中图法分类号:TN919.8　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-23

论文引用格式:Chen Tong, Lu Ming, Shi Junqi, Cong Wuyang, Ding Dandan, Jia Chuanmin, Liu Jiaying, Liu Dong, Song Li, Ma Siwei, Yang You, Liu Wenyu, Cao Xun, Ma Zhan. XXXX. A review and frontier perspectives on end-to-end learned image and video coding. Journal of Image and Graphics, XX(XX):0001-0023(陈彤, 陆明, 石峻奇, 丛吾洋, 丁丹丹, 贾川民, 刘家瑛, 刘东, 宋利, 马思伟, 杨铀, 刘文予, 曹汛, 马展. XXXX. 端到端智能图像视频编码的发展回顾与前沿展望. 中国图象图形学报, XX(XX):0001-0023)[DOI:10.11834/jig.250627]

端到端智能图像视频编码的发展回顾与前沿展望

陈彤[1],陆明[1],石峻奇[1],丛吾洋[1],丁丹丹[2],贾川民[3],刘家瑛[3],刘东[4],宋利[5],马思伟[3],杨铀[6],刘文予[6],曹汛[1],马展[1]

1. 南京大学,江苏省南京市　210093; 2. 杭州师范大学,浙江省杭州市,311121; 3. 北京大学,北京市　100871; 4. 中国科学技术大学,安徽省合肥市　230026; 5. 上海交通大学,上海市,200230; 6. 华中科技大学,湖北省武汉市,430074

摘　要:图像与视频编码及相应标准自诞生以来,一直支撑着点播、直播、视频会议等核心多媒体服务。过去三十余年,主流技术路线围绕规则驱动的模块化工具(如变换、预测、熵编码、环路滤波等)的精细化设计与协同优化展开,并借助标准化组织形成生态。近十年,随着深度学习表征能力、公共数据集累积、以及高效训练/推理框架的成熟,端到端智能编码技术快速迭代,在若干测试集与应用场景中展现出超越传统标准的压缩性能。本报告围绕图像编码,第一部分概述端到端智能编码主流框架演化主线;第二部分阐述率失真性能指标之外的可实用性功能,包括可变码率与码率控制、模型量化与鲁棒性;第三部分总结智能编码纳入/影响标准化进程的努力与现状;第四部分探讨从智能图像编码到智能视频编码的进一步拓展。希望本文能够为研究者与工程实践者提供系统化的思考视角,促进智能图像视频编码方法在产业级场景中的有序落地。

关键词:智能图像压缩;变分自编码器;率失真;实用性;标准化

A review and frontier perspectives on end-to-end learned image and video coding

Chen Tong[1], Lu Ming[1], Shi Junqi[1], Cong Wuyang[1], Ding Dandan[2], Jia Chuanmin[3], Liu Jiaying[3], Liu Dong[4], Song Li[5], Ma Siwei[3], Yang You[6], Liu Wenyu[6], Cao Xun[1], Ma Zhan[1]

1. Nanjing University, Nanjing, Jiangsu Province, 210093; 2. Hangzhou Normal University, Hangzhou, Zhejiang Province 311121; 3. Peking University, Beijing 100871; 4. University of Science and Technology of China, Hefei, Anhui Province 230026; 5. Shanghai Jiao Tong University, Shanghai 200230; 6. Huazhong University of Science and Technology, Wuhan, Hubei Province 430074

Abstract: For over three decades, image and video coding technologies and their associated international standards have served as the foundational compression engines underpinning core internet-scale multimedia services, ranging from on-demand streaming and live broadcasting to real-time video conferencing and social media sharing. Traditional approaches have predominantly followed a rule-driven, modular paradigm where carefully engineered components—such as intra and inter prediction, block-based transforms like DCT and DWT, scalar quantization, entropy coding, and in-loop filtering—are jointly optimized under classical Rate-Distortion (R-D) theory. This methodology, refined through successive genera-

