployed on FPGA can achieve the same compression effect with lower power consumption.

Sun et al. [224] proposed a non-linear memory-free codebook to quantize the network weights in groups. They set the bit budget to 8 bits, which can be represented as fixed-point 1, α , β , where the symbol occupies one bit, the integer occupies α bits, and the decimal occupies β bits. During the training process, α and β are set to 1 and 6 bits respectively to complete quantization. Each weight parameter is quantized to the n -th power of 2. The quantified model saves 75% of memory usage.

The whole quantization process [223] is shown in Fig. 10.2. The process begins with the Quantization of the model in its floating-point representation. The Calibration step follows, during which the model’s minimum and maximum values are determined to calculate the appropriate Scale and Zero-point parameters. The model is then subjected to the quantization of its Weights and Activation values, transforming them into their respective integer values (INT8 Model). This process is aided by Training Data, which is crucial for the Quantization-Aware Training (QAT) stage to ensure model accuracy. After the quantization steps, the model is integrated into Model Integration. This is followed by the DPU Configuration process, which configures the hardware platform interface for deployment onto the FPGA. Finally, the quantized model is loaded onto the FPGA for execution, completing the process from model training to FPGA deployment.

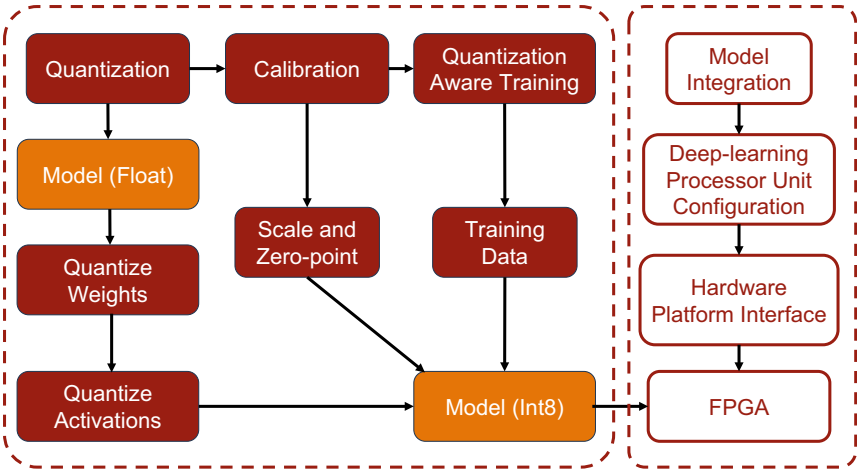


Fig. 10.2 The deployment process of the quantized model on FPGA. ©2022 IEEE. Reprinted, with permission, from Ref. [223]

10.5 Summary

Table 10.2 presents a comparison of GPUs, ASICs, and FPGAs in the context of deep learning acceleration, emphasizing their respective advantages and disadvantages. GPUs are versatile and widely supported by deep learning frameworks, offering high parallel processing capabilities suitable for a broad range of applications. However, they consume more power and may not be the most efficient for specialized tasks. ASICs provide exceptional performance and energy efficiency for specific applications due to their custom-designed nature, but they lack flexibility and involve high initial development costs and longer time to market. FPGAs offer a balance with reconfigurable hardware that can be optimized for particular tasks, potentially achieving lower latency and power efficiency, but they require specialized hardware design expertise and generally have lower throughput compared to GPUs.

The AutoML platforms in Table 10.3 offer diverse features catering to various user needs in automating machine learning tasks. Auto-Keras and Auto-Sklearn simplify model development by leveraging popular libraries (Keras and Scikit-Learn) and making advanced techniques accessible to non-experts. TPOT distinguishes itself by using genetic algorithms to optimize machine learning pipelines across numerous configurations. Kubeflow excels in managing complex, distributed machine learning systems within Kubernetes environments, ideal for MLOps practices. NNI provides a suite of tools for model tuning and architecture search, supporting a wide range

Table 10.2 Comparison of GPUs, ASICs, and FPGAs in deep learning accelerations

Technology	Advantages	Disadvantages	Typical architectures
GPU	• High parallel processing capability • Suitable for a wide range of applications • High data throughput • Well-supported by deep learning frameworks	• High power consumption • Less efficient for task-specific applications	• NVIDIA Tesla V100 • NVIDIA Tesla A100 • NVIDIA RTX Series • AMD Radeon Series
ASIC	• Highly efficient for specific tasks • Low power consumption • High performance for tailored applications	• Inflexible for different tasks • Long development cycle • High initial development cost	• Google TPU • DianNao Series • Eyeriss Series • Cambricon Series
FPGA	• Reconfigurable hardware • Efficient for specific tasks after configuration • Lower latency	• Requires hardware design expertise • Lower throughput compared to GPUs • Can be more power-efficient, but depends on the design	• Xilinx Virtex Series • Xilinx UltraScale+ Series • Intel Altera Stratix Series

Source Author

Table 10.3 Comprehensive overview of autoML platforms

AutoML platform	Key features and contributions
Auto-Keras	• Leverages the Keras library • Simplifies neural network architecture search • Makes the process more accessible to non-experts
Auto-Sklearn	• Built on Scikit-Learn • Automates the selection of models and optimization of hyperparameters • Simplifies and optimizes the end-to-end machine learning workflow
TPOT	• Uses genetic algorithms for automation • Automates design and optimization of machine learning pipelines • Explores a wide array of configurations
Kubeflow	• Kubernetes-native platform • Manages complex machine learning systems • Suitable for MLOps in distributed environments
NNI (Neural Network Intelligence)	• Developed by Microsoft • Offers tools for model tuning, architecture search, and model compression • Supports a wide range of optimization techniques
Google AutoML	• Intuitive and accessible interface design • Effortless interoperability with Google Cloud infrastructure • Specialized solutions for vision, language, and translation tasks
H2O AutoML	• Open-source platform • Automates the entire machine learning workflow • Suitable for rapid prototyping and complex model development
Amazon SageMaker	• Offers a fully integrated service • Simplifies the entire machine learning process • Effective in cloud-based environments

Source Author

of optimization techniques. Cloud-based solutions like Google AutoML and Amazon SageMaker offer user-friendly interfaces with seamless integration into their respective cloud ecosystems, specializing in tasks like vision, language, and providing fully integrated services. H2O AutoML, as an open-source platform, automates the entire machine learning workflow and suits both rapid prototyping and complex model development. The optimal choice among these platforms depends on specific requirements such as ease of use, level of automation, cloud integration, and customization needs.

This chapter introduces lightweight methods for image coding, focusing on model compression techniques such as pruning, quantization, distillation, and NAS. The examination reveals that, while these methods are effective in general deep learning applications, they face limitations in image coding tasks where maintaining high-quality reconstructions with reduced computational overhead is critical. Additionally, various deployment platforms, including GPUs, FPGAs, ASICs, and AutoML, are discussed, as they support the implementation of lightweight models. These platforms

enhance the efficiency of image coding by providing parallel processing, customizable hardware, and automation for model optimization. Finally, specialized acceleration methods developed specifically for image coding networks are presented, offering a more efficient balance between performance and resource usage.

Exercises

1. Please explain the difference between a CPU and a GPU in the context of DNN training. Why is a GPU generally preferred for this task?
2. Please discuss the role of ASICs in accelerating deep neural networks. Please provide an example of an ASIC designed for this purpose.
3. Please explain how FPGAs can be used for neural network acceleration. What are the advantages of using FPGAs over traditional CPUs and GPUs?
4. Please explain what network pruning is and how it contributes to model lightweighting.
5. Please explain the concept of quantization in neural networks. Why is it important for lightweight models?
6. Please explain how knowledge distillation helps in creating lightweight neural networks.
7. Please outline the trade-offs involved in using lightweight architectures like MobileNet or SqueezeNet.
8. Please discuss the impact of transfer learning on lightweight model deployment.
9. Please explain the significance of model compression techniques in lightweight neural networks.
10. Please describe the concept of NAS in the context of lightweight networks.

References

1. X. Zang, W. Gao, G. Li, H. Fang, C. Ban, Z. He, H. Sun, A baseline investigation: transformer-based cross-view baseline for text-based person search, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 7737–7746
2. G. Liao, W. Gao, G. Li, J. Wang, S. Kwong, Cross-collaborative fusion-encoder network for robust rgb-thermal salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **32**(11), 7646–7661 (2022)
3. W. Gao, G. Liao, S. Ma, G. Li, Y. Liang, W. Lin, Unified information fusion network for multi-modal rgb-d and rgb-t salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **32**(4), 2091–2106 (2021)
4. Y. Chen, S. Sun, G. Li, W. Gao, T.H. Li, Closing the gap between theory and practice during alternating optimization for gans. *IEEE Trans. Neural Netw. Learn. Syst.* (2023)
5. Y. Chen, C. Jin, G. Li, T.H. Li, W. Gao, Mitigating label noise in gans via enhanced spectral normalization. *IEEE Trans. Circuits Syst. Video Technol.* (2023)

6. X. Zang, G. Li, W. Gao, Multidirection and multiscale pyramid in transformer for video-based pedestrian retrieval. *IEEE Trans. Ind. Inform.* **18**(12), 8776–8785 (2022)
7. X. Zang, G. Li, W. Gao, X. Shu, Learning to disentangle scenes for person re-identification. *Image Vis. Comput.* **116**, 104330 (2021)
8. X. Zang, G. Li, W. Gao, X. Shu, Exploiting robust unsupervised video person re-identification. *IET Image Process.* **16**(3), 729–741 (2022)
9. Z. Yue, G. Li, W. Gao, Cross-level guided attention for human-object interaction detection, in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2023), pp. 284–289
10. S. Sun, J. Liu, H. Li, G. Liu, T.H. Li, W. Gao, Streamflow: streamlined multi-frame optical flow estimation for video sequences, in *Advances in Neural Information Processing Systems* (2024)
11. Z. Yao, W. Gao, Iterative saliency aggregation and assignment network for efficient salient object detection in optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* (2024)
12. G. Liao, W. Gao, Rethinking feature mining for light field salient object detection. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(10), 1–24 (2024)
13. B. Yang, W. Gao, G. Li, Focus and adjust: progressive refinement network for human object interaction detection, in *2022 26th International Conference on Pattern Recognition (ICPR)* (IEEE, 2022), pp. 2546–2552
14. G. Liao, W. Gao, Q. Jiang, R. Wang, G. Li, Mmnet: multi-stage and multi-scale fusion network for rgb-d salient object detection, in *Proceedings of the 28th ACM International Conference on Multimedia* (2020), pp. 2436–2444
15. W. Gao, S. Kwong, Y. Zhou, Y. Jia, J. Zhang, W. Wu, Multiscale phase congruency analysis for image edge visual saliency detection, in *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 1 (IEEE, 2016), pp. 75–80
16. W. Gao, S. Kwong, Y. Zhou, X. Wang, Phase congruency analysis of down-sampled and blurring images for foreground extraction, in *2015 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)* (IEEE, 2015), pp. 152–157
17. Y. Guo, W. Gao, G. Li, Interpretable task-inspired adaptive filter pruning for neural networks under multiple constraints. *Int. J. Comput. Vis.* 1–17 (2024)
18. W. Gao, Y. Guo, S. Ma, G. Li, S. Kwong, Efficient neural network compression inspired by compressive sensing. *IEEE Trans. Neural Netw. Learn. Syst.* (20220)
19. Y. Guo, W. Gao, Semantic-driven automatic filter pruning for neural networks, in *IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
20. L. Tao, W. Gao, Efficient channel pruning based on architecture alignment and probability model bypassing, in *Ieee International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2021), pp. 3232–3237
21. Y. Li, Z. Yu, X. Chen, S. Wang, H. Yuan, . Gao, Zero-shot quantization for large-kernels via shape-based distribution and diversity self-distillation, in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2025)
22. S. Fan, W. Gao, G. Li, Salient object detection for point clouds, in *European conference on computer vision* (Springer, 2022), pp. 1–19
23. S. Luo, W. Gao, A general framework for rotation invariant point cloud analysis, in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2024), pp. 3665–3669
24. X. Lu, W. Gao, Attentivenet: Detecting small objects for lidar point clouds by attending to important points, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
25. Z. Pan, N. Zhang, W. Gao, S. Liu, G. Li, Less is more: label recommendation for weakly supervised point cloud semantic segmentation. *Proc. AAAI Conf. Artif. Intell.* **38**(5), 4397–4405 (2024)
26. Z. Pan, G. Liu, W. Gao, T.H. Li, Epcontrast: effective point-level contrastive learning for large-scale point cloud understanding, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024), pp. 1–6

27. N. Zhang, Z. Pan, T. H. Li, W. Gao, G. Li, Improving graph representation for point cloud segmentation via attentive filtering, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 1244–1254
28. K. Wen, N. Zhang, G. Li, W. Gao, Mpvnn: multi-resolution point-voxel non-parametric network for 3d point cloud processing, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024), pp. 1–6
29. Z. Pan, N. Zhang, W. Gao, S. Liu, G. Li, Point cloud semantic segmentation with sparse and inhomogeneous annotations, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
30. R. Zhang, G. Li, W. Gao, S. Liu, A quantum-inspired framework in master-servant mode for large-scale multi-modal place recognition. *IEEE Trans. Intell. Transp. Syst.* (2024)
31. D. Yang, W. Gao, Pointchd: a point cloud benchmark for congenital heart disease classification and segmentation. *IEEE J. Biomed. Health Inform.* (2024)
32. S. Fan, W. Gao, G. Li, Point-mpp: point cloud self-supervised learning from masked position prediction. *IEEE Trans. Neural Netw. Learn. Syst.* (2024)
33. Z. Pan, W. Gao, G. Li, Distribution guidance network for weakly supervised point cloud semantic segmentation. *Adv. Neural Inf. Process. Syst.* (2024)
34. R. Zhang, G. Li, W. Gao, T.H. Li, Compoin: can complex-valued representation benefit point cloud place recognition? *IEEE Trans. Intell. Transp. Syst.* (2024)
35. W. Gao, G. Li, *Deep Learning for 3D Point Clouds* (Springer Nature, 2024)
36. W. Gao, G. Li, Introduction to 3d point clouds: datasets and perception, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 1–27
37. W. Gao, G. Li, Learning basics for 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 29–70
38. W. Gao, G. Li, Deep-learning-based point cloud enhancement i, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 71–97
39. W. Gao, G. Li, Deep-learning-based point cloud enhancement ii, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 99–130
40. W. Gao, G. Li, Deep-learning-based point cloud analysis i, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 131–162
41. W. Gao, G. Li, Deep-learning-based point cloud analysis ii, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 163–193
42. W. Gao, G. Li, Point cloud pre-trained models and large models, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 195–225
43. W. Gao, G. Li, Point cloud-language multi-modal learning, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 227–254
44. W. Gao, G. Li, Open-source projects for 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 255–272
45. W. Gao, G. Li, Typical engineering applications of 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2025), pp. 273–299
46. W. Gao, G. Li, Future work on deep learning-based point cloud technologies, in *Deep Learning for 3D Point Clouds* (Springer, 2025), pp. 301–315
47. Y. Wu, L. Xie, S. Sun, W. Gao, Y. Yan, Adaptive intra period size for deep learning-based screen content video coding, in *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2024), pp. 1–6
48. H. Zheng, W. Gao, End-to-end rgb-d image compression via exploiting channel-modality redundancy. *Proc. AAAI Conf. Artif. Intell.* **38**(7), 7562–7570 (2024)
49. L. Tao, W. Gao, G. Li, C. Zhang, Adanic: towards practical neural image compression via dynamic transform routing, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 16 879–16 888
50. Y. Wu, W. Gao, End-to-end lossless compression of high precision depth maps guided by pseudo-residual (2022). [arXiv:2201.03195](https://arxiv.org/abs/2201.03195)
51. Y. Wu, Z. Qi, H. Zheng, L. Tao, W. Gao, Deep image compression with latent optimization and piece-wise quantization approximation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 1926–1930

52. W. Gao, L. Tao, L. Zhou, D. Yang, X. Zhang, Z. Guo, Low-rate image compression with super-resolution learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2020), pp. 154–155
53. W. Gao, S. Sun, H. Zheng, Y. Wu, H. Ye, Y. Zhang, Opendmc: an open-source library and performance evaluation for deep-learning-based multi-frame compression, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9685–9688
54. W. Gao, H. Zheng, C. Zhang, K. Zheng, Z. Yu, Y. Li, H. Ye, Y. Zhang, Opendic: an open-source library and performance evaluation for deep-learning-based image compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 202–11 205
55. Y. Li, W. Gao, G. Li, S. Ma, Saliency segmentation oriented deep image compression with novel bit allocation. *IEEE Trans. Image Process.* **34**, 16–29 (2025)
56. C. Zhang, W. Gao, Learned rate control for frame-level adaptive neural video compression via dynamic neural network, in *European Conference on Computer Vision* (Springer, 2024)
57. L. Xie, W. Gao, H. Zheng, G. Li, Spcgc: scalable point cloud geometry compression for machine vision, in *2024 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2024), pp. 17 272–17 278
58. L. Xie, W. Gao, H. Zheng, H. Ye, Semantic-aware visual decomposition for point cloud geometry compression, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 595–595
59. Z. Qi, W. Gao, Variable-rate point cloud geometry compression based on feature adjustment and interpolation, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 63–72
60. Z. Yu, W. Gao, When dynamic neural network meets point cloud compression: computation-aware variable rate and checkerboard context, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 600–600
61. L. Xie, W. Gao, S. Fan, Z. Yao, Pdnet: parallel dual-branch network for point cloud geometry compression and analysis, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 596–596
62. L. Xie, W. Gao, H. Zheng, End-to-end point cloud geometry compression and analysis with sparse tensor, in *International Workshop on Advances in Point Cloud Compression, Processing and Analysis* (2022), pp. 27–32
63. C. Fu, G. Li, R. Song, W. Gao, S. Liu, Octattention: octree-based large-scale contexts model for point cloud compression. *Proc. AAAI Conf. Artif. Intell.* **36**(1), 625–633 (2022)
64. C. Zhang, W. Gao, Adapcc: adaptive rate control and rate-distortion-complexity optimization for dynamic point cloud compression, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
65. K. Wang, W. Gao, Unipcgc: towards practical point cloud geometry compression via an efficient unified approach, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
66. K. Zheng, W. Gao, H. Zheng, Octree-based learned point cloud geometry compression: a lossy perspective, in *2025 Data Compression Conference (DCC)* (IEEE, 2025)
67. Z. Qi, H. Liu, W. Gao, Variable-rate point cloud compression with channel-wise entropy modeling, in *2025 Data Compression Conference (DCC)* (IEEE, 2025)
68. H. Zheng, W. Gao, Z. Yu, T. Zhao, G. Li, Viewpcgc: view-guided learned point cloud geometry compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 7152–7161
69. L. Xie, W. Gao, H. Zheng, G. Li, Roi-guided point cloud geometry compression towards human and machine vision, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 3741–3750
70. W. Liu, W. Gao, X. Mu, Fast inter-frame motion prediction for compressed dynamic point cloud attribute enhancement. *Proc. AAAI Conf. Artif. Intell.* **38**(4), 3720–3728 (2024)
71. Z. Yang, W. Gao, X. Lu, Danet: density-adaptive network for geometry-based point cloud compression artifacts removal, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5