收稿日期:2025-12-12;**修回日期**:2025-12-31

*通信作者:照片(一寸,300dpi以上)

基金项目:国家自然科学基金重点项目(批准号:62431011);国家自然科学基金(批准号:62501262);江苏省基础研究计划(自然科学基金)类青年项目(批准号:BK20251179)

Supported by: State Key Program of National Natural Science of China (Grant No. 62431011); National Natural Science Foundation of China (Grant No. 62501262); Natural Science Foundation of Jiangsu Province, China (Grant No. BK20251179)

tions of standards including JPEG, H. 26x, MPEG, and AVS, has achieved remarkable efficiency and interoperability through the coordinated efforts of standardization bodies. However, the past decade has witnessed a paradigm shift catalyzed by the rapid advancement of deep learning, leading to the emergence of end-to-end learned image and video compression. Empowered by expressive neural architectures, large-scale public datasets, and mature training ecosystems, end-to-end trainable systems have demonstrated R-D performance that consistently surpasses conventional codecs on benchmark datasets. These learned systems are primarily built upon Variational Autoencoders (VAEs), which replace the handcrafted rules of traditional pipelines with a unified differentiable framework. In this architecture, an encoder utilizes analysis transforms to map image data into compact latent representations, while a decoder employs synthesis transforms to reconstruct the image. Unlike linear transforms in traditional coding, these transforms leverage powerful neural networks, evolving from early convolutional neural networks (CNNs) to advanced architectures incorporating attention mechanisms, Transformers, and Mamba-based state-space models. A critical challenge in this framework is the non-differentiable nature of quantization. To enable end-to-end optimization, methods such as Additive Uniform Noise (AUN) are used during training to approximate quantization errors while maintaining differentiability, or Straight-Through Estimators (STE) are employed to pass gradients directly through quantization layers. While uniform quantization remains standard, recent advancements explore vector quantization and non-uniform quantization strategies to further refine feature representation. The core of compression efficiency in these systems lies in entropy modeling, which estimates the probability distribution of latent variables to minimize the bitrate. This field has evolved significantly from early factorized models that assumed statistical independence among latents. The introduction of the hyperprior structure, which utilizes auxiliary latent variables to model the spatial distribution parameters of the primary latents, marked a significant milestone in capturing dependencies. Subsequent innovations introduced autoregressive modeling, which predicts current features based on causal contexts in spatial or channel dimensions, further enhancing probability estimation accuracy. Most recently, hierarchical autoregressive models have been developed to capture global and local contexts in a coarse-to-fine manner, pushing the boundaries of feature compactness and coding efficiency. In parallel, the optimization objectives have expanded beyond pixel-level fidelity metrics like MSE and MS-SSIM to include perceptual metrics and adversarial losses, allowing for a trade-off between signal distortion and perceptual quality. Beyond theoretical performance, the transition of learned coding from academic exploration to industrial application requires addressing practical dimensions such as variable rate control, hardware efficiency, and robustness. To support variable bitrates within a single model, researchers have developed mechanisms involving multi-scale decomposition and feature modulation, where quality factors or maps scale latent variables or intermediate features. Rate control algorithms have also advanced, utilizing iterative search strategies or deep modeling of the rate-parameter relationship to meet specific bandwidth constraints. To facilitate deployment on commodity hardware, model quantization techniques, including Quantization-Aware Training (QAT) and Post-Training Quantization (PTQ), are employed to convert floating-point models into fixed-point integer operations, ensuring cross-platform consistency and reducing computational overhead. Furthermore, addressing the vulnerability of neural networks to perturbations, robust coding frameworks are being designed to defend against adversarial attacks and transmission errors through training regularization and input preprocessing. These technical advancements have culminated in the integration of learned compression into formal international standards. Two landmark standards have recently emerged: JPEG AI, ratified by ITU-T as T. 840. 1, and IEEE 1857. 11-2024. While both adopt VAE-based architectures, they differ in design philosophy. JPEG AI employs a multi-branch network design and emphasizes subjective quality and machine-task compatibility, optimizing for perception-oriented metrics like MS-SSIM and VMAF. In contrast, IEEE 1857. 11 focuses on objective gains in PSNR and MS-SSIM, offering tiered complexity profiles (Base, Main, High) to adapt to different computational capabilities. Both standards have established rigorous training and evaluation protocols, including the use of specific datasets like Kodak and dedicated robustness benchmarks, to ensure fair comparison and reproducibility. The principles of learned image coding have naturally extended to video coding, although with unique challenges in temporal modeling. The evolution of neural video coding can be categorized into three developmental phases. The first phase involved hybrid approaches that replaced specific modules, such as intra prediction, with learned networks while retaining the traditional motion-compensated residual coding framework. The second phase moved toward conditional inter-frame coding, utilizing learned optical flow networks for