72. X. Fan, G. Li, D. Li, Y. Ren, W. Gao, T.H. Li, Deep geometry post-processing for decompressed point clouds, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
73. X. Zhang, G. Liao, W. Gao, G. Li, Tdnet: transformer-based dual-branch restoration network for geometry based point cloud compression artifacts, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
74. Z. Li, G. Li, T.H. Li, S. Liu, W. Gao, Semantic point cloud upsampling. *IEEE Trans. Multimedia* (2022)
75. Z. Li, S. Liu, W. Gao, G. Li, G. Li, S4r: rethinking point cloud sampling via guiding upsampling-aware perception. *IEEE Trans. Multimedia* (2025)
76. R. Zhang, W. Gao, G. Li, T.H. Li, Qinet: decision surface learning and adversarial enhancement for quasi-immune completion of diverse corrupted point clouds. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–14 (2022)
77. R. Bao, Y. Ren, G. Li, W. Gao, S. Liu, Flow-based point cloud completion network with adversarial refinement, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2559–2563
78. J. Chen, G. Li, R. Zhang, T.H. Li, W. Gao, Pointvae: invertible variational autoencoder framework for 3d point cloud generation, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 3216–3220
79. R. Zhang, J. Chen, W. Gao, G. Li, T.H. Li, Pointot: interpretable geometry-inspired point cloud generative model via optimal transport. *IEEE Trans. Circuits Syst. Video Technol.* **32**(10), 6792–6806 (2022)
80. Z. Li, G. Li, T. Li, S. Liu, W. Gao, Information-growth attention network for image super-resolution, in *Proceedings of the 29th ACM International Conference on Multimedia* (2021), pp. 544–552
81. L. Zhou, W. Gao, G. Li, H. Yuan, T. Zhao, G. Yue, Disentangled feature distillation for light field super-resolution with degradations, in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2023), pp. 116–121
82. L. Zhou, W. Gao, G. Li, End-to-end spatial-angular light field super-resolution using parallax structure preservation strategy, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 3396–3400
83. W. Gao, L. Zhou, L. Tao, A fast view synthesis implementation method for light field applications. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* **17**(4), 1–20 (2021)
84. X. Zhang, W. Gao, G. Li, Q. Jiang, R. Cong, Image quality assessment-driven reinforcement learning for mixed distorted image restoration. *ACM Trans. Multimedia Comput. Commun. Appl.* **19**(1s), 1–23 (2023)
85. X. Zhang, W. Gao, H. Yuan, G. Li, Je 2 net: joint exploitation and exploration in reinforcement learning based image restoration, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2090–2094
86. X. Zhang, W. Gao, Hirl: hybrid image restoration based on hierarchical deep reinforcement learning via two-step analysis, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2445–2449
87. S. Wang, Y. Lu, W. Xia, P. Xia, Z. Wang, W. Gao, Light field angular super-resolution by view-specific queries. *Visual Comput.* 1–16 (2024)
88. Y. Sun, Z. Li, S. Wang, W. Gao, Depth-assisted calibration on learning-based factorization for a compressive light field display. *Opt. Express* **31**(4), 5399–5413 (2023)
89. Y. Sun, Z. Li, L. Li, S. Wang, W. Gao, Optimization of compressive light field display in dual-guided learning, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2075–2079
90. W. Gao, S. Fan, G. Li, W. Lin, A thorough benchmark and a new model for light field saliency detection. *IEEE Trans. Pattern Anal. Mach. Intell.* (2023)
91. Z. Guo, W. Gao, H. Wang, J. Wang, S. Fan, No-reference deep quality assessment of compressed light field images, in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6

92. G. Li, G. Wei, W. Gao, *Point Cloud Compression: Technologies and Standardization* (Springer Nature, 2024)
93. G. Li, W. Gao, W. Gao, Introduction, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 1–28
94. G. Li, W. Gao, W. Gao, Background knowledge, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 29–51
95. G. Li, W. Gao, W. Gao, Predictive coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 53–70
96. G. Li, W. Gao, W. Gao, Transform coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 71–96
97. G. Li, W. Gao, W. Gao, Quantization techniques, in *Point Cloud Compression: Technologies and Standardization* (2024), pp. 97–112
98. G. Li, W. Gao, W. Gao, Entropy coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 113–133
99. G. Li, W. Gao, W. Gao, Mpeg geometry-based point cloud compression (g-pcc) standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 135–165
100. G. Li, W. Gao, W. Gao, Avs point cloud compression standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 167–197
101. G. Li, W. Gao, W. Gao, Mpeg video-based point cloud compression (v-pcc) standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 199–218
102. G. Li, W. Gao, W. Gao, Mpeg ai-based 3d graphics coding standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 219–241
103. G. Li, W. Gao, W. Gao, Future work, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 243–250
104. T. Qin, G. Li, W. Gao, S. Liu, Multi-grained point cloud geometry compression via dual-model prediction with extended octree. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(9), 1–30 (2024)
105. Y. Shao, W. Gao, S. Liu, G. Li, Advanced patch-based affine motion estimation for dynamic point cloud geometry compression. *Sensors* **24**(10), 3142 (2024)
106. Y. Shao, F. Song, W. Gao, S. Liu, G. Li, Texture-guided graph transform optimization for point cloud attribute compression. *Appl. Sci.* **14**(10), 4094 (2024)
107. Y. Shao, X. Yang, W. Gao, S. Liu, G. Li, 3d point cloud attribute compression using diffusion-based texture-aware intra prediction. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
108. J. Zhang, Y. Chen, G. Liu, W. Gao, G. Li, Efficient point cloud attribute compression framework using attribute-guided graph fourier transform, in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2024), pp. 8426–8430
109. W. Gao, H. Yuan, G. Li, Z. Li, H. Yuan, Low complexity coding unit decision for video-based point cloud compression. *IEEE Trans. Image Process.* **33**, 149–162 (2023)
110. Y. Shao, G. Li, Q. Zhang, W. Gao, S. Liu, Nonrigid registration-based progressive motion compensation for point cloud geometry compression. *IEEE Trans. Geosci. Remote Sens.* (2023)
111. X. Chen, Y. Wang, Y. Li, X. Ling, M. Li, R. Liu, M. Ouyang, K. Zhao, T. Guan, Y. He, Low bit-width zero-shot quantization with soft feature-infused hints for iot systems. *IEEE Internet Things J.* (2024)
112. F. Song, G. Li, X. Yang, W. Gao, S. Liu, Block-adaptive point cloud attribute coding with region-aware optimized transform. *IEEE Trans. Circuits Syst. Video Technol.* (2023)
113. Y. An, Y. Shao, G. Li, W. Gao, S. Liu, A fast motion estimation method with hamming distance for lidar point cloud compression, in *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2022), pp. 1–5
114. H. Yuan, W. Gao, G. Li, Z. Li, Rate-distortion-guided learning approach with cross-projection information for v-pcc fast cu decision, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 3085–3093

115. F. Song, G. Li, W. Gao, T.H. Li, Rate-distortion optimized graph for point cloud attribute coding. *IEEE Signal Process. Lett.* **29**, 922–926 (2022)
116. F. Song, G. Li, X. Yang, W. Gao, T.H. Li, Fine-grained correlation representation for graph-based point cloud attribute compression, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
117. F. Shen, W. Gao, A rate control algorithm for video-based point cloud compression, in *2021 International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2021), pp. 1–5
118. F. Song, Y. Shao, W. Gao, H. Wang, T. Li, Layer-wise geometry aggregation framework for lossless lidar point cloud compression. *IEEE Trans. Circuits Syst. Video Technol.* **31**(12), 4603–4616 (2021)
119. Z. Cai, W. Gao, G. Li, W. Gao, Distortion propagation model-based v-pcc rate control for 3d point cloud broadcasting. *IEEE Trans. Broadcast.* 1–13 (2024)
120. X. Mu, W. Gao, H. Yuan, S. Wang, G. Li, Adaptive lpu decision for dynamic point cloud compression. *IEEE Signal Process. Lett.* (2024)
121. C. Peng, W. Gao, Laplacian matrix learning for point cloud attribute compression with ternary search-based adaptive block partition, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 10 412–10 420
122. Y. Wang, W. Gao, X. Mu, H. Yuan, Rate control optimization for joint geometry and attribute coding of lidar point clouds, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
123. T. Guo, H. Yuan, T. Wang, W. Gao, Graph filter-based fast motion matching for inter frame coding of mpeg g-pcc, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 1151–1155
124. Z. Yang, W. Gao, G. Li, Y. Yan, Sur-driven video coding rate control for jointly optimizing perceptual quality and buffer control. *IEEE Trans. Image Process.* (2023)
125. F. Shen, Z. Cai, W. Gao, An efficient rate control algorithm for intra frame coding in avs3, in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2021), pp. 3164–3169
126. H. Yuan, W. Gao, J. Wang, Dynamic computational resource allocation for fast inter frame coding in video conferencing applications, in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6
127. W. Gao, Q. Jiang, R. Wang, S. Ma, G. Li, S. Kwong, Consistent quality oriented rate control in hevc via balancing intra and inter frame coding. *IEEE Trans. Ind. Inform.* **18**(3), 1594–1604 (2021)
128. H. Yuan, W. Gao, A new coding unit partitioning mode for screen content video coding, in *Proceedings of the 2021 5th International Conference on Digital Signal Processing* (2021), pp. 66–72
129. W. Gao, On the performance evaluation of state-of-the-art rate control algorithms for practical video coding and transmission systems, in *Proceedings of the 2020 4th International Conference on Video and Image Processing* (2020), pp. 179–185
130. W. Gao, S. Kwong, Q. Jiang, C.-K. Fong, P.H. Wong, W.Y. Yuen, Data-driven rate control for rate-distortion optimization in hevc based on simplified effective initial qp learning. *IEEE Trans. Broadcast.* **65**(1), 94–108 (2018)
131. W. Gao, A multi-objective optimization perspective for joint consideration of video coding quality, in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (IEEE, 2019), pp. 986–991
132. W. Gao, S. Kwong, Y. Jia, Joint machine learning and game theory for rate control in high efficiency video coding. *IEEE Trans. Image Process.* **26**(12), 6074–6089 (2017)
133. W. Gao, S. Kwong, Y. Zhou, H. Yuan, Ssim-based game theory approach for rate-distortion optimized intra frame ctu-level bit allocation. *IEEE Trans. Multimedia* **18**(6), 988–999 (2016)
134. W. Gao, S. Kwong, H. Yuan, X. Wang, Dct coefficient distribution modeling and quality dependency analysis based frame-level bit allocation for hevc. *IEEE Trans. Circuits Syst. Video Technol.* **26**(1), 139–153 (2015)

135. W. Gao, S. Kwong, Phase congruency based edge saliency detection and rate control for perceptual image and video coding, in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2016), pp. 000 264–000 269
136. H. Yuan, W. Gao, Openfastvc: An open source library for video coding fast algorithm implementation, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9660–9663
137. H. Yuan, W. Gao, S. Ma, Y. Yan, Divide-and-conquer-based rdo-free cu partitioning for 8k video compression. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(4), 1–20 (2024)
138. L. Tao, W. Gao, A hardware implementation of entropy encoder for 8k video coding, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
139. Y. Guo, W. Gao, S. Ma, G. Li, Accelerating transform algorithm implementation for efficient intra coding of 8k uhd videos. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* **18**(4), 1–20 (2022)
140. Z. Cai, W. Gao, Efficient fast algorithm and parallel hardware architecture for intra prediction of avs3, in *IEEE International Symposium on Circuits and Systems (ISCAS)* (IEEE, 2021), pp. 1–5
141. W. Gao, H. Yuan, Y. Guo, L. Tao, Z. Cai, G. Li, Openhardwarevc: an open source library for 8k uhd video coding hardware implementation, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 7339–7342
142. W. Gao, H. Yuan, G. Liao, Z. Guo, J. Chen, Pp8k: a new dataset for 8k uhd video compression and processing. *IEEE MultiMedia* (2023)
143. W. Liu, W. Gao, G. Li, S. Ma, T. Zhao, H. Yuan, Enlarged motion-aware and frequency-aware network for compressed video artifact reduction. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
144. L. Xie, X. Mu, G. Li, W. Gao et al., Pku-dpcc: a new dataset for dynamic point cloud compression. *APSIPA Trans. Signal Inf. Process.* **13**(6) (2024)
145. H. Yuan, W. Gao, W. Gao, Opensep: an open source subjective experiment platform, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 248–11 251
146. S. Fan, W. Gao, Screen-based 3d subjective experiment software, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9672–9675
147. J. Wang, W. Gao, G. Li, Zoom to perceive better: No-reference point cloud quality assessment via exploring effective multiscale feature. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
148. J. Wang, W. Gao, G. Li, Applying collaborative adversarial learning to blind point cloud quality measurement. *IEEE Trans. Instrum. Meas.* (2023)
149. H. Li, W. Gao, Information exploration of projected views for point cloud quality measurement. *IEEE Trans. Instrum. Meas.* (2024)
150. B. Qu, X. Liang, S. Sun, W. Gao, Exploring aigc video quality: a focus on visual harmony, video-text consistency and domain distribution gap, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2024)
151. B. Qu, H. Li, W. Gao, Bringing textual prompt to ai-generated image quality assessment, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024)
152. S. Sun, X. Liang, S. Fan, W. Gao, W. Gao, Ve-bench: subjective-aligned benchmark suite for text-driven video editing quality assessment, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
153. P. Molchanov, S. Tyree, T. Karras, T. Aila, J. Kautz, Pruning convolutional neural networks for resource efficient inference (2016). [arXiv:1611.06440](https://arxiv.org/abs/1611.06440)
154. Y. He, X. Zhang, J. Sun, Channel pruning for accelerating very deep neural networks, in *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 1389–1397
155. Z. Liu, M. Sun, T. Zhou, G. Huang, T. Darrell, Rethinking the value of network pruning (2018). [arXiv:1810.05270](https://arxiv.org/abs/1810.05270)
156. W. Su, Z. Li, M. Xu, J. Kang, D. Niyato, S. Xie, Compressing deep reinforcement learning networks with a dynamic structured pruning method for autonomous driving. *IEEE Trans. Veh. Technol.* (2024)

157. T. Liang, J. Glossner, L. Wang, S. Shi, X. Zhang, Pruning and quantization for deep neural network acceleration: A survey. *Neurocomputing* **461**, 370–403 (2021)
158. H. Cheng, M. Zhang, J.Q. Shi, A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations. *IEEE Trans. Pattern Anal. Mach. Intell.* (2024)
159. W. Wen, C. Wu, Y. Wang, Y. Chen, H. Li, Learning structured sparsity in deep neural networks. *Adv. Neural Inf. Process. Syst.* **29**, 2074–2082 (2016)
160. Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, C. Zhang, Learning efficient convolutional networks through network slimming, in *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 2736–2744
161. M. Tan, Q. Le, Efficientnet: rethinking model scaling for convolutional neural networks, in *International Conference on Machine Learning* (2019), pp. 6105–6114
162. S. Han, J. Pool, J. Tran, W. Dally, Learning both weights and connections for efficient neural network. *Adv. Neural Inf. Process. Syst.* **28**, 1135–1143 (2015)
163. H. Hakkak, Auto deep compression by reinforcement learning based actor-critic structure (2018). [arXiv:1807.02886](https://arxiv.org/abs/1807.02886)
164. M. Zhu, S. Gupta, To prune, or not to prune: exploring the efficacy of pruning for model compression (2017). [arXiv:1710.01878](https://arxiv.org/abs/1710.01878)
165. Y. Gong, L. Liu, M. Yang, L.D. Bourdev, Compressing deep convolutional networks using vector quantization (2014). [arXiv:abs/1412.6115](https://arxiv.org/abs/1412.6115)
166. Y. Guo, A survey on methods and theories of quantized neural networks (2018). [arXiv:1808.04752](https://arxiv.org/abs/1808.04752)
167. L. Wei, Z. Ma, C. Yang, Q. Yao, Advances in the neural network quantization: a comprehensive review. *Appl. Sci.* **14**(17), 7445 (2024)
168. M. Li, Z. Huang, L. Chen, J. Ren, M. Jiang, F. Li, J. Fu, C. Gao, Contemporary advances in neural network quantization: a survey, in *2024 International Joint Conference on Neural Networks (IJCNN)* (IEEE, 2024), pp. 1–10
169. Y. Bhalgat, J. Lee, M. Nagel, T. Blankevoort, N. Kwak, Lsq+: improving low-bit quantization through learnable offsets and better initialization, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2020), pp. 696–697
170. B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, D. Kalenichenko, Quantization and training of neural networks for efficient integer-arithmetic-only inference, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 2704–2713
171. S.K. Esser, J.L. McKinstry, D. Bablani, R. Appuswamy, D.S. Modha, Learned step size quantization (2019). [arXiv:1902.08153](https://arxiv.org/abs/1902.08153)
172. J. Yim, D. Joo, J. Bae, J. Kim, A gift from knowledge distillation: Fast optimization, network minimization and transfer learning, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 4133–4141
173. X. Zhu, S. Gong et al., Knowledge distillation by on-the-fly native ensemble. *Adv. Neural Inf. Process. Syst.* **31** (2018)
174. J. Gou, B. Yu, S.J. Maybank, D. Tao, Knowledge distillation: a survey. *Int. J. Comput. Vis.* **129**(6), 1789–1819 (2021)
175. G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network (2015). [arXiv:1503.02531](https://arxiv.org/abs/1503.02531)
176. A. Romero, N. Ballas, S.E. Kahou, A. Chassang, C. Gatta, Y. Bengio, Fitnets: hints for thin deep nets (2014). [arXiv:1412.6550](https://arxiv.org/abs/1412.6550)
177. Z. Huang, N. Wang, Like what you like: Knowledge distill via neuron selectivity transfer (2017). [arXiv:1707.01219](https://arxiv.org/abs/1707.01219)
178. J. Kim, S. Park, N. Kwak, Paraphrasing complex network: network compression via factor transfer. *Adv. Neural Inf. Process. Syst.* **31**, 2765–2774 (2018)
179. S. Zagoruyko, N. Komodakis, Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer (2016). [arXiv:1612.03928](https://arxiv.org/abs/1612.03928)
180. B. Heo, M. Lee, S. Yun, J.Y. Choi, Knowledge transfer via distillation of activation boundaries formed by hidden neurons. *Proc. AAAI Conf. Artif. Intell.* **33**(01), 3779–3787 (2019)

181. L. Xie, W. Gao, Pchmvision: an open-source library of point cloud compression for human and machine vision, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 239–11 243
182. L. Xie, W. Gao, Learningpcc: a pytorch library for learning-based point cloud compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 234–11 238
183. T. Elsken, J.H. Metzen, F. Hutter, Neural architecture search: a survey. *J. Mach. Learn. Res.* **20**(1), 1997–2017 (2019)
184. A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, J. Kepner, Ai accelerator survey and trends, in *IEEE High Performance Extreme Computing Conference* (2021), pp. 1–9
185. W. Gao, H. Ye, G. Li, H. Zheng, Y. Wu, L. Xie, OpenPointCloud: an open-source algorithm library of deep learning based point cloud compression, in *ACM International Conference on Multimedia* (2022), pp. 7347–7350
186. L. Xie, M. Hu, X. Bai, towards hardware-friendly and robust facial landmark detection method, in *International Conference on Neural Information Processing* (Springer, 2022), pp. 432–444
187. L. Xie, M. Hu, X. Bai, Online improved vehicle tracking accuracy via unsupervised route generation, in *IEEE 34th International Conference on Tools with Artificial Intelligence* (IEEE, 2022), pp. 788–792
188. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D.G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, X. Zheng, TensorFlow: a system for Large-Scale machine learning, in *12th USENIX Symposium on Operating Systems Design and Implementation* (2016), pp. 265–283
189. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: an imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32**, 8024–8035 (2019)
190. Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, in *Proceedings of the 22nd ACM International Conference on Multimedia* (2014), pp. 675–678
191. S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanzaro, E. Shelhamer, cudnn: efficient primitives for deep learning (20140. [arXiv:1410.0759](https://arxiv.org/abs/1410.0759))
192. S. Markidis, S.W.D. Chien, E. Laure, I.B. Peng, J.S. Vetter, Nvidia tensor core programmability, performance and precision, in *IEEE International Parallel and Distributed Processing Symposium Workshops* (2018), pp. 522–531
193. J. Choquette, W. Gandhi, O. Giroux, N. Stam, R. Krashinsky, Nvidia a100 tensor core gpu: Performance and innovation. *IEEE Micro* **41**(2), 29–35 (2021)
194. Y. Chen, T. Luo, S. Liu, S. Zhang, L. He, J. Wang, L. Li, T. Chen, Z. Xu, N. Sun, O. Temam, Dadiannao: a machine-learning supercomputer, in *2014 47th Annual IEEE/ACM International Symposium on Microarchitecture* (2014), pp. 609–622
195. T. Chen, Z. Du, N. Sun, J. Wang, C. Wu, Y. Chen, O. Temam, Diannao: a small-footprint high-throughput accelerator for ubiquitous machine-learning, in *Proceedings of the 19th International Conference on Architectural Support for Programming Languages and Operating Systems* (2014), pp. 269–284
196. Z. Du, R. Fasthuber, T. Chen, P. Ienne, L. Li, T. Luo, X. Feng, Y. Chen, O. Temam, Shidiannao: shifting vision processing closer to the sensor, in *Proceedings of the 42nd Annual International Symposium on Computer Architecture* (2015), pp. 92–104
197. Y. Zhao, Z. Du, Q. Guo, S. Liu, L. Li, Z. Xu, T. Chen, Y. Chen, Cambricon-f: Machine learning computers with fractal von neumann architecture, in *Proceedings of the 46th International Symposium on Computer Architecture* (2019), pp. 788–801
198. Y. Zhao, C. Liu, Z. Du, Q. Guo, X. Hu, Y. Zhuang, Z. Zhang, X. Song, W. Li, X. Zhang, L. Li, Z. Xu, T. Chen, Cambricon-q: a hybrid architecture for efficient training, in *48th ACM/IEEE Annual International Symposium on Computer Architecture* (2021), pp. 706–719

199. X. Zhou, Z. Du, Q. Guo, S. Liu, C. Liu, C. Wang, X. Zhou, L. Li, T. Chen, Y. Chen, Cambricon-s: addressing irregularity in sparse neural networks through a cooperative software/hardware approach, in *51st Annual IEEE/ACM International Symposium on Microarchitecture* (2018), pp. 15–28
200. N.P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, R. Boyle, P.-I. Cantin, C. Chao, C. Clark, J. Coriell, M. Daley, M. Dau, J. Dean, B. Gelb, T.V. Ghaemmaghami, R. Gottipati, W. Gulland, R. Hagmann, C.R. Ho, D. Hogberg, J. Hu, R. Hundt, D. Hurt, J. Ibarz, A. Jaffey, A. Jaworski, A. Kaplan, H. Khaitan, D. Killebrew, A. Koch, N. Kumar, S. Lacy, J. Laudon, J. Law, D. Le, C. Leary, Z. Liu, K. Lucke, A. Lundin, G. MacKean, A. Maggiore, M. Mahony, K. Miller, R. Nagarajan, R. Narayanaswami, R. Ni, K. Nix, T. Norrie, M. Omernick, N. Penukonda, A. Phelps, J. Ross, M. Ross, A. Salek, E. Samadiani, C. Severn, G. Sizikov, M. Snellman, J. Souter, D. Steinberg, A. Swing, M. Tan, G. Thorson, B. Tian, H. Toma, E. Tuttle, V. Vasudevan, R. Walter, W. Wang, E. Wilcox, D.H. Yoon, In-datacenter performance analysis of a tensor processing unit, in *Proceedings of the 44th Annual International Symposium on Computer Architecture* (2017), pp. 1–12
201. N. Jouppi, G. Kurian, S. Li, P. Ma, R. Nagarajan, L. Nai, N. Patil, S. Subramanian, A. Swing, B. Towles, C. Young, X. Zhou, Z. Zhou, D.A. Patterson, Tpu v4: an optically reconfigurable supercomputer for machine learning with hardware support for embeddings, in *Proceedings of the 50th Annual International Symposium on Computer Architecture* (2023), pp. 1–14
202. H. Sharma, J. Park, N. Suda, L. Lai, B. Chau, V. Chandra, H. Esmailzadeh, Bit fusion: bit-level dynamically composable architecture for accelerating deep neural network, in *45th ACM/IEEE Annual International Symposium on Computer Architecture* (2018), pp. 764–775
203. S. Yin, P. Ouyang, J. Yang, T. Lu, X. Li, L. Liu, S. Wei, An ultra-high energy-efficient reconfigurable processor for deep neural networks with binary/ternary weights in 28nm cmos, in *IEEE Symposium on VLSI Circuits* (2018), pp. 37–38
204. A. Gondimalla, N. Chesnut, M. Thottethodi, T.N. Vijaykumar, Sparten: a sparse tensor accelerator for convolutional neural networks, in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture* (2019), pp. 151–165
205. Y. Chen, J. He, X. Zhang, C. Hao, D. Chen, Cloud-dnn: an open framework for mapping dnn models to cloud fpgas, in *Proceedings of the 2019 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays* (2019), pp. 73–82
206. M. Blott, T.B. Preußer, N.J. Fraser, G. Gambardella, K. O'brien, Y. Umuroglu, M. Leeser, K. Vissers, Finn-r: an end-to-end deep-learning framework for fast exploration of quantized neural networks. *ACM Trans. Reconfigurable Technol. Syst.* **11**(3), 1–23 (2018)
207. L. Lu, Y. Liang, Spwa: an efficient sparse winograd convolutional neural networks accelerator on fpgas, in *Proceedings of the 55th Annual Design Automation Conference* (2018), pp. 135:1–135:6
208. V. Kathail, Xilinx vitis unified software platform, in *The ACM/SIGDA International Symposium on Field-Programmable Gate Arrays* (2020), pp. 173–174
209. Amazon SageMaker Documentation, <https://docs.aws.amazon.com/sagemaker/>
210. H. Jin, F. Chollet, Q. Song, X. Hu, Autokeras: an automl library for deep learning. *J. Mach. Learn. Res.* **24**(6), 1–6 (2023). [Online]. Available: <http://jmlr.org/papers/v24/20-1355.html>
211. M. Feurer, A. Klein, K. Eggenberger, J.T. Springenberg, M. Blum, F. Hutter, Efficient and robust automated machine learning, in *Advances in Neural Information Processing Systems* (2015), pp. 2962–2970
212. T.T. Le, W. Fu, J.H. Moore, Scaling tree-based automated machine learning to biomedical big data with a feature set selector. *Bioinformatics* **36**(1), 250–256 (2020)
213. Kubeflow, <https://www.kubeflow.org/>
214. Microsoft, Neural Network Intelligence, 1 2021. [Online]. Available: <https://github.com/microsoft/nni>
215. W. Gao, G. Li, H. Yuan, R. Hamzaoui, Z. Li, S. Liu, Apccpa'22: 1st international workshop on advances in point cloud compression, processing and analysis, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 7392–7393

216. W. Gao, G. Li, Point cloud compression, enhancement and applications: from 3d perception to large models, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 292–11 293
217. H. Sun, G. Li, S. Chen, L. Li, W. Gao, Special topic on 3d point cloud processing and applications. *ZTE Commun.* **21**(4), 1 (2023)
218. W. Gao, W. Gao, X. Mu, C. Peng, G. Li, Overview and comparison of avs point cloud compression standard. *APSIPA Trans. Signal Inf. Process.* (2024)
219. W. Gao, S. Wang, X. Zhang, S. Kwong, Editorial for special issue on deep learning-based data compression. *APSIPA Trans. Signal Inf. Process.* **13**(6) (2024)
220. Y. Zhang, W. Gao, G. Li, ‘Openpointcloud-v2: a deep learning based open-source algorithm library of point cloud processing, in *Proceedings of the 1st International Workshop on Advances in Point Cloud Compression, Processing and Analysis* (2022), pp. 51–55
221. F. Yang, L. Herranz, Y. Cheng, M.G. Mozerov, Slimmable compressive autoencoders for practical neural image compression, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 4998–5007
222. J. Guo, D. Xu, G. Lu, Cbanet: toward complexity and bitrate adaptive deep image compression using a single network. *IEEE Trans. Image Process.* **32**, 2049–2062 (2023)
223. C. Jia, X. Hang, S. Wang, Y. Wu, S. Ma, W. Gao, Fpx-nic: an fpga-accelerated 4k ultra-high-definition neural video coding system. *IEEE Trans. Circuits Syst. Video Technol.* **32**(9), 6385–6399 (2022)
224. H. Sun, Z. Cheng, M. Takeuchi, J. Katto, End-to-end learned image compression with fixed point weight quantization, in *IEEE International Conference on Image Processing* (2020), pp. 3359–3363
225. T. George, S. Wenzhe, T. Radu, T. Lucas, B. Johannes, A. Eirikur, J. Nick, M. Fabian, ‘Workshop and challenge on learned image compression (2020) [Online]. Available: <http://www.compression.cc>
226. L. Xie, X. Mu, G. Li, W. Gao et al., Pku-dpcc: a new dataset for dynamic point cloud compression. *APSIPA Trans. Signal Inf. Process.* **13**(6) (2024)
227. L. Xie, W. Gao, S. Fan, Z. Yao, Pdnnet: parallel dual-branch network for point cloud geometry compression and analysis, in *Data Compression Conference* (2024), pp. 596–596
228. L. Xie, W. Gao, H. Zheng, H. Ye, Semantic-aware visual decomposition for point cloud geometry compression, in *Data Compression Conference* (2024), pp. 595–595
229. L. Xie, W. Gao, H. Zheng, G. Li, SPCGC: scalable Point Cloud Geometry Compression for Machine Vision, in *IEEE International Conference on Robotics and Automation* (2024), pp. 594–595
230. F. Al-Ali, T.D. Gamage, H.W. Nanayakkara, F. Mehdipour, S.K. Ray, Novel case study and benchmarking of alexnet for edge ai: from cpu and gpu to fpga, in *IEEE Canadian Conference on Electrical and Computer Engineering* (2020), pp. 1–4
231. Y. Chen, H. Fan, B. Xu, Z. Yan, Y. Kalantidis, M. Rohrbach, Y. Shuicheng, J. Feng, Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution, in *IEEE/CVF International Conference on Computer Vision* (2019), pp. 3434–3443
232. Y.-H. Chen, T. Krishna, J.S. Emer, V. Sze, Eyeriss: an energy-efficient reconfigurable accelerator for deep convolutional neural networks. *IEEE J. Solid-State Circuits* **52**(1), 127–138 (2017)
233. Y.-H. Chen, T.-J. Yang, J.S. Emer, V. Sze, Eyeriss v2: a flexible accelerator for emerging deep neural networks on mobile devices. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **9**(2), 292–308 (2019)
234. J. Choquette, O. Giroux, D. Foley, Volta: performance and programmability. *IEEE Micro* **38**(2), 42–52 (2018). (Mar.)
235. C. Choy, J. Gwak, S. Savarese, 4d spatio-temporal convnets: Minkowski convolutional neural networks, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 3075–3084

236. M. Sun, Z. Li, A. Lu, Y. Li, S.-E. Chang, X. Ma, X. Lin, Z. Fang, Film-qnn: efficient fpga acceleration of deep neural networks with intra-layer, mixed-precision quantization, in *Proceedings of the 2022 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays* (2022), pp. 134–145
237. M. Quach, G. Valenzise, F. Dufaux, Improved deep point cloud geometry compression, in *IEEE International Workshop on Multimedia Signal Processing* (2020), pp. 1–6
238. J. Wang, D. Ding, Z. Li, Z. Ma, Multiscale point cloud geometry compression, in *Data Compression Conference* (2021), pp. 73–82
239. A.F. Guarda, N.M. Rodrigues, F. Pereira, Adaptive deep learning-based point cloud geometry coding. *IEEE J. Sel. Top. Signal Process.* **15**(2), 415–430 (2020)
240. J. Wang, D. Ding, Z. Li, X. Feng, C. Cao, Z. Ma, Sparse Tensor-based Multiscale Representation for Point Cloud Geometry Compression (2021). [ArXiv:2111.10633](https://arxiv.org/abs/2111.10633)
241. A. Luo, H. Sun, J. Liu, J. Katto, Memory-efficient learned image compression with pruned hyperprior module, in *IEEE International Conference on Image Processing* (2022), pp. 3061–3065
242. M. Akbari, J. Liang, J. Han, C. Tu, Generalized octave convolutions for learned multi-frequency image compression (2020). [arXiv:2002.10032](https://arxiv.org/abs/2002.10032)
243. D. He, Y. Zheng, B. Sun, Y. Wang, H. Qin, Checkerboard context model for efficient learned image compression, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 14 771–14 780
244. D. Minnen, S. Singh, Channel-wise autoregressive entropy models for learned image compression, in *IEEE International Conference on Image Processing* (2020), pp. 3339–3343
245. T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, A. Anandkumar, Born again neural networks, in *International Conference on Machine Learning* (2018), pp. 1607–1616
246. L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, K. Ma, Be your own teacher: Improve the performance of convolutional neural networks via self distillation, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 3713–3722
247. H. Liu, K. Simonyan, Y. Yang, Darts: differentiable architecture search (2018). [arXiv:1806.09055](https://arxiv.org/abs/1806.09055)
248. H. Cai, L. Zhu, S. Han, Proxylessnas: direct neural architecture search on target task and hardware (2018). [arXiv:1812.00332](https://arxiv.org/abs/1812.00332)
249. L. Tao, W. Gao, G. Li, C. Zhang, Adanic: Towards Practical Neural Image Compression Via Dynamic Transform Routing (2023), pp. 16 833–16 842
250. D. Minnen, J. Ballé, G. Toderici, Joint autoregressive and hierarchical priors for learned image compression, in *Advances in Neural Information Processing Systems* (2018), pp. 10 794–10 803
251. X. Wei, J. Lin, J. Xu, W. Gao, T. Zhao, Rdvc: efficient deep video compression with regulable rate and complexity optimization. *IEEE Trans. Multimedia* (2024)
252. X. Liu, H. Chang, J. Wei, L. Zhu, L. Liu, L. Li, S. Zhou, C. Li, D. Xu, W. Gao, Implanting robust watermarks in latent diffusion models for video generation, in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2025)
253. J. Wei, H. Chang, X. Liu, L. Liu, L. Li, S. Zhou, C. Li, D. Xu, W. Gao, R. Liao, Dual-process watermarked diffusion: integrating watermarking with denoising in point clouds, in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2025)
254. H. Liu, Y. Xu, S. Tao, Z. Huang, B. Xie, X. Li, W. Gao, Simultaneous conjugate gradient and iaff-unet for accurate ir drop calculation, in *IEEE International Conference on Computer Design* (2024)
255. X. Mao, H. Yuan, X. Lu, R. Hamzaoui, W. Gao, Pcac-gan: a sparse-tensor-based generative adversarial network for 3d point cloud attribute compression. *Comput. Visual Media* (2024)
256. J. Lin, S. Pei, B. Chen, N. Jiang, W. Gao, T. Zhao, Ldrm: degradation rectify model for low-light imaging via color-monochrome cameras, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 8406–8414

257. J. Chen, A. Huang, W. Gao, Y. Niu, T. Zhao, Joint shared-and-specific information for deep multi-view clustering. *IEEE Trans. Circuits Syst. Video Technol.* (2023)
258. H. Liu, H. Yuan, J. Hou, R. Hamzaoui, W. Gao, Pufa-gan: a frequency-aware generative adversarial network for 3d point cloud upsampling. *IEEE Trans. Image Process.* **31**, 7389–7402 (2022)
259. J. Xing, H. Yuan, C. Chen, W. Gao, Wiener filter-based color attribute quality enhancement for geometry-based point cloud compression, in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (IEEE, 2022), pp. 1208–1212
260. J.-X. Zhuang, X. Huang, Y. Yang, J. Chen, Y. Yu, W. Gao, G. Li, J. Chen, T. Zhang, Open-media: open-source medical image analysis toolbox and benchmark under heterogeneous ai computing platforms, in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)* (Springer, 2022), pp. 356–367
261. R. Cong, H. Yang, Q. Jiang, W. Gao, H. Li, C. Wang, Y. Zhao, S. Kwong, Bcs-net: boundary, context, and semantic for automatic covid-19 lung infection segmentation from ct images. *IEEE Trans. Instrum. Meas.* **71**, 1–11 (2022)
262. G. Yue, S. Li, T. Zhou, M. Wang, J. Du, Q. Jiang, W. Gao, T. Wang, J. Lv, Adaptive context exploration network for polyp segmentation in colonoscopy images. *IEEE Trans. Emerg. Top. Comput. Intell.* **7**(2), 487–499 (2022)
263. Y. Chen, X. Yu, S. Liu, W. Gao, G. Li, Zero-shot unsupervised image-to-image translation via exploiting semantic attributes. *Image Vis. Comput.* **124**, 104489 (2022)
264. D. Yang, W. Gao, G. Li, H. Yuan, J. Hou, S. Kwong, Exploiting manifold feature representation for efficient classification of 3d point clouds. *ACM Trans. Multimedia Comput. Commun. Appl.* **19**(1s), 1–21 (2023)
265. W. Zhao, X. Liu, Z. Zhong, J. Jiang, W. Gao, G. Li, X. Ji, Self-supervised arbitrary-scale point clouds upsampling via implicit neural representation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 1999–2007
266. H. Liu, H. Yuan, R. Hamzaoui, W. Gao, S. Li, Pu-refiner: a geometry refiner with adversarial learning for point cloud upsampling, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2270–2274
267. Z. Peng, Q. Jiang, F. Shao, W. Gao, W. Lin, Lggd+: image retargeting quality assessment by measuring local and global geometric distortions. *IEEE Trans. Circuits Syst. Video Technol.* **32**(6), 3422–3437 (2021)
268. J. Hu, Q. Jiang, R. Cong, W. Gao, F. Shao, Two-branch deep neural network for underwater image enhancement in hsv color space. *IEEE Signal Process. Lett.* **28**, 2152–2156 (2021)
269. Y. Mao, Q. Jiang, R. Cong, W. Gao, F. Shao, S. Kwong, Cross-modality fusion and progressive integration network for saliency prediction on stereoscopic 3d images. *IEEE Trans. Multimedia* **24**, 2435–2448 (2021)
270. Q. Jiang, W. Gao, S. Wang, G. Yue, F. Shao, Y.-S. Ho, S. Kwong, Blind image quality measurement by exploiting high-order statistics with deep dictionary encoding network. *IEEE Trans. Instrum. Meas.* **69**(10), 7398–7410 (2020)
271. M. Zhou, X. Wei, S. Wang, S. Kwong, C.-K. Fong, P.H. Wong, W.Y. Yuen, W. Gao, Ssim-based global optimization for ctu-level rate control in hevc. *IEEE Trans. Multimedia* **21**(8), 1921–1933 (2019)
272. Q. Jiang, F. Shao, W. Gao, H. Li, Y.-S. Ho, A risk-aware pairwise rank learning approach for visual discomfort prediction of stereoscopic 3d. *IEEE Signal Process. Lett.* **26**(11), 1588–1592 (2019)
273. Q. Jiang, F. Shao, W. Gao, Z. Chen, G. Jiang, Y.-S. Ho, Unified no-reference quality assessment of singly and multiply distorted stereoscopic images. *IEEE Trans. Image Process.* **28**(4), 1866–1881 (2018)
274. Y. Jia, S. Kwong, W. Wu, R. Wang, W. Gao, Sparse bayesian learning-based kernel poisson regression. *IEEE Trans. Cybern.* **49**(1), 56–68 (2017)
275. W. Gao, S. Kwong, H. Sang, Low-cost memory data scheduling method for reconfigurable fft bit-reversal circuits. *Electron. Lett.* **51**(3), 217–219 (2015)

276. W. Wu, S. Kwong, Y. Zhou, Y. Jia, W. Gao, Nonnegative matrix factorization with mixed hypergraph regularization for community detection. *Inf. Sci.* **435**, 263–281 (2018)
277. Y. Zhou, S. Kwong, W. Gao, X. Wang, A phase congruency based patch evaluator for complexity reduction in multi-dictionary based single-image super-resolution. *Inf. Sci.* **367**, 337–353 (2016)
278. Y. Zhou, S. Kwong, H. Guo, W. Gao, X. Wang, Bilevel optimization of block compressive sensing with perceptually nonlocal similarity. *Inf. Sci.* **360**, 1–20 (2016)
279. H. Yuan, S. Kwong, X. Wang, W. Gao, Y. Zhang, Rate distortion optimized inter-view frame level bit allocation method for mv-hevc. *IEEE Trans. Multimedia* **17**(12), 2134–2146 (2015)
280. Y. Jia, S. Kwong, W. Wu, W. Gao, R. Wang, Generalized relevance vector machine, in *Intelligent Systems Conference (IntelliSys)* (IEEE, 2017), pp. 638–645
281. Y. Zhou, S. Kwong, W. Gao, X. Zhang, X. Wang, Complexity reduction in multi-dictionary based single-image superresolution reconstruction via phase congruency, in *2015 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)* (IEEE, 2015), pp. 146–151
282. X. Wang, S. Kwong, W. Gao, Y. Zhou, H. Yuan, Y. Zhang, Smooth view quality oriented bit allocation optimization for 3d video coding, in *2015 IEEE International Conference on Systems, Man, and Cybernetics* (IEEE, 2015), pp. 1764–1769

Chapter 11

Open Source Projects for AI-based Image and Video Coding



11.1 Introduction

The significance of AI open source is multifaceted and profound, encompassing a range of economic, social, and technological aspects. Firstly, AI open source provides a platform for knowledge sharing and collaboration, enabling individuals and organizations to contribute to the development of AI technology. By openly sharing algorithms, models, and datasets, the AI community can benefit from collective intelligence and accelerate the pace of innovation. Through the transparent and open process of code development, contributors can easily access and understand the workings of AI models, thus increasing trust and adoption. Secondly, AI open source promotes fairness and inclusivity in the development of AI technology. By making AI resources more accessible to a wider range of users, it enables individuals and organizations with limited resources to benefit from AI technology. This helps to level the playing field, enabling everyone to participate in and benefit from the AI revolution. Thirdly, AI open source stimulates innovation and creates new opportunities for businesses and individuals. By openly sharing their research and development achievements, organizations can attract more attention and support from the community, increasing their influence and impact. In addition, open source projects often attract a large number of developers and users, who provide valuable feedback and suggestions for further improvement. This can help organizations to identify new market opportunities and product ideas, enabling them to innovate more effectively. Fourthly, AI open source contributes to the development of a more sustainable global economy. By enabling knowledge sharing and collaboration, it reduces redundant research and development efforts, leading to more efficient use of resources. This can help to reduce costs and increase the efficiency of the global economy. In addition, open source projects often involve participants from different countries and cultures, promoting cross-border cooperation and enhancing the global connectivity of the AI ecosystem. Finally, AI open source plays a crucial role in addressing important societal challenges. By enabling access to cutting-edge AI technology, it can help

to address pressing issues such as climate change, poverty, disease, and inequality. Open source projects can bring together diverse stakeholders, including governments, non-profit organizations, and communities, to collaborate on solutions that benefit everyone. This can promote sustainable development and enhance the well-being of humanity. It should be noted that, besides the tasks of quality assessment [1–10], quality enhancement [10–31], machine analysis [32–77], there have been already many different open source projects for diverse compression algorithms [65, 78–161].