motion estimation and warping to generate temporal contexts. The third and most recent phase represents a shift toward unified probabilistic frameworks that eliminate explicit motion estimation entirely. These systems leverage hierarchical spatial-temporal priors to perform joint intra and inter prediction within a single model, achieving performance that rivals or exceeds the latest H. 266/VVC standard while approaching real-time processing speeds on GPUs. Looking forward, the field is converging toward two major trends: task-aware coding and generative integration. Task-aware coding aims to support both human vision and machine perception from a single bitstream, aligning with the biological principle of "compression as intelligence" where compact representations facilitate diverse downstream cognitive tasks. Simultaneously, the integration of generative models, such as diffusion models and large multimodal models, is enabling ultra-low-bitrate reconstruction with high semantic fidelity, fundamentally altering the rate-distortion-perception trade-off. This report synthesizes these technical, practical, and standardization advances to provide a comprehensive perspective. Ultimately, the future of intelligent compression lies in establishing a new foundation for multimodal, task-agnostic, and semantically aware visual communication.

Key words: Neural Image Compression; Variational Autoencoder; Rate-Distortion; Practicality; Standardization

0 引　言

30多年以来，图像视频编码技术与相关标准，例如图像相关的JPEG（Joint Photographic Experts Group）（Hudson 等，2018）、JPEG 2000（Marcellin 等，2000）、HEIC（High Efficiency Image Container）（Hannuksela 等，2015）、视频相关的H. 26x 系列（Bross 等，2021；Sullivan 等，2005）、MPEG-x（Moving Picture Experts Group）系列、AVS（Audio Video coding Standard）系列（Zhang 等，2019），AVx 系列（Han 等，2021），构成了互联网海量多媒体业务（如社交网络、点播、互动直播、视频会议）的底层“压缩引擎”。在传统范式下，图像视频编码通过应用信号处理的物理规则来设计和优化工具集（时域/空域预测、块级变换、量化、熵编码与滤波等）实现率失真性能的提升和技术方案的迭代；同时通过不同的国际和国家标准组织制定相应的音视频编码压缩标准，来确保不同设备以及系统实现之间的互联互通。

近十年来，深度神经网络（Deep Neural Network，DNN）或深度网络以端到端或部分模块替换的方式融入传统的图像视频编码框架：从针对单一编码工具（如超分、滤波去噪、变换、帧内帧间预测）到全流程的端到端可学习编码系统。得益于深度网络的强大表征能力、相关训练数据的丰富与日益成熟的深度学习软硬件生态，基于深度网络的智能压缩方案在多个公开数据集上展现出优于传统音视频标准的压缩效率（Ding 等，2021；朱文武等，2022；贾川民等，2024）。然而，上述相关技术的蓬勃发展和热烈讨论仍主要集中于学术界。尽管智能图像编码方向已形成了两个国际标准JPEG AI（Ascenso 等，2023）和IEEE 1857. 11（IEEE Std 1857. 11-2024，2024），但工业界对将深度学习技术应用到实际图像视频编码场景仍存顾虑。其中，算法复杂度、跨平台部署难度以及场景应用的灵活适配度等等是焦点问题。