To summarize, AI open source holds great promise for accelerating innovation, promoting fairness and inclusivity, stimulating economic development, and addressing important societal challenges. As the AI era progresses, maintaining an open and collaborative approach is essential to promote the development of a more just, inclusive, and sustainable future for all.

11.2 CompressAI

CompressAI is an open source library that provides various compression algorithms and techniques for data compression. CompressAI aims at encouraging more researchers to contribute to the learned image and video compression domain, by providing resources to research, implement and evaluate machine learning based compression codecs. The library provides a set of tools for data compression, including entropy coding, entropy decoding, and entropy modeling. It is widely used in various applications, such as video compression, audio compression, and data storage.

Content coding is a process that removes unnecessary redundancy in the data to be compressed, while entropy coding is a process that assigns short codes to symbols that occur more frequently and long codes to symbols that occur less frequently to achieve lossless compression. The CompressAI library provides a set of tools for content coding and entropy coding, including arithmetic coding, Huffman coding, and range coding. In addition, the library also provides functions for data simulation and evaluation of compressed data quality. It allows users to flexibly combine different compression algorithms and techniques to achieve better compression results.

The CompressAI library is widely used in various applications, such as video compression, audio compression, and data storage. For example, in video compression, the library can be used for entropy coding and entropy decoding of video data to achieve lossless compression. In audio compression, the library can be used for entropy coding and entropy decoding of audio data to achieve lossless compression. In data storage, the library can be used for compressing data to reduce storage space usage and transmission costs [162, 163].

CompressAI provides several end-to-end deep neural networks which are implemented on PyTorch. These models are available to be imported from the zoo library. The details of the models are listed in the Table 11.1. If you want to learn more about the CompressAI library or use it in your project, you can refer to the official website or other relevant resources for more information:

Table 11.1 Overview of models in compressAI

Model name	Model description
bmshj2018-factorized	Complete entropy model utilizing a factorized prior [164]
bmshj2018-hyperprior	Comprehensive hyperprior entropy model [164]
mbt2018-mean	Hyperprior model incorporating a Gaussian mixture [165]
mbt2018	Combines autoregressive processes with hyperprior [165]
cheng2020-anchor	Builds on mbt2018 adding residual blocks and sub-pixel deconvolution [166]
cheng2020-attn	Enhancement of cheng2020-anchor with inclusion of attention blocks [166]

Source Author

<https://interdigitalinc.github.io/CompressAI/intro.html>. Meanwhile, deep learning-based video compression models are considered to be included into CompressAI in the future.

11.3 OpenVVC

OpenVVC is an open source software designed for real-time decoding, fully compliant with the ITU-T H.266 / MPEG-I Part 3 VVC standard. Written from scratch in C, it provides efficient, low-power decoding capabilities across a range of operating systems, including macOS, Windows, Linux, and Android. It is optimized for low-power, real-time decoding of 4K versatile video coding (VVC) videos on Intel x86 and ARM platforms, making it a versatile solution for high-performance video processing.

OpenVVC supports decoding of 8-bit bitstreams and Frame + Tile parallelism. It can also be embedded into FFmpeg. One of the key advantages of OpenVVC is its cross-platform compatibility, supporting desktops, laptops, tablets, and smartphones across different operating systems. This ensures broad accessibility for users regardless of their device. The project is actively developed and maintained by a community of contributors, which fosters continuous improvement and innovation. You can refer to the official website or other relevant resources for more information: <https://github.com/OpenVVC/OpenVVC>.

11.4 OpenDIC

OpenDIC is an open source code library of deep learning-based image compression (DIC) constructed by Peking University. We collect DIC methods from recent researches, with MindSpore, PyTorch and TensorFlow frameworks.

In particular, MindSpore reproduction is done for all algorithms in this open source library. You can refer to the official website for more information: <https://openi.pcl.ac.cn/OpenDIC/OpenDIC>.

Table 11.2 will give an overview of the OpenDIC open source project. The algorithm names with their source publications, available frameworks and brief introductions are illustrated [83, 112].

11.5 OpenDMC

OpenDMC is a pioneering open source library dedicated to deep learning-based multi-frame coding (DMC) constructed by Peking University. It serves as a comprehensive repository of state-of-the-art DMC methods, offering a rich collection of source codes implemented in various leading frameworks such as MindSpore, PyTorch, and TensorFlow. This initiative not only facilitates access to cutting-edge DMC techniques but also provides a platform for performance evaluations, enabling a direct comparison of different methods under a unified benchmark. You can refer to the official website for more information: <https://openi.pcl.ac.cn/OpenDMC/OpenDMC>.

Table 11.3 gives an overview of the OpenDMC open source project. The algorithm names with their source publications, available frameworks and brief introductions are illustrated.

11.6 Summary

This chapter introduces open source projects for AI-based coding. First, the significance of open source AI is emphasized, highlighting the motivation for this chapter and the associated projects. A brief introduction to current open source projects on AI-based image and video coding is then provided, including popular compression algorithm libraries such as CompressAI and OpenVVC. The main focus of this chapter is the introduction of the open source projects for AI-based image and video coding, including OpenDIC and OpenDMC. These projects are presented with an overview of all algorithms.

Exercises

1. Please explain why open source AI is important.
2. Please briefly introduce the open source library CompressAI.
3. Please specify which aspects the open source project includes.

Table 11.2 Overview of OpenDIC

Index	Name and publication	Available frameworks	Brief introduction
1	Coarse2Fine, AAAI 2020, [167]	PyTorch and MindSpore	Coarse2Fine introduces a hierarchical strategy that implements a coarse-to-fine approach, integrating multiple layers of hyper-priors to facilitate comprehensive image analysis with the aim of effectively reducing spatial redundancy. This approach significantly enhances the rate-distortion performance in image compression
2	Cheng2020, CVPR 2020, [166]	PyTorch and MindSpore	Cheng2020 extends the application of hyperpriors from lossy to lossless compression and integrates an L2-norm term into the loss function to expedite the training process. Additionally, the research explores various parameterized models for latent codes, advocating for the utilization of Gaussian mixture likelihoods to create adaptable and versatile context models
3	Checkerboard-Context-Model, CVPR 2021, [168]	PyTorch and MindSpore	The dual-phase checkerboard context computation within the Checkerboard-Context-Model effectively addresses spatial constraints encountered during the decoding process by modifying the order of decoding operations. Empirical evidence indicates that this methodology accelerates the decoding process by over 40-fold, markedly improving computational efficiency while maintaining comparable rate-distortion performance
4	e2e_gdn, ICLR 2017, [169]	PyTorch and MindSpore	e2e_gdn presents an image compression approach that consists of three main stages: a nonlinear analysis transformation, a uniform quantizer, and a nonlinear synthesis transformation. This method employs convolutional linear filters along with nonlinear activation functions throughout the process
5	CAE, ICLR 2016, [170]	PyTorch and MindSpore	CAE presents a novel approach to optimizing autoencoders for lossy image compression
6	Imgcomp, CVPR 2018, [171]	TensorFlow and MindSpore	Imgcomp presents an innovative approach to managing the rate-distortion trade-off in image compression autoencoders. The key innovation lies in the direct modeling of the latent representation's entropy through the use of a context model. This context model is implemented as a 3D Convolutional Neural Network (3D-CNN) that is trained to learn a conditional probability model of the latent space distribution produced by the autoencoder. By doing so, Imgcomp is able to more effectively navigate the balance between compression rate and image quality, which is a critical aspect of image compression technology

(continued)

Table 11.2 (continued)

Index	Name and publication	Available frameworks	Brief introduction
7	bmsbj2018, ICLR 2018, [164]	PyTorch and MindSpore	bmsbj2018 uses variational autoencoders for image compression, outperforming other neural network methods. It delivers better visual quality, as shown by higher MS-SSIM and PSNR scores
8	iwave, TPAMI 2022, [172]	PyTorch and MindSpore	iwave presents a novel end-to-end image compression framework that employs a specialized, wavelet-inspired transform. This transform, known as iwave, processes images to produce a set of coefficients that retain all the original image information, enabling lossless compression

Source Author

Table 11.3 Overview of OpenDMC

Index	Name and publication	Available frameworks	Brief introduction
1	DCVC, NeurIPS 2021, [173]	PyTorch and MindSpore	DCVC introduces a deep contextual video compression framework aimed at shifting the paradigm from predictive coding to conditional coding. Specifically, it tackles key questions about the way to define, utilize, and learn conditions within a deep learning-based video compression framework
2	ssf2020, CVPR 2020, [174]	PyTorch and MindSpore	ssf2020 shows that a generalized warping operator can significantly simplify the model and training process while still achieving competitive compression results. This operator tackles challenges like disocclusions and fast motion effectively. It introduces scale-space flow, a scale parameter as an extension of optical flow. This addition boosts the network's capacity to handle indeterminacy, making it more robust
3	DVC, CVPR 2019, [175]	PyTorch and MindSpore	DVC combines the established frameworks of conventional video compression techniques with the robust nonlinear representational power of neural networks, introducing the inaugural end-to-end deep learning model designed for video compression. This model holistically optimizes every element within the compression process
4	DVC_P, VCIP 2021, [176]	TensorFlow and MindSpore	DVC_P expands on the foundation of OpenDVC by introducing perceptual enhancements. It integrates a discriminator network and employs a hybrid loss function to assist the model in achieving a balance between distortion, perceptual quality, and bit rate. Furthermore, it utilizes nearest-neighbor interpolation to mitigate checkerboard artifacts

(continued)

Table 11.3 (continued)

Index	Name and publication	Available frameworks	Brief introduction
5	HESIC, CVPR 2021, [177]	PyTorch and MindSpore	The HESIC framework employs a deep regression model to ascertain the homography matrix, facilitating an in-depth analysis of the mutual information found within stereo image pairs. For the optimization of entropy coding, it leverages two conditional stereo entropy models: a Gaussian mixture model alongside context-based entropy models. This combination allows for a precise depiction of the inter-image correlations, leading to a substantial decrease in the bitrates required for effective encoding
6	SASIC, CVPR 2022, [178]	PyTorch and MindSpore	SASIC introduces an innovative approach for compressing stereo images by exploiting the shared characteristics present in the overlapping perspectives of left and right images. The left image is compressed through a data-driven methodology that incorporates an autoencoder in conjunction with a hyperprior entropy model. Simultaneously, the encoding and decoding operations for the right image are guided by information derived from the already encoded left image
7	FRVC, ICIP 2022, [179]	PyTorch and MindSpore	FRVC introduces substantial improvements and innovative enhancements to the forefront of end-to-end optimized hierarchical bidirectional video compression, thereby advancing the prevailing paradigms in learned video compression. A noteworthy improvement is the integration of motion estimation and prediction modules, complemented by the compression of precisely adjusted residual motion vectors, collectively resulting in enhanced rate-distortion performance
8	DSIC, ICCV 2019, [180]	PyTorch and MindSpore	DSIC introduces an advanced technique for stereo image compression by exploiting the commonalities inherent in the overlapping fields of view of the left and right images. The compression of the left image is achieved through a learning-based approach employing an autoencoder, supplemented by a hyperprior entropy model. The processes of encoding and decoding the right image utilize the data derived from the pre-encoded left image. Additionally, a stereo attention module is incorporated to enhance the correlation between the left and right images during the decoding phase

Source Author

4. Please specify the deep learning framework reproduction that is a characteristic of OpenDIC and OpenDMC.
5. Please specify the image compression algorithms demonstrated in OpenDIC.
6. Please specify the video compression algorithms demonstrated in OpenDMC.
7. Please describe the future of the open source project.
8. Please provide a brief summary of the chapter.
9. Please describe the main features of the CompressAI library and explain how it supports research in AI-based image and video coding.
10. Please outline the advantages that OpenVVC provides as an open source project for video compression.

References

1. L. Xie, X. Mu, G. Li, W. Gao, et al., Pku-dpcc: a new dataset for dynamic point cloud compression. *APSIPA Trans. Signal Inf. Process.* **13**(6) (2024)
2. H. Yuan, W. Gao, W. Gao, Opensep: an open source subjective experiment platform, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 248–11 251
3. S. Fan, W. Gao, Screen-based 3d subjective experiment software, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9672–9675
4. J. Wang, W. Gao, G. Li, Zoom to perceive better: No-reference point cloud quality assessment via exploring effective multiscale feature. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
5. J. Wang, W. Gao, G. Li, Applying collaborative adversarial learning to blind point cloud quality measurement. *IEEE Trans. Instrum. Meas.* (2023)
6. H. Li, W. Gao, Information exploration of projected views for point cloud quality measurement. *IEEE Trans. Instrum. Meas.* (2024)
7. B. Qu, X. Liang, S. Sun, W. Gao, Exploring aigc video quality: a focus on visual harmony, video-text consistency and domain distribution gap, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2024)
8. B. Qu, H. Li, W. Gao, Bringing textual prompt to ai-generated image quality assessment, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024)
9. S. Sun, X. Liang, S. Fan, W. Gao, W. Gao, Ve-bench: subjective-aligned benchmark suite for text-driven video editing quality assessment, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
10. Z. Guo, W. Gao, H. Wang, J. Wang, S. Fan, No-reference deep quality assessment of compressed light field images, in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6
11. W. Liu, W. Gao, X. Mu, Fast inter-frame motion prediction for compressed dynamic point cloud attribute enhancement. *Proc. AAAI Conf. Artif. Intell.* **38**(4), 3720–3728 (2024)
12. Z. Yang, W. Gao, X. Lu, Danet: density-adaptive network for geometry-based point cloud compression artifacts removal, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
13. X. Fan, G. Li, D. Li, Y. Ren, W. Gao, T.H. Li, Deep geometry post-processing for decompressed point clouds, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
14. X. Zhang, G. Liao, W. Gao, G. Li, Tdnet: transformer-based dual-branch restoration network for geometry based point cloud compression artifacts in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6

15. Z. Li, G. Li, T.H. Li, S. Liu, W. Gao, Semantic point cloud upsampling. *IEEE Trans. Multimedia* (2022)
16. Z. Li, S. Liu, W. Gao, G. Li, G. Li, S4r: rethinking point cloud sampling via guiding upsampling-aware perception. *IEEE Trans. Multimedia* (2025)
17. R. Zhang, W. Gao, G. Li, T.H. Li, Qinet: decision surface learning and adversarial enhancement for quasi-immune completion of diverse corrupted point clouds. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–14 (2022)
18. R. Bao, Y. Ren, G. Li, W. Gao, S. Liu, Flow-based point cloud completion network with adversarial refinement, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2559–2563
19. J. Chen, G. Li, R. Zhang, T.H. Li, W. Gao, Pointvae: invertible variational autoencoder framework for 3d point cloud generation, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 3216–3220
20. R. Zhang, J. Chen, W. Gao, G. Li, T.H. Li, Pointot: interpretable geometry-inspired point cloud generative model via optimal transport. *IEEE Trans. Circuits Syst. Video Technol.* **32**(10), 6792–6806 (2022)
21. Z. Li, G. Li, T. Li, S. Liu, W. Gao, Information-growth attention network for image super-resolution, in *Proceedings of the 29th ACM International Conference on Multimedia* (2021), pp. 544–552
22. L. Zhou, W. Gao, G. Li, H. Yuan, T. Zhao, G. Yue, Disentangled feature distillation for light field super-resolution with degradations, in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2023), pp. 116–121
23. L. Zhou, W. Gao, G. Li, End-to-end spatial-angular light field super-resolution using parallax structure preservation strategy, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 3396–3400
24. W. Gao, L. Zhou, L. Tao, A fast view synthesis implementation method for light field applications. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* **17**(4), 1–20 (2021)
25. X. Zhang, W. Gao, G. Li, Q. Jiang, R. Cong, Image quality assessment-driven reinforcement learning for mixed distorted image restoration. *ACM Trans. Multimedia Comput. Commun. Appl.* **19**(1s), 1–23 (2023)
26. X. Zhang, W. Gao, H. Yuan, G. Li, Je 2 net: joint exploitation and exploration in reinforcement learning based image restoration, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2090–2094
27. X. Zhang, W. Gao, Hirl: hybrid image restoration based on hierarchical deep reinforcement learning via two-step analysis, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2445–2449
28. S. Wang, Y. Lu, W. Xia, P. Xia, Z. Wang, W. Gao, Light field angular super-resolution by view-specific queries. *Visual Comput.* 1–16 (2024)
29. Y. Sun, Z. Li, S. Wang, W. Gao, Depth-assisted calibration on learning-based factorization for a compressive light field display. *Opt. Express* **31**(4), 5399–5413 (2023)
30. Y. Sun, Z. Li, L. Li, S. Wang, W. Gao, Optimization of compressive light field display in dual-guided learning, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2075–2079
31. W. Gao, S. Fan, G. Li, W. Lin, A thorough benchmark and a new model for light field saliency detection. *IEEE Trans. Pattern Anal. Mach. Intell.* (2023)
32. S. Fan, W. Gao, G. Li, Salient object detection for point clouds, in *European Conference on Computer Vision* (Springer, 2022), pp. 1–19
33. S. Luo, W. Gao, A general framework for rotation invariant point cloud analysis, in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2024), pp. 3665–3669
34. X. Lu, W. Gao, Attentivenet: detecting small objects for lidar point clouds by attending to important points, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5

35. Z. Pan, N. Zhang, W. Gao, S. Liu, G. Li, Less is more: label recommendation for weakly supervised point cloud semantic segmentation. *Proc. AAAI Conf. Artif. Intell.* **38**(5), 4397–4405 (2024)
36. Z. Pan, G. Liu, W. Gao, T.H. Li, Epcontrast: effective point-level contrastive learning for large-scale point cloud understanding, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024), pp. 1–6
37. N. Zhang, Z. Pan, T.H. Li, W. Gao, G. Li, Improving graph representation for point cloud segmentation via attentive filtering, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 1244–1254
38. K. Wen, N. Zhang, G. Li, W. Gao, Mpvnn: multi-resolution point-voxel non-parametric network for 3d point cloud processing, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024), pp. 1–6
39. Z. Pan, N. Zhang, W. Gao, S. Liu, G. Li, Point cloud semantic segmentation with sparse and inhomogeneous annotations. in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
40. R. Zhang, G. Li, W. Gao, S. Liu, A quantum-inspired framework in master-servant mode for large-scale multi-modal place recognition. *IEEE Trans. Intell. Transp. Syst.* (2024)
41. D. Yang, W. Gao, Pointchd: a point cloud benchmark for congenital heart disease classification and segmentation. *IEEE J. Biomed. Health Inform.* (2024)
42. S. Fan, W. Gao, G. Li, Point-mpp: point cloud self-supervised learning from masked position prediction. *IEEE Trans. Neural Netw. Learn. Syst.* (2024)
43. Z. Pan, W. Gao, G. Li, Distribution guidance network for weakly supervised point cloud semantic segmentation, in *Advances in Neural Information Processing Systems* (2024)
44. R. Zhang, G. Li, W. Gao, T.H. Li, Compoint: can complex-valued representation benefit point cloud place recognition? *IEEE Trans. Intell. Transp. Syst.* (2024)
45. X. Zang, W. Gao, G. Li, H. Fang, C. Ban, Z. He, H. Sun, A baseline investigation: Transformer-based cross-view baseline for text-based person search, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 7737–7746
46. G. Liao, W. Gao, G. Li, J. Wang, S. Kwong, Cross-collaborative fusion-encoder network for robust rgb-thermal salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **32**(11), 7646–7661 (2022)
47. W. Gao, G. Liao, S. Ma, G. Li, Y. Liang, W. Lin, Unified information fusion network for multi-modal rgb-d and rgb-t salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **32**(4), 2091–2106 (2021)
48. Y. Chen, S. Sun, G. Li, W. Gao, T.H. Li, Closing the gap between theory and practice during alternating optimization for gans. *IEEE Trans. Neural Netw. Learn. Syst.* (2023)
49. Y. Chen, C. Jin, G. Li, T.H. Li, W. Gao, Mitigating label noise in gans via enhanced spectral normalization. *IEEE Trans. Circuits Syst. Video Technol.* (2023)
50. X. Zang, G. Li, W. Gao, Multidirection and multiscale pyramid in transformer for video-based pedestrian retrieval. *IEEE Trans. Ind. Inform.* **18**(12), 8776–8785 (2022)
51. X. Zang, G. Li, W. Gao, X. Shu, Learning to disentangle scenes for person re-identification. *Image Vis. Comput.* **116**, 104330 (2021)
52. X. Zang, G. Li, W. Gao, X. Shu, Exploiting robust unsupervised video person re-identification. *IET Image Process.* **16**(3), 729–741 (2022)
53. Z. Yue, G. Li, W. Gao, Cross-level guided attention for human-object interaction detection, in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2023), pp. 284–289
54. S. Sun, J. Liu, H. Li, G. Liu, T.H. Li, W. Gao, Streamflow: streamlined multi-frame optical flow estimation for video sequences, in *Advances in Neural Information Processing Systems* (2024)
55. Z. Yao, W. Gao, Iterative saliency aggregation and assignment network for efficient salient object detection in optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* (2024)
56. G. Liao, W. Gao, Rethinking feature mining for light field salient object detection. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(10), 1–24 (2024)

57. B. Yang, W. Gao, G. Li, Focus and adjust: progressive refinement network for human object interaction detection, in *2022 26th International Conference on Pattern Recognition (ICPR)* (IEEE, 2022), pp. 2546–2552
58. G. Liao, W. Gao, Q. Jiang, R. Wang, G. Li, Mmnet: multi-stage and multi-scale fusion network for rgb-d salient object detection, in *Proceedings of the 28th ACM International Conference on Multimedia* (2020), pp. 2436–2444
59. W. Gao, S. Kwong, Y. Zhou, Y. Jia, J. Zhang, W. Wu, Multiscale phase congruency analysis for image edge visual saliency detection, in *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 1 (IEEE, 2016), pp. 75–80
60. W. Gao, S. Kwong, Y. Zhou, X. Wang, Phase congruency analysis of down-sampled and blurring images for foreground extraction, in *2015 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)* (IEEE, 2015), pp. 152–157
61. Y. Guo, W. Gao, G. Li, Interpretable task-inspired adaptive filter pruning for neural networks under multiple constraints. *Int. J. Comput. Vis.* 1–17 (2024)
62. W. Gao, Y. Guo, S. Ma, G. Li, S. Kwong, Efficient neural network compression inspired by compressive sensing. *IEEE Trans. Neural Netw. Learn. Syst.* (2022)
63. Y. Guo, W. Gao, Semantic-driven automatic filter pruning for neural networks, in *IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
64. L. Tao, W. Gao, Efficient channel pruning based on architecture alignment and probability model bypassing, in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2021), pp. 3232–3237
65. Y. Li, Z. Yu, X. Chen, S. Wang, H. Yuan, W. Gao, Zero-shot quantization for large-kernels via shape-based distribution and diversity self-distillation, in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2025)
66. W. Gao, G. Li, *Deep Learning for 3D Point Clouds* (Springer Nature, 2024)
67. W. Gao, G. Li, Introduction to 3d point clouds: datasets and perception, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 1–27
68. W. Gao, G. Li, Learning basics for 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 29–70
69. W. Gao, G. Li, Deep-learning-based point cloud enhancement i, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 71–97
70. W. Gao, G. Li, Deep-learning-based point cloud enhancement ii, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 99–130
71. W. Gao, G. Li, Deep-learning-based point cloud analysis i, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 131–162
72. W. Gao, G. Li, Deep-learning-based point cloud analysis ii, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 163–193
73. W. Gao, G. Li, Point cloud pre-trained models and large models, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 195–225
74. W. Gao, G. Li, Point cloud-language multi-modal learning, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 227–254
75. W. Gao, G. Li, Open-source projects for 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 255–272
76. W. Gao, G. Li, Typical engineering applications of 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2025), pp. 273–299
77. W. Gao, G. Li, Future work on deep learning-based point cloud technologies, in *Deep Learning for 3D Point Clouds* (Springer, 2025), pp. 301–315
78. W. Gao, G. Li, H. Yuan, R. Hamzaoui, Z. Li, S. Liu, Apccpa'22: 1st international workshop on advances in point cloud compression, processing and analysis, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 7392–7393
79. W. Gao, G. Li, Point cloud compression, enhancement and applications: from 3d perception to large models, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 292–11 293

80. H. Sun, G. Li, S. Chen, L. Li, W. Gao, Special topic on 3d point cloud processing and applications. *ZTE Commun.* **21**(4), 1 (2023)
81. W. Gao, W. Gao, X. Mu, C. Peng, G. Li, Overview and comparison of avs point cloud compression standard. *APSIPA Trans. Signal Inf. Process.* (2024)
82. W. Gao, S. Wang, X. Zhang, S. Kwong, Editorial for special issue on deep learning-based data compression. *APSIPA Trans. Signal Inf. Process.* **13**(6) (2024)
83. W. Gao, H. Ye, G. Li, H. Zheng, Y. Wu, L. Xie, OpenPointCloud: an Open-source algorithm library of deep learning based point cloud compression, in *ACM International Conference on Multimedia* (2022), pp. 7347–7350
84. Y. Zhang, W. Gao, G. Li, Openpointcloud-v2: a deep learning based open-source algorithm library of point cloud processing, in *Proceedings of the 1st International Workshop on Advances in Point Cloud Compression, Processing and Analysis* (2022), pp. 51–55
85. L. Xie, W. Gao, Learningpcc: a pytorch library for learning-based point cloud compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 234–11 238
86. L. Xie, W. Gao, Pchmvision: an open-source library of point cloud compression for human and machine vision, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 239–11 243
87. T. Qin, G. Li, W. Gao, S. Liu, Multi-grained point cloud geometry compression via dual-model prediction with extended octree. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(9), 1–30 (2024)
88. Y. Shao, W. Gao, S. Liu, G. Li, Advanced patch-based affine motion estimation for dynamic point cloud geometry compression. *Sensors* **24**(10), 3142 (2024)
89. Y. Shao, F. Song, W. Gao, S. Liu, G. Li, Texture-guided graph transform optimization for point cloud attribute compression. *Appl. Sci.* **14**(10), 4094 (2024)
90. Y. Shao, X. Yang, W. Gao, S. Liu, G. Li, 3d point cloud attribute compression using diffusion-based texture-aware intra prediction. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
91. J. Zhang, Y. Chen, G. Liu, W. Gao, G. Li, Efficient point cloud attribute compression framework using attribute-guided graph fourier transform, in *ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2024), pp. 8426–8430
92. W. Gao, H. Yuan, G. Li, Z. Li, H. Yuan, Low complexity coding unit decision for video-based point cloud compression. *IEEE Trans. Image Process.* **33**, 149–162 (2023)
93. Y. Shao, G. Li, Q. Zhang, W. Gao, S. Liu, Nonrigid registration-based progressive motion compensation for point cloud geometry compression. *IEEE Trans. Geosci. Remote Sens.* (2023)
94. F. Song, G. Li, X. Yang, W. Gao, S. Liu, Block-adaptive point cloud attribute coding with region-aware optimized transform. *IEEE Trans. Circuits Syst. Video Technol.* (2023)
95. Y. An, Y. Shao, G. Li, W. Gao, S. Liu, A fast motion estimation method with hamming distance for lidar point cloud compression, in *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2022), pp. 1–5
96. X. Chen, Y. Wang, Y. Li, X. Ling, M. Li, R. Liu, M. Ouyang, K. Zhao, T. Guan, Y. He, Low bit-width zero-shot quantization with soft feature-infused hints for iot systems. *IEEE Internet Things J.* (2024)
97. H. Yuan, W. Gao, G. Li, Z. Li, Rate-distortion-guided learning approach with cross-projection information for v-pcc fast cu decision, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 3085–3093
98. F. Song, G. Li, W. Gao, T.H. Li, Rate-distortion optimized graph for point cloud attribute coding. *IEEE Signal Process. Lett.* **29**, 922–926 (2022)
99. F. Song, G. Li, X. Yang, W. Gao, T.H. Li, Fine-grained correlation representation for graph-based point cloud attribute compression, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
100. F. Shen, W. Gao, A rate control algorithm for video-based point cloud compression, in *2021 International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2021), pp. 1–5

101. F. Song, Y. Shao, W. Gao, H. Wang, T. Li, Layer-wise geometry aggregation framework for lossless lidar point cloud compression. *IEEE Trans. Circuits Syst. Video Technol.* **31**(12), 4603–4616 (2021)
102. Z. Cai, W. Gao, G. Li, W. Gao, Distortion propagation model-based v-pcc rate control for 3d point cloud broadcasting. *IEEE Trans. Broadcast.* 1–13 (2024)
103. X. Mu, W. Gao, H. Yuan, S. Wang, G. Li, Adaptive lpu decision for dynamic point cloud compression. *IEEE Signal Process. Lett.* (2024)
104. C. Peng, W. Gao, Laplacian matrix learning for point cloud attribute compression with ternary search-based adaptive block partition, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 10 412–10 420
105. Y. Wang, W. Gao, X. Mu, H. Yuan, Rate control optimization for joint geometry and attribute coding of lidar point clouds, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
106. T. Guo, H. Yuan, T. Wang, W. Gao, Graph filter-based fast motion matching for inter frame coding of mpeg g-pcc, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 1151–1155
107. L. Xie, W. Gao, H. Zheng, G. Li, Spcgc: scalable point cloud geometry compression for machine vision, in *2024 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2024), pp. 17 272–17 278
108. L. Xie, W. Gao, H. Zheng, H. Ye, Semantic-aware visual decomposition for point cloud geometry compression, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 595–595
109. Z. Qi, W. Gao, Variable-rate point cloud geometry compression based on feature adjustment and interpolation, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 63–72
110. Z. Yu, W. Gao, When dynamic neural network meets point cloud compression: computation-aware variable rate and checkerboard context, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 600–600
111. L. Xie, W. Gao, S. Fan, Z. Yao, Pdnet: parallel dual-branch network for point cloud geometry compression and analysis, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 596–596
112. L. Xie, W. Gao, H. Zheng, End-to-end point cloud geometry compression and analysis with sparse tensor, in *International Workshop on Advances in Point Cloud Compression, Processing and Analysis* (2022), pp. 27–32
113. C. Fu, G. Li, R. Song, W. Gao, S. Liu, Octattention: octree-based large-scale contexts model for point cloud compression. *Proc. AAAI Conf. Artif. Intell.* **36**(1), 625–633 (2022)
114. C. Zhang, W. Gao, Adapcc: adaptive rate control and rate-distortion-complexity optimization for dynamic point cloud compression, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
115. K. Wang, W. Gao, Unipcgc: towards practical point cloud geometry compression via an efficient unified approach, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
116. K. Zheng, W. Gao, H. Zheng, Octree-based learned point cloud geometry compression: a lossy perspective, in *2025 Data Compression Conference (DCC)* (IEEE, 2025)
117. Z. Qi, H. Liu, W. Gao, Variable-rate point cloud compression with channel-wise entropy modeling, in *2025 Data Compression Conference (DCC)* (IEEE, 2025)
118. H. Zheng, W. Gao, Z. Yu, T. Zhao, G. Li, Viewpcgc: view-guided learned point cloud geometry compression in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 7152–7161
119. L. Xie, W. Gao, H. Zheng, G. Li, Roi-guided point cloud geometry compression towards human and machine vision, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 3741–3750
120. G. Li, G. Wei, W. Gao, *Point Cloud Compression: Technologies and Standardization* (Springer Nature, 2024)
121. G. Li, W. Gao, W. Gao, Introduction, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 1–28

122. G. Li, W. Gao, W. Gao, Background knowledge, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 29–51
123. G. Li, W. Gao, W. Gao, Predictive coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 53–70
124. G. Li, W. Gao, W. Gao, Transform coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 71–96
125. G. Li, W. Gao, W. Gao, Quantization techniques, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 97–112
126. G. Li, W. Gao, W. Gao, Entropy coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 113–133
127. G. Li, W. Gao, W. Gao, Mpeg geometry-based point cloud compression (g-pcc) standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 135–165
128. G. Li, W. Gao, W. Gao, Avs point cloud compression standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 167–197
129. G. Li, W. Gao, W. Gao, Mpeg video-based point cloud compression (v-pcc) standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024) pp. 199–218
130. G. Li, W. Gao, W. Gao, Mpeg ai-based 3d graphics coding standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 219–241
131. G. Li, W. Gao, W. Gao, Future work, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 243–250
132. Y. Wu, L. Xie, S. Sun, W. Gao, Y. Yan, Adaptive intra period size for deep learning-based screen content video coding, in *IEEE International Conference on Multimedia and Expo Workshops* (2024), pp. 1–6
133. H. Zheng, W. Gao, End-to-end rgb-d image compression via exploiting channel-modality redundancy. *Proc. AAAI Conf. Artif. Intell.* **38**(7), 7562–7570 (2024)
134. L. Tao, W. Gao, G. Li, C. Zhang, Adanic: towards practical neural image compression via dynamic transform routing, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 16 879–16 888
135. Y. Wu, W. Gao, End-to-end lossless compression of high precision depth maps guided by pseudo-residual (2022). [arXiv:2201.03195](https://arxiv.org/abs/2201.03195)
136. Y. Wu, Z. Qi, H. Zheng, L. Tao, W. Gao, Deep image compression with latent optimization and piece-wise quantization approximation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 1926–1930
137. W. Gao, L. Tao, L. Zhou, D. Yang, X. Zhang, Z. Guo, Low-rate image compression with super-resolution learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2020), pp. 154–155
138. W. Gao, S. Sun, H. Zheng, Y. Wu, H. Ye, Y. Zhang, Opendmc: an open-source library and performance evaluation for deep-learning-based multi-frame compression, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9685–9688
139. W. Gao, H. Zheng, C. Zhang, K. Zheng, Z. Yu, Y. Li, H. Ye, Y. Zhang, Opendic: an open-source library and performance evaluation for deep-learning-based image compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 202–11 205
140. Y. Li, W. Gao, G. Li, S. Ma, Saliency segmentation oriented deep image compression with novel bit allocation. *IEEE Trans. Image Process.* **34**, 16–29 (2025)
141. C. Zhang, W. Gao, Learned rate control for frame-level adaptive neural video compression via dynamic neural network, in *European Conference on Computer Vision* (Springer, 2024)
142. Z. Yang, W. Gao, G. Li, Y. Yan, Sur-driven video coding rate control for jointly optimizing perceptual quality and buffer control. *IEEE Trans. Image Process.* (2023)
143. F. Shen, Z. Cai, W. Gao, An efficient rate control algorithm for intra frame coding in avs3, in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2021), pp. 3164–3169
144. H. Yuan, W. Gao, J. Wang, Dynamic computational resource allocation for fast inter frame coding in video conferencing applications, in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6

145. W. Gao, Q. Jiang, R. Wang, S. Ma, G. Li, S. Kwong, Consistent quality oriented rate control in hevc via balancing intra and inter frame coding. *IEEE Trans. Ind. Inform.* **18**(3), 1594–1604 (2021)
146. H. Yuan, W. Gao, A new coding unit partitioning mode for screen content video coding, in *Proceedings of the 2021 5th International Conference on Digital Signal Processing* (2021), pp. 66–72
147. W. Gao, On the performance evaluation of state-of-the-art rate control algorithms for practical video coding and transmission systems, in *Proceedings of the 2020 4th International Conference on Video and Image Processing* (2020), pp. 179–185
148. W. Gao, S. Kwong, Q. Jiang, C.-K. Fong, P.H. Wong, W.Y. Yuen, Data-driven rate control for rate-distortion optimization in hevc based on simplified effective initial qp learning. *IEEE Trans. Broadcast.* **65**(1), 94–108 (2018)
149. W. Gao, A multi-objective optimization perspective for joint consideration of video coding quality, in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (IEEE, 2019), pp. 986–991
150. W. Gao, S. Kwong, Y. Jia, Joint machine learning and game theory for rate control in high efficiency video coding. *IEEE Trans. Image Process.* **26**(12), 6074–6089 (2017)
151. W. Gao, S. Kwong, Y. Zhou, H. Yuan, Ssim-based game theory approach for rate-distortion optimized intra frame ctu-level bit allocation. *IEEE Trans. Multimedia* **18**(6), 988–999 (2016)
152. W. Gao, S. Kwong, H. Yuan, X. Wang, Dct coefficient distribution modeling and quality dependency analysis based frame-level bit allocation for hevc. *IEEE Trans. Circuits Syst. Video Technol.* **26**(1), 139–153 (2015)
153. W. Gao, S. Kwong, Phase congruency based edge saliency detection and rate control for perceptual image and video coding, in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2016), pp. 000 264–000 269
154. H. Yuan, W. Gao, Openfastvc: an open source library for video coding fast algorithm implementation, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9660–9663
155. H. Yuan, W. Gao, S. Ma, Y. Yan, Divide-and-conquer-based rdo-free cu partitioning for 8k video compression. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(4), 1–20 (2024)
156. L. Tao, W. Gao, A hardware implementation of entropy encoder for 8k video coding, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
157. Y. Guo, W. Gao, S. Ma, G. Li, Accelerating transform algorithm implementation for efficient intra coding of 8k uhd videos. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* **18**(4), 1–20 (2022)
158. Z. Cai, W. Gao, Efficient fast algorithm and parallel hardware architecture for intra prediction of avs3, in *IEEE International Symposium on Circuits and Systems (ISCAS)* (IEEE, 2021), pp. 1–5
159. W. Gao, H. Yuan, Y. Guo, L. Tao, Z. Cai, G. Li, Openhardwarevc: an open source library for 8k uhd video coding hardware implementation, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 7339–7342
160. W. Gao, H. Yuan, G. Liao, Z. Guo, J. Chen, Pp8k: a new dataset for 8k uhd video compression and processing. *IEEE MultiMedia* (2023)
161. W. Liu, W. Gao, G. Li, S. Ma, T. Zhao, H. Yuan, Enlarged motion-aware and frequency-aware network for compressed video artifact reduction. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
162. L. Xie, M. Hu, X. Bai, Towards hardware-friendly and robust facial landmark detection method, in *International Conference on Neural Information Processing* (Springer, 2022), pp. 432–444
163. L. Xie, M. Hu, X. Bai, Online improved vehicle tracking accuracy via unsupervised route generation, in *IEEE 34th International Conference on Tools with Artificial Intelligence* (IEEE, 2022), pp. 788–792
164. J. Ballé, D. Minnen, S. Singh, S.J. Hwang, N. Johnston, Variational image compression with a scale hyperprior, in *International Conference on Learning Representations* (2018)

165. D. Minnen, J. Ballé, G.D. Toderici, Joint autoregressive and hierarchical priors for learned image compression. *Adv. Neural Inf. Process. Syst.* **31**, 10 794–10 803 (2018)
166. Z. Cheng, H. Sun, M. Takeuchi, J. Katto, Learned image compression with discretized gaussian mixture likelihoods and attention modules, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 7939–7948
167. Y. Hu, W. Yang, J. Liu, Coarse-to-fine hyper-prior modeling for learned image compression. *Proc. AAAI Conf. Artif. Intell.* **34**(07), 11 013–11 020 (2020)
168. D. He, Y. Zheng, B. Sun, Y. Wang, H. Qin, Checkerboard context model for efficient learned image compression, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 14 771–14 780
169. J. Ballé, V. Laparra, E.P. Simoncelli, End-to-end optimized image compression, in *International Conference on Learning Representations* (2017)
170. L. Theis, W. Shi, A. Cunningham, F. Huszár, Lossy image compression with compressive autoencoders, in *International Conference on Learning Representations* (2016)
171. F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, L. Van Gool, Conditional probability models for deep image compression, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 4394–4402
172. H. Ma, D. Liu, N. Yan, H. Li, F. Wu, End-to-end optimized versatile image compression with wavelet-like transform. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(3), 1247–1263 (2020)
173. J. Li, B. Li, Y. Lu, Deep contextual video compression. *Adv. Neural Inf. Process. Syst.* **34**, 18 114–18 125 (2021)
174. E. Agustsson, D. Minnen, N. Johnston, J. Balle, S.J. Hwang, G. Toderici, Scale-space flow for end-to-end optimized video compression, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 8503–8512
175. G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, Z. Gao, Dvc: an end-to-end deep video compression framework, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 11 006–11 015
176. S. Zhang, M. Mrak, L. Herranz, M.G. Blanch, S. Wan, F. Yang, Dvc-p: deep video compression with perceptual optimizations, in *International Conference on Visual Communications and Image Processing* (2021), pp. 1–5
177. X. Deng, W. Yang, R. Yang, M. Xu, E. Liu, Q. Feng, R. Timofte, Deep homography for efficient stereo image compression, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 1492–1501
178. M. Wödlinger, J. Kotera, J. Xu, R. Sablatnig, Sasic: stereo image compression with latent shifts and stereo attention, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 661–670
179. E. Cetin, M.A. Yilmaz, A.M. Tekalp, Flexible-rate learned hierarchical bi-directional video compression with motion refinement and frame-level bit allocation, in *IEEE International Conference on Image Processing* (2022), pp. 1206–1210
180. J. Liu, S. Wang, R. Urtasun, Dsic: deep stereo image compression, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 3136–3145

Chapter 12

Future Works for AI-based Image and Video Coding



12.1 Introduction

The accelerated evolution of AI technologies has fundamentally reshaped the paradigms of visual data coding, encompassing both still images and video sequences. The AI developments introduce novel methods and technologies for image and video coding, which have outperformed traditional coding solutions. Deep learning has emerged as a powerful tool, enabling efficient compression algorithms, accurate quality assessment and perceptual modeling, and adaptability for diverse practical applications. This book delves into these innovations, exploring the development of cutting-edge AI-driven solutions for image and video coding.

The integration of neural networks into image and video compression frameworks can significantly increase rate-distortion performances. The learning-based paradigm can better optimize the human and machine perception, as well as the improvements of artifact removal and perceptual modeling. Furthermore, the advancements in 3D and multimodal coding has been evolved steadily, enabling handling of complex datasets. Due to the growing demands for real-time processing and energy-efficient solutions, the investigation further elucidates how the synergistic integration of cutting-edge technologies - particularly edge computing for distributed processing and quantum computing for exponential acceleration - can substantially augment the operational efficacy of neural codecs. Complementing these advancements, open source projects and standardization can foster collaborations and drive innovations in this field.

12.2 Future Work on Fundamentals for Deep Learning-based Image and Video Coding

The research prospects for deep encoding techniques are vast and dynamic, promising exciting developments in new architectures, theories, and methods [1–46]. Several key areas are expected to be the focus of future research, driven by advancements in neural networks, emerging technologies, and the growing need for efficient and interpretable AI models.

First, with the continuous evolution of neural networks, the emergence of novel architectures that offer more efficient, accurate, and robust encoding capabilities can be anticipated. These might include advanced variations of existing models like CNNs and RNNs, or entirely new paradigms that challenge current understanding. The fusion of deep learning-based coding with cutting-edge technologies such as quantum computing, edge computing, and IoT devices can revolutionize data processing particularly in terms of processing speed and energy efficiency.

Second, leveraging deep learning for more sophisticated compression algorithms, especially in fields like video compression and real-time streaming, can significantly enhance performance. This includes exploring lossless compression techniques that maintain data integrity while reducing size. The need for model interpretability grows as AI models become more complex. Developing explainable AI models for deep learning-based coding will be crucial, which is particularly important. Applying deep encoding techniques in diverse fields such as healthcare, autonomous vehicles, and smart cities could lead to breakthroughs in how data is processed and utilized.

Third, tailoring deep encoding models to specific types of data (e.g., medical images, astronomical data, etc.) can yield more efficient compression, meeting the unique needs of different scenarios. As the negative environmental impact of massive computing grows, creating energy-efficient deep learning models that can scale with increasing data sizes becomes imperative. Deepening the theoretical understanding of deep learning-based coding models will give rise to more principled and robust research approaches. Innovations in training methodologies, including unsupervised and semi-supervised techniques, can enhance the efficiency and effectiveness of deep learning-based coding. In the face of exponential data growth, models capable of efficiently processing and compressing massive datasets will become increasingly critical.

In conclusion, deep encoding is poised for significant advancements, driven by technological innovation and a deeper understanding of theoretical principles. Integrating these new developments promises to enhance the capabilities of data compression [47–84] and encoding far beyond current limits. Future research will likely focus on improving efficiency, interpretability, and scalability while exploring novel applications across various domains.

12.3 Future Work on Quality Assessment and Perception Models for Image and Video

Recent advances in visual perception research have highlighted critical gaps in current quality assessment methodologies, particularly in their ability to emulate human visual processing. As imaging technologies evolve toward higher complexity and diversity, there is an urgent need to develop more sophisticated assessment frameworks that can bridge the gap between computational models and biological vision systems. The fields of Image/Video Quality Assessment (IQA/VQA), Just-Noticeable-Distortion (JND) and Salient Object Detection (SOD) modeling [85–96] present critical opportunities for advancement by deepening their integration with human visual perception as follows.

Future IQA [97–106] systems can develop advanced Human Visual System (HVS) models that better simulate neural responses to distortions while improving cross-domain generalization across diverse content types and resolutions. Similarly, VQA systems require enhanced spatiotemporal modeling to capture motion artifacts and temporal inconsistencies, leveraging motion perception principles to align with human sensitivity.

JND modeling stands to benefit from refined masking effect quantification and neural correlates of perceptual thresholds, enabling more accurate prediction of imperceptible distortions. Practical applications should focus on integrating JND into compression and processing pipelines, optimizing bit allocation and computational efficiency while preserving perceived quality. Additionally, adaptive techniques can leverage saliency to prioritize visually important regions, reducing redundancy in less perceptible areas.

For SOD, a promising research direction lies in unifying SOD with quality assessment, creating hierarchical evaluation frameworks that couple detection confidence with perceptual metrics. Neural-inspired architectures could further bridge these domains by incorporating ventral and dorsal stream processing principles, enabling task-specific optimization. By advancing these interconnected areas, future systems can achieve more robust, efficient, and human-aligned quality assessment for emerging visual technologies.

By establishing stronger connections between computational models and neural visual processing, future quality assessment systems will not only achieve greater accuracy but also adapt more effectively to emerging display technologies and viewing conditions. Such advancements will ultimately enable more efficient, human-centric visual communication across diverse applications.

12.4 Future Work on Deep Learning-based Image Coding

As deep learning-based image coding approaches performance saturation [1–9, 22–26, 29–46], the field is shifting focus toward three key evolutionary directions: adaptability across diverse applications, computational efficiency improvements, and specialized compression for emerging use cases. This transition reflects the growing need for intelligent compression systems that can dynamically respond to varying content types, network conditions, and hardware constraints.

Future works can concentrate on these issues: (1) Content-based adaptive compression that automatically adjusts strategies based on scene complexity and image type (natural images, medical images, or remote sensing images). (2) Dynamic bitrate adaptation responding to real-time network bandwidth conditions and edge device capabilities. (3) Perception-driven optimization aligning with human visual perception and task-specific requirements like object detection. (4) Lightweight algorithm development for efficient deployment on IoT and edge computing platforms. These directions will be supported by continued improvements in model flexibility and computational efficiency.

The next generation of image compression systems will increasingly emphasize intelligent adaptation, balancing visual fidelity with practical constraints across diverse application scenarios. By integrating content-aware compression, network-aware bitrate control, and hardware-aware optimizations, future research can deliver versatile solutions that maintain high performance while addressing the complex demands of modern visual communication systems. This holistic approach promises to extend the boundaries of what's achievable in deep image compression beyond current saturation points.

12.5 Future Work on Deep Learning-based Video Coding

Recent years have witnessed remarkable progress in deep learning-based video coding [1–9, 22–26, 29–46], with neural network approaches demonstrating superior rate-distortion performance compared to traditional coding methods. While these techniques show particular promise in large-scale model applications, several critical challenges must be addressed before they can fully replace conventional video coding standards like H.265/H.266.

Three research directions emerge to advance the field: (1) Developing lightweight architectures to reduce algorithm complexity, enabling real-time execution on conventional CPUs and mobile processors without specialized GPU requirements. (2) Creating content-adaptive frameworks that enhance generalization capabilities across diverse video formats, with particular emphasis on handling challenging screen content videos through advanced domain adaptation techniques. (3) Establishing

interpretable neural coding schemes that incorporate explainable AI principles and hybrid syntax structures to ensure robust bitstream decoding while maintaining compression efficiency.

These research avenues collectively address the fundamental limitations in current neural video coding systems, paving the way for more practical, reliable, and universally applicable solutions that could potentially surpass traditional coding standards in both performance and versatility. Future work should particularly focus on developing hybrid approaches that combine the strengths of neural networks with established coding principles to achieve optimal balance between innovation and reliability.

12.6 Future Work on Deep Learning-based 3D Image Compression

With the rapid development of virtual reality (VR), augmented reality (AR), and autonomous driving, the demand and importance of 3D image compression technology is also increasing. In the future, the development trend of 3D image compression includes the following aspects.

Next-generation 3D image compression technologies are evolving along several key dimensions to address the growing demands of modern applications. First, at the algorithmic level, these systems combine enhanced conventional approaches with novel deep learning techniques to achieve superior rate-distortion performance, particularly in encoding complex spatial structures and textural features. Second, they incorporate intelligent adaptation mechanisms that dynamically optimize compression parameters based on specific application requirements and content characteristics. Third, to accommodate diverse usage scenarios, next-generation solutions feature multi-resolution capabilities that automatically adjust output quality to match display device specifications and network conditions while maintaining visual fidelity. Finally, for latency-sensitive applications like VR and autonomous driving, these technologies leverage specialized hardware accelerators (including AI chips and advanced GPUs) coupled with high-throughput storage architectures to achieve real-time processing with minimal delay. Together, these advancements form a comprehensive framework that addresses the full spectrum of technical challenges in modern 3D image compression.

These hardware-aware optimizations, combined with advancements in computational efficiency through novel algorithms, adaptive compression frameworks, multi-resolution support, and sub-frame latency performance, promise to deliver comprehensive improvements that meet the sophisticated demands of next-generation visual applications while significantly enhancing compression fidelity for 3D visual data. These technological advancements will fundamentally transform how we capture, store and visualize three-dimensional content across next-generation applications.

12.7 Future Work on Human and Machine Vision Oriented Image and Video Coding

Despite the significant advancements in human and machine perception oriented compression [1–9, 22–26, 29–46], several challenges remain. Current techniques largely focus on conventional computer vision tasks, while future research must address more complex scenarios and explore innovative approaches to enhance compression efficiency.

First, current compression techniques need to evolve beyond conventional computer vision tasks to address more complex applications [27, 107–139]. The integration of large-scale models with extensive prior information presents a particularly promising avenue for enhancing compression efficiency, especially in handling challenging low-bitrate content scenarios.

Second, cross-modal compression represents another significant frontier, where simultaneous encoding of image, video, audio, and textual data could revolutionize multimedia applications. This approach has the potential to dramatically improve performance in emerging applications such as video captioning and multimodal search systems. Furthermore, hybrid methodologies that strategically combine traditional feature engineering with data-driven deep learning architectures demonstrate considerable potential. These hybrid approaches can achieve optimal trade-offs between coding efficiency and cross-domain generalization in machine vision systems.

In conclusion, while perception-oriented compression has made substantial progress, significant opportunities exist in exploring complex task optimization, large-scale model integration, and cross-modal compression techniques. Future research should focus on developing hybrid approaches that leverage both traditional and deep learning methods to achieve superior compression performance across diverse applications and domains.

12.8 Future Work on Compression Artifacts Removal for Image and Video Coding

The development of image and video compression artifact reduction methods has evolved significantly alongside advancements in feature extraction paradigms [106, 140–160]. These methods have progressed through several key stages, including 2D convolution, 3D convolution, LSTM, deformable convolution, and attention mechanisms. Recently, diffusion-based models have emerged as a promising approach for compression artifact reduction.

First, the migration of diffusion-based models from denoising and generation tasks to other enhancement applications, such as super-resolution and frame interpolation, has opened new possibilities for compression artifact reduction. Compared to traditional CNN or transformer-based image restoration methods, diffusion-based approaches demonstrate superior subjective visual experience, particularly

in applications with low latency transmission requirements like real-time online conferencing systems, where facial clarity is paramount.

Second, the inherent nature of compression artifacts, which can be viewed as noise introduced by quantization parameters, makes the diffusion model's framework particularly suitable for compression distortion removal tasks. The diffusion and inference processes in these models can potentially be adapted to establish relationships between distorted images corresponding to adjacent quantization parameters during the inference stage. This approach could lead to more effective artifact reduction while maintaining computational efficiency.

In conclusion, diffusion-based models represent a significant advancement in compression artifact reduction, offering improved visual quality compared to traditional methods. Their ability to model the relationship between quantization-induced distortions and clean images presents a promising research direction. Future work should focus on optimizing these models for real-time applications while maintaining their superior subjective quality performance.

12.9 Future Work on AI-based Image and Video Coding Standards

The IEEE AI standard is actively exploring several core experiments to advance image compression technology, with plans to release reference software v7.0 in the near future. Concurrently, the JPEG AI and MPEG VCM standards are undergoing significant updates, while MPAI-EEV has made notable progress in neural video coding for drone applications.

First, the IEEE AI standard's future research agenda includes nine key experiments: (1) Fixed-point implementation for parsing; (2) Fixed-point optimization for main transform; (3) Entropy decoding acceleration techniques; (4) Enhanced Block architecture based on feedback mechanisms; (5) Integration of BEE, iWave and NIC frameworks; (6) Potential replacement of nonlocal attention with local attention in BEE/NIC; (7) Implementation of global hyperprior; (8) Simplification of main decoder structure; and (9) Block-wise encoding/decoding for UHD images. These investigations will continue to shape the development of effective image compression standards.

Second, the JPEG AI standard is progressing through its version updates, with completion targets set for 2025 (Version 1) and 2027 (Version 2). Similarly, MPEG VCM maintains an active development cycle, incorporating the five core experiments mentioned above. Meanwhile, MPAI-EEV has achieved significant milestones in learned video coding for drone footage, culminating in the initiation of neural video codec standardization. The group's current focus is developing the EEV-0.5 model during the latest meeting cycle.

In conclusion, the AI-based compression standardization landscape demonstrates vigorous activity across multiple organizations. The IEEE AI standard's upcoming

v7.0 release and continued experimentation, combined with JPEG AI's versioned updates and MPAI-EEV's neural codec standardization efforts, collectively represent significant advancements in media compression technology. These parallel developments promise to deliver more efficient and capable compression standards in the coming years.

12.10 Future Work on Implementations for AI-based Image and Video Coding

The field of AI-based image and video coding is poised for significant advancements through several key research directions. Current challenges in model compression and optimization present opportunities to develop more efficient neural network architectures that maintain accuracy while reducing computational demands, particularly for resource-constrained devices.

Three primary areas will drive progress in AI-based image and video coding, including: (1) Development of adaptive pruning algorithms that automatically optimize network architectures based on task-specific requirements and data characteristics; (2) Advancement of quantization techniques including integer quantization methods and layer-specific adaptive strategies to better balance model size and precision; (3) System-level optimizations through end-to-end learning approaches that jointly optimize coding models with complementary components like sensors and communication systems. Additionally, cross-modal lightweight approaches show particular promise for handling diverse data types such as combined image and depth information, while hardware-aware optimizations will ensure efficient deployment across heterogeneous computing environments.

These interconnected research directions will collectively enable more adaptable and efficient AI-based image and video coding methods. By pursuing intelligent model compression, comprehensive system optimization, and hardware-aware implementations, future work can achieve the dual objectives of maintaining high performance while meeting the stringent requirements of emerging applications and evolving computing platforms.

12.11 Future Work on Open Source Projects for AI-based Image and Video Coding

The rapid advancement of AI-based coding has spurred significant growth in open-source projects for visual media compression, spanning images, videos, 3D data, and point clouds. Researchers are actively replicating published works for validating non-open-sourced papers, translating existing open-source implementations across deep learning frameworks, benchmarking performance and visualizing results to ensure

accuracy. This open-source movement is critical as AI systems grapple with exponentially increasing data volumes, demanding more efficient compression algorithms capable of large-scale processing.

Future progress hinges on three fronts: (1) optimizing compression algorithms for speed and efficiency, (2) integrating AI compression with complementary technologies like machine learning pipelines, and (3) leveraging distributed computing through open-source models to overcome intensive training requirements. The community-driven approach enables researchers worldwide to pool computational resources, i.e., utilizing GPU/TPU clusters, while fostering interdisciplinary collaboration through shared models and codebases. This collective effort not only reduces development costs but also accelerates breakthroughs in advanced coding/decoding techniques for real-world applications like machine translation and conversational AI.

As the field matures, open-source frameworks will remain indispensable for driving innovation in AI-based compression. The synergy between algorithm optimization, cross-domain integration, and global collaboration through shared resources promises to deliver next-generation solutions that balance computational efficiency with superior compression performance. This collaborative paradigm positions the open-source community at the forefront of solving tomorrow's data compression challenges.

12.12 Future Work on Large Language Model-based Image and Video Coding

Recent research efforts [161–165] have demonstrated that large language models (LLMs) are not only powerful text generators but also effective data compressors, reinforcing the concept of “Generator is Compressor”. This dual capability of LLMs opens up exciting possibilities for their application in image compression. However, several challenges and limitations in current approaches must be addressed to fully realize their potential. Below, key directions for future work in LLM-based image compression are outlined.

First, while existing methods [166–168] leverage LLMs to assist in compression through textual descriptions of images, they primarily rely on traditional end-to-end training strategies for small-scale models. These approaches do not fully treat LLMs as compressors, limiting their ability to overcome performance constraints imposed by scaling laws [169]. Future work should focus on developing frameworks that directly utilize LLMs as compressors, enabling them to harness their full potential for image compression. Additionally, the accurate textual descriptions provided by LLMs can be further exploited to enhance compression performance, particularly in semantic-aware compression tasks.

Second, although [161, 170] position LLMs as direct compressors and explore their potential for cross-modal data compression through in-context learning without

parameter training, their applicability to image compression remains limited. Future research should investigate methods to bridge the gap between textual and visual domains, enabling LLMs to better understand and process image data. Techniques such as multimodal embeddings or pre trained vision-language models could be explored to enhance LLMs' ability to compress images effectively. Moreover, the powerful in-context learning capabilities of LLMs, which have already achieved state-of-the-art performance in lossless compression, should be further optimized for perceptual-lossy image compression.

Third, while [171] employs Low-Rank Adaptation (LoRA) [172] to fine-tune LLMs for image compression, their approach is constrained by a focus on ASCII character string compression, which limits its applicability to one-dimensional data formats. Future work should extend this approach to high-dimensional image data, developing more sophisticated mappings between visual data and LLM input representations. Although LoRA fine-tuning enhances LLMs' ability to comprehend semantic information in images, the method's relatively slow speed remains a challenge. Future research should focus on optimizing inference time and reducing computational overhead to make LLM-based image compression more practical for real-world applications.

Experimental results have shown that the aforementioned methods achieve superior image compression performance, highlighting the potential of LLMs in this domain. However, several challenges remain, including high inference time, large memory consumption, and the need for improved signal fidelity. Future research efforts should prioritize addressing these issues by exploring efficient optimization techniques, such as quantization, pruning, or distillation, to reduce computational demands while maintaining high compression performance. Additionally, the development of perceptual-lossy compression techniques that prioritize semantically important regions of images could further enhance compression efficiency and visual quality.

In conclusion, the integration of LLMs into image compression holds significant promise, but further research is needed to overcome existing limitations and fully exploit their capabilities. By addressing these challenges, future work can pave the way for innovative and efficient LLM-based image compression methods, advancing the field of data compression in the era of large-scale generative models.

12.13 Summary

This chapter explores the directions of future advancements in the field of AI-based image and video coding, emphasizing the transformative role of deep learning. Key areas include developing novel neural network architectures, integrating emerging technologies like quantum and edge computing, and improving quality assessment and perception models. It also highlights the importance of perception-driven compression techniques tailored for specific practical applications, such as human vision, machine vision, and real-time scenarios. Additionally, the developments of 3D and

multimodal coding frameworks, optimization in removing compression artifacts, achieving scalable and energy-efficient models are also future research focuses. It also addresses the evolution of coding standards and the pivotal role of open source projects in fostering innovations. By leveraging large-scale generative models and addressing existing challenges, the document outlines a roadmap for improving efficiency, quality, and adaptability in multimedia coding technologies across diverse applications.

Exercises

1. Please explain how deep learning is transforming image and video coding.
2. Please describe the role of quality assessment models in AI-based coding.
3. Please highlight the key challenges in deep learning-based video compression.
4. Please outline the future potential of AI in real-time image and video processing.
5. Please explain the significance of adaptive compression in deep learning.
6. Please describe the advancements in 3D and multimodal image compression.
7. Please discuss the importance of artifact removal in AI-based coding.
8. Please detail the role of open source projects in advancing AI-based image and video coding.
9. Please explain how large language models contribute to image compression.
10. Please summarize the future trends in AI-based image and video coding standards.

References

1. W. Gao, G. Li, H. Yuan, R. Hamzaoui, Z. Li, S. Liu, Apccpa'22: 1st international workshop on advances in point cloud compression, processing and analysis, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 7392–7393
2. W. Gao, G. Li, Point cloud compression, enhancement and applications: from 3d perception to large models, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 292–11 293
3. H. Sun, G. Li, S. Chen, L. Li, W. Gao, Special topic on 3d point cloud processing and applications. *ZTE Commun.* **21**(4), 1 (2023)
4. W. Gao, W. Gao, X. Mu, C. Peng, G. Li, Overview and comparison of avs point cloud compression standard. *APSIPA Trans. Signal Inf. Process.* (2024)
5. W. Gao, S. Wang, X. Zhang, S. Kwong, Editorial for special issue on deep learning-based data compression. *APSIPA Trans. Signal Inf. Process.* **13**(6) (2024)
6. W. Gao, H. Ye, G. Li, H. Zheng, Y. Wu, L. Xie, Openpointcloud: an open-source algorithm library of deep learning based point cloud compression, in *Proceedings of the 30th ACM international conference on multimedia* (2022), pp. 7347–7350
7. Y. Zhang, W. Gao, G. Li, Openpointcloud-v2: a deep learning based open-source algorithm library of point cloud processing, in *Proceedings of the 1st International Workshop on Advances in Point Cloud Compression, Processing and Analysis* (2022), pp. 51–55

8. L. Xie, W. Gao, Learningpcc: a pytorch library for learning-based point cloud compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 234–11 238
9. L. Xie, W. Gao, Pchmvision: An open-source library of point cloud compression for human and machine vision, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 239–11 243
10. W. Gao, G. Li, *Deep Learning for 3D Point Clouds* (Springer Nature, 2024)
11. W. Gao, G. Li, Introduction to 3d point clouds: datasets and perception, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 1–27
12. W. Gao, G. Li, Learning basics for 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 29–70
13. W. Gao, G. Li, Deep-learning-based point cloud enhancement i, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 71–97
14. W. Gao, G. Li, Deep-learning-based point cloud enhancement ii, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 99–130
15. W. Gao, G. Li, Deep-learning-based point cloud analysis i, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 131–162
16. W. Gao, G. Li, Deep-learning-based point cloud analysis ii, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 163–193
17. W. Gao, G. Li, Point cloud pre-trained models and large models, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 195–225
18. W. Gao, G. Li, Point cloud-language multi-modal learning, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 227–254
19. W. Gao, G. Li, Open-source projects for 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2024), pp. 255–272
20. W. Gao, G. Li, Typical engineering applications of 3d point clouds, in *Deep Learning for 3D Point Clouds* (Springer, 2025), pp. 273–299
21. W. Gao, G. Li, Future work on deep learning-based point cloud technologies, in *Deep Learning for 3D Point Clouds* (Springer, 2025), pp. 301–315
22. L. Xie, W. Gao, H. Zheng, G. Li, Spcgc: scalable point cloud geometry compression for machine vision, in *2024 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2024), pp. 17 272–17 278
23. L. Xie, W. Gao, H. Zheng, H. Ye, Semantic-aware visual decomposition for point cloud geometry compression, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 595–595
24. Z. Qi, W. Gao, Variable-rate point cloud geometry compression based on feature adjustment and interpolation, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 63–72
25. Z. Yu, W. Gao, When dynamic neural network meets point cloud compression: Computation-aware variable rate and checkerboard context, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 600–600
26. L. Xie, W. Gao, S. Fan, Z. Yao, Pdnet: parallel dual-branch network for point cloud geometry compression and analysis, in *Data Compression Conference (DCC)* (IEEE, 2024), pp. 596–596
27. Y. Li, Z. Yu, X. Chen, S. Wang, H. Yuan, W. Gao, Zero-shot quantization for large-kernels via shape-based distribution and diversity self-distillation, in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2025)
28. X. Chen, Y. Wang, Y. Li, X. Ling, M. Li, R. Liu, M. Ouyang, K. Zhao, T. Guan, Y. He, Low bit-width zero-shot quantization with soft feature-infused hints for iot system. *IEEE Internet Things J.* (2024)
29. L. Xie, W. Gao, H. Zheng, End-to-end point cloud geometry compression and analysis with sparse tensor, in *Proceedings of the 1st International Workshop on Advances in Point Cloud Compression, Processing and Analysis* (2022), pp. 27–32
30. C. Fu, G. Li, R. Song, W. Gao, S. Liu, Octattention: octree-based large-scale contexts model for point cloud compression. *Proc. AAAI Conf. Artif. Intell.* **36**(1), 625–633 (2022)

31. C. Zhang, W. Gao, Adapcc: adaptive rate control and rate-distortion-complexity optimization for dynamic point cloud compression, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
32. K. Wang, W. Gao, Unipcg: towards practical point cloud geometry compression via an efficient unified approach, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
33. K. Zheng, W. Gao, H. Zheng, Octree-based learned point cloud geometry compression: a lossy perspective, in *2025 Data Compression Conference (DCC)* (IEEE, 2025)
34. Z. Qi, H. Liu, W. Gao, Variable-rate point cloud compression with channel-wise entropy modeling, in *2025 Data Compression Conference (DCC)* (IEEE, 2025)
35. H. Zheng, W. Gao, Z. Yu, T. Zhao, G. Li, Viewpcgc: view-guided learned point cloud geometry compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 7152–7161
36. L. Xie, W. Gao, H. Zheng, G. Li, Roi-guided point cloud geometry compression towards human and machine vision, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 3741–3750
37. Y. Wu, L. Xie, S. Sun, W. Gao, Y. Yan, Adaptive intra period size for deep learning-based screen content video coding, in *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2024), pp. 1–6
38. H. Zheng, W. Gao, End-to-end rgb-d image compression via exploiting channel-modality redundancy, *Proc. AAAI Conf. Artif. Intell.* **38**(7), 7562–7570 (2024)
39. L. Tao, W. Gao, G. Li, C. Zhang, Adanic: towards practical neural image compression via dynamic transform routing, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 16 879–16 888
40. Y. Wu, W. Gao, End-to-end lossless compression of high precision depth maps guided by pseudo-residual (2022). [arXiv:2201.03195](https://arxiv.org/abs/2201.03195)
41. Y. Wu, Z. Qi, H. Zheng, L. Tao, W. Gao, Deep image compression with latent optimization and piece-wise quantization approximation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 1926–1930
42. W. Gao, L. Tao, L. Zhou, D. Yang, X. Zhang, Z. Guo, Low-rate image compression with super-resolution learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2020), pp. 154–155
43. W. Gao, S. Sun, H. Zheng, Y. Wu, H. Ye, Y. Zhang, Opendmc: an open-source library and performance evaluation for deep-learning-based multi-frame compression, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9685–9688
44. W. Gao, H. Zheng, C. Zhang, K. Zheng, Z. Yu, Y. Li, H. Ye, Y. Zhang, Opendic: an open-source library and performance evaluation for deep-learning-based image compression, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 202–11 205
45. Y. Li, W. Gao, G. Li, S. Ma, Saliency segmentation oriented deep image compression with novel bit allocation. *IEEE Trans. Image Process.* **34**, 16–29 (2025)
46. C. Zhang, W. Gao, Learned rate control for frame-level adaptive neural video compression via dynamic neural network, in *European Conference on Computer Vision* (Springer, 2024)
47. T. Qin, G. Li, W. Gao, S. Liu, Multi-grained point cloud geometry compression via dual-model prediction with extended octree. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(9), 1–30 (2024)
48. Y. Shao, W. Gao, S. Liu, G. Li, Advanced patch-based affine motion estimation for dynamic point cloud geometry compression. *Sensors* **24**(10), 3142 (2024)
49. Y. Shao, F. Song, W. Gao, S. Liu, G. Li, Texture-guided graph transform optimization for point cloud attribute compression. *Appl. Sci.* **14**(10), 4094 (2024)
50. Y. Shao, X. Yang, W. Gao, S. Liu, G. Li, 3d point cloud attribute compression using diffusion-based texture-aware intra prediction. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
51. W. Gao, H. Yuan, G. Li, Z. Li, H. Yuan, Low complexity coding unit decision for video-based point cloud compression. *IEEE Trans. Image Process.* **33**, 149–162 (2023)

52. Y. Shao, G. Li, Q. Zhang, W. Gao, S. Liu, Nonrigid registration-based progressive motion compensation for point cloud geometry compression. *IEEE Trans. Geosci. Remote Sens.* (2023)
53. F. Song, G. Li, X. Yang, W. Gao, S. Liu, Block-adaptive point cloud attribute coding with region-aware optimized transform. *IEEE Trans. Circuits Syst. Video Technol.* (2023)
54. Y. An, Y. Shao, G. Li, W. Gao, S. Liu, A fast motion estimation method with hamming distance for lidar point cloud compression, in *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2022), pp. 1–5
55. H. Yuan, W. Gao, G. Li, Z. Li, Rate-distortion-guided learning approach with cross-projection information for v-pcc fast cu decision, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 3085–3093
56. F. Song, G. Li, W. Gao, T.H. Li, Rate-distortion optimized graph for point cloud attribute coding. *IEEE Signal Process. Lett.* **29**, 922–926 (2022)
57. F. Song, G. Li, X. Yang, W. Gao, T.H. Li, Fine-grained correlation representation for graph-based point cloud attribute compression, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
58. F. Shen, W. Gao, A rate control algorithm for video-based point cloud compression, in *2021 International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2021), pp. 1–5
59. F. Song, Y. Shao, W. Gao, H. Wang, T. Li, Layer-wise geometry aggregation framework for lossless lidar point cloud compression. *IEEE Trans. Circuits Syst. Video Technol.* **31**(12), 4603–4616 (2021)
60. Z. Cai, W. Gao, G. Li, W. Gao, Distortion propagation model-based v-pcc rate control for 3d point cloud broadcasting. *IEEE Trans. Broadcast.* 1–13 (2024)
61. X. Mu, W. Gao, H. Yuan, S. Wang, G. Li, Adaptive lpu decision for dynamic point cloud compression. *IEEE Signal Process. Lett.* (2024)
62. C. Peng, W. Gao, Laplacian matrix learning for point cloud attribute compression with ternary search-based adaptive block partition, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 10 412–10 420
63. Y. Wang, W. Gao, X. Mu, H. Yuan, Rate control optimization for joint geometry and attribute coding of lidar point clouds, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
64. T. Guo, H. Yuan, T. Wang, W. Gao, Graph filter-based fast motion matching for inter frame coding of mpeg g-pcc, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 1151–1155
65. Z. Yang, W. Gao, G. Li, Y. Yan, Sur-driven video coding rate control for jointly optimizing perceptual quality and buffer control. *IEEE Trans. Image Process.* (2023)
66. F. Shen, Z. Cai, W. Gao, An efficient rate control algorithm for intra frame coding in avs3, in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2021), pp. 3164–3169
67. H. Yuan, W. Gao, J. Wang, Dynamic computational resource allocation for fast inter frame coding in video conferencing applications, in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6
68. W. Gao, Q. Jiang, R. Wang, S. Ma, G. Li, S. Kwong, Consistent quality oriented rate control in hevc via balancing intra and inter frame coding. *IEEE Trans. Ind. Inform.* **18**(3), 1594–1604 (2021)
69. H. Yuan, W. Gao, A new coding unit partitioning mode for screen content video coding, in *Proceedings of the 2021 5th International Conference on Digital Signal Processing* (2021), pp. 66–72
70. W. Gao, On the performance evaluation of state-of-the-art rate control algorithms for practical video coding and transmission systems, in *Proceedings of the 2020 4th International Conference on Video and Image Processing* (2020), pp. 179–185
71. W. Gao, S. Kwong, Q. Jiang, C.-K. Fong, P.H. Wong, W.Y. Yuen, Data-driven rate control for rate-distortion optimization in hevc based on simplified effective initial qp learning. *IEEE Trans. Broadcast.* **65**(1), 94–108 (2018)

72. W. Gao, A multi-objective optimization perspective for joint consideration of video coding quality, in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (IEEE, 2019), pp. 986–991
73. W. Gao, S. Kwong, Y. Jia, Joint machine learning and game theory for rate control in high efficiency video coding. *IEEE Trans. Image Process.* **26**(12), 6074–6089 (2017)
74. W. Gao, S. Kwong, Y. Zhou, H. Yuan, Ssim-based game theory approach for rate-distortion optimized intra frame ctu-level bit allocation. *IEEE Trans. Multimedia* **18**(6), 988–999 (2016)
75. W. Gao, S. Kwong, H. Yuan, X. Wang, Dct coefficient distribution modeling and quality dependency analysis based frame-level bit allocation for hevc. *IEEE Trans. Circuits Syst. Video Technol.* **26**(1), 139–153 (2015)
76. W. Gao, S. Kwong, Phase congruency based edge saliency detection and rate control for perceptual image and video coding, in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2016), pp. 000 264–000 269
77. H. Yuan, W. Gao, Openfastvc: an open source library for video coding fast algorithm implementation, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9660–9663
78. H. Yuan, W. Gao, S. Ma, Y. Yan, Divide-and-conquer-based rdo-free cu partitioning for 8k video compression. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(4), 1–20 (2024)
79. L. Tao, W. Gao, A hardware implementation of entropy encoder for 8k video coding, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
80. Y. Guo, W. Gao, S. Ma, G. Li, Accelerating transform algorithm implementation for efficient intra coding of 8k uhd videos. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* **18**(4), 1–20 (2022)
81. Z. Cai, W. Gao, Efficient fast algorithm and parallel hardware architecture for intra prediction of avs3, in *IEEE International Symposium on Circuits and Systems (ISCAS)* (IEEE, 2021), pp. 1–5
82. W. Gao, H. Yuan, Y. Guo, L. Tao, Z. Cai, G. Li, Openhardwarevc: an open source library for 8k uhd video coding hardware implementation, in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 7339–7342
83. W. Gao, H. Yuan, G. Liao, Z. Guo, J. Chen, Pp8k: a new dataset for 8k uhd video compression and processing. *IEEE MultiMedia* (2023)
84. W. Liu, W. Gao, G. Li, S. Ma, T. Zhao, H. Yuan, Enlarged motion-aware and frequency-aware network for compressed video artifact reduction. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
85. G. Li, G. Wei, W. Gao, *Point Cloud Compression: Technologies and Standardization* (Springer Nature, 2024)
86. G. Li, W. Gao, W. Gao, Introduction, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 1–28
87. G. Li, W. Gao, W. Gao, Background knowledge, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 29–51
88. G. Li, W. Gao, W. Gao, Predictive coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 53–70
89. G. Li, W. Gao, W. Gao, Transform coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 71–96
90. G. Li, W. Gao, W. Gao, Quantization techniques, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 97–112
91. G. Li, W. Gao, W. Gao, Entropy coding, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 113–133
92. G. Li, W. Gao, W. Gao, Mpeg geometry-based point cloud compression (g-pcc) standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 135–165
93. G. Li, W. Gao, W. Gao, Avs point cloud compression standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 167–197
94. G. Li, W. Gao, W. Gao, Mpeg video-based point cloud compression (v-pcc) standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 199–218

95. G. Li, W. Gao, W. Gao, Mpeg ai-based 3d graphics coding standard, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 219–241
96. G. Li, W. Gao, W. Gao, Future work, in *Point Cloud Compression: Technologies and Standardization* (Springer, 2024), pp. 243–250
97. L. Xie, X. Mu, G. Li, W. Gao et al., Pku-dpcc: a new dataset for dynamic point cloud compression. *APSIPA Trans. Signal Inf. Process.* **13**(6) (2024)
98. H. Yuan, W. Gao, W. Gao, Opensep: an open source subjective experiment platform, in *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 11 248–11 251
99. S. Fan, W. Gao, ‘creen-based 3d subjective experiment software, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 9672–9675
100. J. Wang, W. Gao, G. Li, Zoom to perceive better: No-reference point cloud quality assessment via exploring effective multiscale feature. *IEEE Trans. Circuits Syst. Video Technol.* (2024)
101. J. Wang, W. Gao, G. Li, Applying collaborative adversarial learning to blind point cloud quality measurement. *IEEE Trans. Instrum. Meas.* (2023)
102. H. Li, W. Gao, Information exploration of projected views for point cloud quality measurement. *IEEE Trans. Instrum. Meas.* (2024)
103. B. Qu, X. Liang, S. Sun, W. Gao, Exploring aigc video quality: a focus on visual harmony, video-text consistency and domain distribution gap, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2024)
104. B. Qu, H. Li, W. Gao, Bringing textual prompt to ai-generated image quality assessment, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024)
105. S. Sun, X. Liang, S. Fan, W. Gao, W. Gao, Ve-bench: subjective-aligned benchmark suite for text-driven video editing quality assessment, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
106. Z. Guo, W. Gao, H. Wang, J. Wang, S. Fan, No-reference deep quality assessment of compressed light field images, in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6
107. X. Zang, W. Gao, G. Li, H. Fang, C. Ban, Z. He, H. Sun, A baseline investigation: Transformer-based cross-view baseline for text-based person search, in *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 7737–7746
108. G. Liao, W. Gao, G. Li, J. Wang, S. Kwong, Cross-collaborative fusion-encoder network for robust rgb-thermal salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **32**(11), 7646–7661 (2022)
109. W. Gao, G. Liao, S. Ma, G. Li, Y. Liang, W. Lin, Unified information fusion network for multimodal rgb-d and rgb-t salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **32**(4), 2091–2106 (2021)
110. Y. Chen, S. Sun, G. Li, W. Gao, T.H. Li, Closing the gap between theory and practice during alternating optimization for gans. *IEEE Trans. Neural Netw. Learn. Syst.* (2023)
111. Y. Chen, C. Jin, G. Li, T.H. Li, W. Gao, Mitigating label noise in gans via enhanced spectral normalization. *IEEE Trans. Circuits Syst. Video Technol.* (2023)
112. X. Zang, G. Li, W. Gao, Multidirection and multiscale pyramid in transformer for video-based pedestrian retrieval. *IEEE Trans. Ind. Inform.* **18**(12), 8776–8785 (2022)
113. X. Zang, G. Li, W. Gao, X. Shu, Learning to disentangle scenes for person re-identification. *Image Vision Comput.* **116**, 104330 (2021)
114. X. Zang, G. Li, W. Gao, X. Shu, Exploiting robust unsupervised video person re-identification. *IET Image Process.* **16**(3), 729–741 (2022)
115. Z. Yue, G. Li, W. Gao, Cross-level guided attention for human-object interaction detection, in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2023), pp. 284–289
116. S. Sun, J. Liu, H. Li, G. Liu, T.H. Li, W. Gao, Streamflow: streamlined multi-frame optical flow estimation for video sequences, in *Advances in Neural Information Processing Systems* (2024)

117. Z. Yao, W. Gao, Iterative saliency aggregation and assignment network for efficient salient object detection in optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* (2024)
118. G. Liao, W. Gao, Rethinking feature mining for light field salient object detection. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(10), 1–24 (2024)
119. B. Yang, W. Gao, G. Li, Focus and adjust: Progressive refinement network for human object interaction detection, in *2022 26th International Conference on Pattern Recognition (ICPR)* (IEEE, 2022), pp. 2546–2552
120. G. Liao, W. Gao, Q. Jiang, R. Wang, G. Li, Mmnet: multi-stage and multi-scale fusion network for rgb-d salient object detection, in *Proceedings of the 28th ACM International Conference on Multimedia* (2020), pp. 2436–2444
121. W. Gao, S. Kwong, Y. Zhou, Y. Jia, J. Zhang, W. Wu, Multiscale phase congruency analysis for image edge visual saliency detection, in *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 1 (IEEE, 2016), pp. 75–80
122. W. Gao, S. Kwong, Y. Zhou, X. Wang, Phase congruency analysis of down-sampled and blurring images for foreground extraction, in *2015 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)* (IEEE, 2015), pp. 152–157
123. Y. Guo, W. Gao, G. Li, Interpretable task-inspired adaptive filter pruning for neural networks under multiple constraints. *Int. J. Comput. Vis.* 1–17 (2024)
124. W. Gao, Y. Guo, S. Ma, G. Li, S. Kwong, Efficient neural network compression inspired by compressive sensing. *IEEE Trans. Neural Netw. Learn. Syst.* (2022)
125. Y. Guo, W. Gao, Semantic-driven automatic filter pruning for neural networks, in *IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
126. L. Tao, W. Gao, Efficient channel pruning based on architecture alignment and probability model bypassing, in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE 2021), pp. 3232–3237
127. S. Fan, W. Gao, G. Li, Salient object detection for point clouds, in *European Conference on Computer Vision* (Springer, 2022), pp. 1–19
128. S. Luo, W. Gao, A general framework for rotation invariant point cloud analysis, in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2024), pp. 3665–3669
129. X. Lu, W. Gao, Attentionet: Detecting small objects for lidar point clouds by attending to important points, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
130. Z. Pan, N. Zhang, W. Gao, S. Liu, G. Li, Less is more: label recommendation for weakly supervised point cloud semantic segmentation. *Proc. AAAI Conf. Artif. Intell.* **38**(5), 4397–4405 (2024)
131. Z. Pan, G. Liu, W. Gao, T.H. Li, Epcontrast: effective point-level contrastive learning for large-scale point cloud understanding, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024), pp. 1–6
132. N. Zhang, Z. Pan, T.H. Li, W. Gao, G. Li, Improving graph representation for point cloud segmentation via attentive filtering, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 1244–1254
133. K. Wen, N. Zhang, G. Li, W. Gao, Mpvnn: multi-resolution point-voxel non-parametric network for 3d point cloud processing, in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024), pp. 1–6
134. Z. Pan, N. Zhang, W. Gao, S. Liu, G. Li, Point cloud semantic segmentation with sparse and inhomogeneous annotations, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025)
135. R. Zhang, G. Li, W. Gao, S. Liu, A quantum-inspired framework in master-servant mode for large-scale multi-modal place recognition. *IEEE Trans. Intell. Transp. Syst.* (2024)
136. D. Yang, W. Gao, Pointchd: a point cloud benchmark for congenital heart disease classification and segmentation. *IEEE J. Biomed. Health Inform.* (2024)
137. S. Fan, W. Gao, G. Li, Point-mpp: point cloud self-supervised learning from masked position prediction. *IEEE Trans. Neural Netw. Learn. Syst.* (2024)

138. Z. Pan, W. Gao, G. Li, Distribution guidance network for weakly supervised point cloud semantic segmentation, in *Advances in Neural Information Processing Systems* (2024)
139. R. Zhang, G. Li, W. Gao, T.H. Li, Compoinet: can complex-valued representation benefit point cloud place recognition? *IEEE Trans. Intell. Transp. Syst.* (2024)
140. Z. Li, G. Li, T. Li, S. Liu, W. Gao, Information-growth attention network for image super-resolution, in *Proceedings of the 29th ACM International Conference on Multimedia* (2021), pp. 544–552
141. L. Zhou, W. Gao, G. Li, H. Yuan, T. Zhao, G. Yue, Disentangled feature distillation for light field super-resolution with degradations, in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (IEEE, 2023), pp. 116–121
142. L. Zhou, W. Gao, G. Li, End-to-end spatial-angular light field super-resolution using parallax structure preservation strategy, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 3396–3400
143. W. Gao, L. Zhou, L. Tao, A fast view synthesis implementation method for light field applications. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* **17**(4), 1–20 (2021)
144. X. Zhang, W. Gao, G. Li, Q. Jiang, R. Cong, Image quality assessment-driven reinforcement learning for mixed distorted image restoration. *ACM Trans. Multimedia Comput. Commun. Appl.* **19**(1s), 1–23 (2023)
145. X. Zhang, W. Gao, H. Yuan, G. Li, Je 2 net: joint exploitation and exploration in reinforcement learning based image restoration, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2090–2094
146. X. Zhang, W. Gao, Hirl: hybrid image restoration based on hierarchical deep reinforcement learning via two-step analysis, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2445–2449
147. S. Wang, Y. Lu, W. Xia, P. Xia, Z. Wang, W. Gao, Light field angular super-resolution by view-specific queries. *Visual Comput.* 1–16 (2024)
148. Y. Sun, Z. Li, S. Wang, W. Gao, Depth-assisted calibration on learning-based factorization for a compressive light field display. *Opt. Express* **31**(4), 5399–5413 (2023)
149. Y. Sun, Z. Li, L. Li, S. Wang, W. Gao, Optimization of compressive light field display in dual-guided learning, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2075–2079
150. W. Gao, S. Fan, G. Li, W. Lin, A thorough benchmark and a new model for light field saliency detection. *IEEE Trans. Pattern Anal. Mach. Intell.* (2023)
151. W. Liu, W. Gao, X. Mu, Fast inter-frame motion prediction for compressed dynamic point cloud attribute enhancement. *Proc. AAAI Conf. Artif. Intell.* **38**(4), 3720–3728 (2024)
152. Z. Yang, W. Gao, X. Lu, Danet: density-adaptive network for geometry-based point cloud compression artifacts removal, in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2023), pp. 1–5
153. X. Fan, G. Li, D. Li, Y. Ren, W. Gao, T.H. Li, Deep geometry post-processing for decompressed point clouds, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
154. X. Zhang, G. Liao, W. Gao, G. Li, Tdrnet: transformer-based dual-branch restoration network for geometry based point cloud compression artifacts, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2022), pp. 1–6
155. Z. Li, G. Li, T.H. Li, S. Liu, W. Gao, Semantic point cloud upsampling. *IEEE Trans. Multimedia* (2022)
156. Z. Li, S. Liu, W. Gao, G. Li, G. Li, S4r: rethinking point cloud sampling via guiding upsampling-aware perception. *IEEE Trans. Multimedia* (2025)
157. R. Zhang, W. Gao, G. Li, T.H. Li, Qinet: decision surface learning and adversarial enhancement for quasi-immune completion of diverse corrupted point clouds. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–14 (2022)
158. R. Bao, Y. Ren, G. Li, W. Gao, S. Liu, Flow-based point cloud completion network with adversarial refinement, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 2559–2563

159. J. Chen, G. Li, R. Zhang, T.H. Li, W. Gao, Pointvae: invertible variational autoencoder framework for 3d point cloud generation, in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), pp. 3216–3220
160. R. Zhang, J. Chen, W. Gao, G. Li, T.H. Li, Pointot: interpretable geometry-inspired point cloud generative model via optimal transport. *IEEE Trans. Circuits Syst. Video Technol.* **32**(10), 6792–6806 (2022)
161. G. Delétang, A. Ruoss, P. Duquenne, E. Catt, T. Genewein, C. Mattern, J. Grau-Moya, L.K. Wenliang, M. Aitchison, L. Orseau, M. Hutter, J. Veness, Language modeling is compression, in *International Conference on Learning Representations* (2024)
162. Y. Li, Y. Guo, F. Guerin, C. Lin, Evaluating large language models for generalization and robustness via data compression (2024). [arXiv:2402.00861](https://arxiv.org/abs/2402.00861)
163. Y. Yu, S. Buchanan, D. Pai, T. Chu, Z. Wu, S. Tong, B. Haeffele, Y. Ma, White-box transformers via sparse rate reduction, in *Advances in Neural Information Processing Systems* (2023), pp. 9422–9457
164. C.S.K. Valmeekam, K. Narayanan, D. Kalathil, J.-F. Chamberland, S. Shakkottai, Llmzip: lossless text compression using large language models (2023). [arXiv:2306.04050](https://arxiv.org/abs/2306.04050)
165. Y. Ye, W. Gao, Llm-pcgc: large language model-based point cloud geometry compression (2024). [arXiv:2408.08682](https://arxiv.org/abs/2408.08682)
166. C. Li, G. Lu, D. Feng, H. Wu, Z. Zhang, X. Liu, G. Zhai, W. Lin, W. Zhang, Misc: ultra-low bitrate image semantic compression driven by large multimodal model (2024). [arXiv:2402.16749](https://arxiv.org/abs/2402.16749)
167. B. Li, S. Wang, S. Wang, Y. Ye, High efficiency image compression for large visual-language models (2024). [arXiv:2407.17060](https://arxiv.org/abs/2407.17060)
168. C. Li, X. Wu, H. Wu, D. Feng, Z. Zhang, G. Lu, X. Min, X. Liu, G. Zhai, W. Lin, Cmc-bench: towards a new paradigm of visual signal compression (2024). [arXiv:2406.09356](https://arxiv.org/abs/2406.09356)
169. J. Kaplan, S. McCandlish, T. Henighan, T.B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling laws for neural language models (2020). [arXiv:2001.08361](https://arxiv.org/abs/2001.08361)
170. D. Heurtel-Depeiges, A. Ruoss, J. Veness, T. Genewein, ‘ompression via pre-trained transformers: A study on byte-level multimodal data (2024). [arXiv:2410.05078](https://arxiv.org/abs/2410.05078)
171. K. Chen, P. Zhang, H. Liu, J. Liu, Y. Liu, J. Huang, S. Wang, H. Yan, H. Li, Large language models for lossless image compression: Next-pixel prediction in language space is all you need (2024). [arXiv:2411.12448](https://arxiv.org/abs/2411.12448)
172. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: low-rank adaptation of large language models, in *International Conference on Learning Representations* (2022)

Index

A

Adaptive filtering, 204
Adaptive quantization, 130, 238
Additive Gaussian noise, 58
Additive uniform noise, 104
Adversarial loss, 182
AI-based compression, 5
AI-based image compression, 226
AI-based video compression, 226
AI-driven compression techniques, 8
Algorithm complexity, 324
Amazon SageMaker, 288
Application-specific integrated circuits, 278, 280
Applying learning to image and video compression, 33
Arithmetic coding, 306
Attention mechanism, 26, 123, 205
Autoencoder, 29, 91, 309
Auto-Keras, 282
Automatic machine learning, 282
AutoML platforms, 282
Autonomous driving, 325
Autonomous vehicles, 322
Autoregressive model, 90
Auto-Sklearn, 282

B

Background knowledge, 32
BD-rate, 246
Bidirectional context transform module, 160

Bidirectional dense feature propagation, 211

Bidirectional temporal information, 132
Bit rate, 88
Bitrate allocation, 193
Bitstream parsing, 237
Blind/Referenceless Image Spatial QualityEvaluator, 53
Block-based neural codec, 231
Blocking artifacts, 204
Blurring effects, 204

C

Charbonnier loss, 213
Checkerboard context, 309
Compact latent representations, 91
CompressAI, 306, 308
Compression artifact reduction, 326
Compression artifacts, 203
Compression distortion removal, 327
Compression efficiency and quality, 256
Compression framework, 37
Conditional entropy coding, 137
Conditional entropy models, 158
Content-aware encoding, 193
Content-based adaptive compression, 324
Content layer, 99
Context information, 94
Context prediction and fusion, 163
Contrast sensitivity function, 59
Convolutional kernel, 24
Convolutional neural networks, 99

Correlation, 99
 Cross-modal data compression, 329
 Cross-modal redundancy, 162
 Cube-level context model, 234

D

Data simulation, 306
 Data storage, transmission, and processing, 8
 Deep learning, 13, 203
 Deep learning-based video coding, 130
 Deformable convolution, 203, 210, 326
 Denosing, 326
 Differences between AE and VAE, 31
 Diffusion-based models, 326
 Digital images, 1
 Discretized Gaussian mixture distribution, 105
 Display device, 325
 Distance Transformation, 64
 Double Stimulus Categorical Rating, 49
 Dual-layer hyperprior, 231
 Dynamic Computational Load Balancing, 284

E

Edge computing, 322
 Encoder-decoder architecture, 27
 End-to-end optimization, 137
 End-to-end optimized, 33
 End-to-end training, 329
 End-to-end video coding, 226
 Energy-efficient, 322
 Entropy coding, 23, 88, 130, 168, 230
 Entropy model, 35, 87, 155
 Evaluation criteria, 277
 Explainable AI models, 322

F

Feature distillation, 212
 Feature fusion, 186
 Feature interaction network, 65
 Feature matching loss, 183
 Feature reconstruction modules, 283
 Feedback mechanism, 327
 Field-programmable gate arrays, 278, 281
 Filter-level sparsity, 273

Fixation prediction, 61
 Fixed-point, 327
 Flow estimation, 127
 Frame interpolation, 326
 Frame rates, 6
 Frequency-domain, 59

G

Gaussian scale mixtures, 96
 Generalization, 324
 Generative adversarial networks, 37, 134, 166
 Generative compression, 37
 Generative image compression, 87
 Global dependencies, 102
 Global hyperprior, 327
 Google AutoML, 283
 Graphics Processing Units, 278
 Group of Pictures mechanism, 207

H

Healthcare, 322
 Hierarchical quality, 134
 Hierarchical quality layers, 213
 High coding efficiency, 60
 High-dimensional image data, 330
 Homography matrix, 160
 Huffman coding, 306
 Human sensitivity, 323
 Human visual system, 53
 Hyperprior, 93
 Hyperprior model, 36, 190

I

Image and video coding, 4, 8, 9, 12, 13
 Image and video video, 9
 Image coding for machine, 9
 Image compression, 4, 151, 184, 186, 203, 307
 Image quality assessment, 48
 Image restoration, 326
 Immersive media, 153
 Instance segmentation, 188
 Inter-frame coding, 123
 Inter-frame temporal features, 203
 Intermediate features, 276
 Intersection Over Union (IoU) loss, 66

Inter-view context, 158
Inter-view redundancy, 160
Intra-frame coding, 123
IoT and edge computing, 324
Iterative refinement, 127
iWave Transform, 238

J

Joint autoregressive and hierarchical priors, 36
Joint Photographic Experts Group, 226
JPEG AI standard, 247
Just-Noticeable-Distortion, 47, 56
Just Recognizable Distortion, 61

K

Kernel-level sparsity, 273
KL divergence, 88
Knowledge distillation, 276

L

Laplace Multiplicative Factor, 92
Large displacement motion, 209
Large-scale models, 193
Latent representation, 158
Layer normalization, 26
Learned Image Compression (LIC), 87, 90, 96, 100, 106, 109–111
Light field image compression, 165
Light field images, 151
Lightweight network, 283
Lightweight neural networks, 289
LLM-based image compression, 329
Local attention modules, 98
Local dependencies, 102
Loss function, 182
Lossless image compression, 87, 90
Lossy compression, 204
Lossy image compression, 87, 91, 110, 111
Lossy transform coding, 136
Luminance adaptation, 58

M

Machine vision, 186
Masked 3D CNN, 230
Mean opinion score, 49
Medical images, 322, 324
MindSpore, 308, 311
Mixed precision quantization, 275
Motion artifacts, 323

Motion compensation, 125, 129
Motion estimation, 125, 206, 311
Motion perception, 323
Motion trajectories, 126
Motion vectors, 125
MPEG VCM Standard, 250
Multi-frame video compression, 308
Multi-head attention, 68
Multi-head self-attention, 27
Multimodal compression framework, 164
Multimodal embeddings, 330
Multi-modal images, 151
Multi-scale mask convolution, 242
Multiscale pyramid models, 97
Multiscale Structural Similarity Index Measure (MS-SSIM), 310
Multi-view joint enhancement, 166

N

Natural images, 324
Natural scene statistics, 52
Network bandwidth, 324
Network pruning, 272
Neural Architecture Search, 271
Neural Image Coding, 226
Neural network compression, 227
Neural network light weight design, 271
Neural syntax layer, 99
Neural Transformation Unit, 101
Neural video compression, 136
Nonlinear transformations, 91
Non-local attention, 99
Non-local attention module, 227
Non-local means, 204
Non-uniform quantization, 167
No-reference quality assessment, 51

O

Objective quality, 207
OpenDIC, 307, 308
OpenDMC, 308
OpenDVC, 310
Open source, 307
Open source projects, 305
OpenVVC, 307, 308
Optical flow, 126, 206
Optical flow estimation, 125

P

Parallel context model, 245
Parameter skipping function, 158

Peak Signal-to-Noise Ratio (PSNR), 310
 Perceptual-lossy compression, 330
 Perceptual quality, 310
 Positional encoding, 27
 Principles and framework, 32
 Probability distribution estimation, 88
 Pyramid Constrained Self-Attention, 68
 PyTorch, 307, 311

Q

Quadtree-based Partitioning, 138
 Quality assessment, 9, 10, 13
 Quality enhancement, 207
 Quality enhancement module, 158
 Quality metrics, 249
 Quantization, 91, 274
 Quantization noise, 102
 Quantization parameters, 215, 327
 Quantum computing, 322

R

Range coding, 306
 Rate-distortion optimization, 311
 Rate-distortion performance, 36, 98, 324
 Real-time decoding, 307
 Real-time streaming, 322
 Recurrent neural networks, 99
 Reduced-Reference Entropic Differences, 52
 Reference frame, 127
 Region of interest, 100
 Reinforcement learning, 214
 Remote sensing images, 324
 Residual connections, 26
 Residual learning, 205
 Residual networks, 123
 RGB components, 2
 RGB-D image compression, 162
 Ringing artifacts, 204
 Robust encoding, 322

S

Salient object detection, 61, 62
 Scale-space flow, 310
 Scene complexity, 324
 Screen content videos, 324
 Search space, 277
 Search strategy, 277
 Selective generative compression, 37
 Self-attention, 205
 Semantic-aware compression, 329

Semantic coding, 192
 Similarity calculation, 127
 Smart cities, 322
 Sparse Coding, 204
 Sparsity Patterns, 277
 Spatial-angular decorrelation, 165
 Spatial attention, 212
 Spatial redundancy, 93
 Spatial-temporal features, 206
 Spatio-temporal correlation, 137
 Spatio-Temporal Cross-Covariance Transformer, 137
 Spatio-temporal model, 59
 Statistical models, 32
 Stereo image compression, 311
 Stereo images, 151
 Stereo video quality assessment, 55
 Structural Similarity Index, 51
 Sub-Aperture Images, 167
 Subjective and objective assessments, 9
 Subjective visual experience, 326
 Substitutional neural image compression, 231
 Super-resolution, 326
 Support vector machine, 53
 Swin Transformer Block, 101

T

Target Applications, 257
 Target frame, 127
 Temporal attention, 206
 Temporal-Foveated Masking, 60
 Temporal inconsistencies, 323
 Temporal redundancies, 130
 TensorFlow, 307
 Tensor Processing Units, 280
 Text semantic information, 161
 3D convolution, 326
 3D image technology, 151
 360-degree images, 151
 Transform and Quantization, 130
 Transformation, 154
 Transformer, 26
 Transmission overhead, 203
 2D convolution, 326

U

Unidirectional information, 132

V

Variable bitrate models, 233

Variational autoencoders, [30](#), [90](#), [92](#)
Variational image compression, [35](#)
VCM core experiment, [250](#)
Vector-level sparsity, [273](#)
Video coding for machine, [9](#)
Video compression, [6](#), [187](#), [203](#), [306](#)
Video quality assessment, [48](#)
Video saliency detection, [64](#)
Video salient object detection, [67](#)
Vision transformers, [99](#)
Visual fidelity, [324](#)
Visual saliency, [61](#)

Volumetric warping, [158](#)

W

Wavelet, [204](#)
Wavelet transform, [204](#)

Y

YUV sampling techniques, [2](#), [3](#)
YUV space, [247](#)
YUV video data format, [2](#)