基于上述背景，本报告旨在通过系统性梳理端到端图像编码的技术体系与发展趋势，帮助读者形成全景认知与共识，以期促进相关技术和标准在产业场景中的规模化落地应用。报告结构如下：

1）**核心框架发展**：对比梳理端到端图像/视频编码的演进，并以代表性工作与性能对照揭示主体框架和核心技术的演进（分析-合成变换、熵编码概率建模等）。

2）**实用性考量**：构建多维指标体系，讨论压缩率失真外的其他度量指标，包括复杂度（如模型参数量）、跨平台部署难度、鲁棒性、实用性拓展（如码率控制）等。

3）**标准化进程**：回顾已开展的标准化进程并展望未来的发展，例如探讨生成大模型和智能图像视频编码融合的可能性、压缩域多任务的支持等。

此外，报告还对国内外发展路径进行了对比分析。值得一提的是，从2019到2021年陆续发表了三篇类似方向的相关综述（Ding 等，2021；Liu 等，2021；Ma 等，2020）。但他们主要聚焦于回顾与分析在传统音视频编码框架中引入深度网络做前后处理与工具模块替换，对于端到端压缩方案的探讨仍然较为不足。本报告则重点分析和讨论端到端学习的

编码方案，是对现有综述的重要补充。鉴于智能图像编码在技术发展和产业应用方面成熟度更高，本报告重点阐述图像端到端编码，同时简要介绍一些视频方面的延展。

1 国际研究现状

尽管理论上任何输入数据都可以通过矢量量化来编码，但对于高维图像视频信号，由于矢量量化的计算复杂度过高，在实践中难以实现（Ballé 等，2021）。现实可行的编码方案通常采用“变换编码”，借此将图像编码问题分解为变换（Transform）、量化（Quantization）与熵编码（Entropy Coding）等环节，分而治之。然后，通过联合优化整体率失真压缩性能来调优方案设计（Goyal，2001）。其中，常见的变换算法有离散余弦变换（Discrete Cosine Transform，DCT），离散小波变换（Discrete Wavelet Transform，DWT），整数变换（Integer Transform）等。变换操作也通常会和预测结合，例如JPEG对离散余弦变换后的直流系数进行预测；H. 264/AVC等先做空域或时域近邻预测，然后对预测残差做变换。流行的熵编码通常会利用内容上下文（Context）进行概率预测来构建高效的变长码（Variable length codes）或者算术码（Arithmetic codes）。

自上世纪中叶以来，研究者们主要着力应用物理规则和统计方法来设计和优化上述各个模块（如1952年提出的哈夫曼编码（Huffman，1952）和1974年发表的离散余弦变换（Ahmed 等，1974）。尽管早

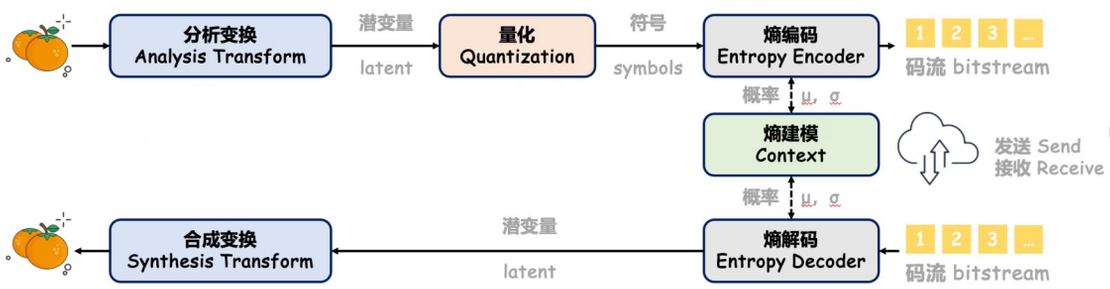

图 1 基于VAE的智能编码方案结构图

Fig. 1 VAE-based Learned Coding Architecture

在上世纪80年代末就有应用神经网络进行图像编码的尝试（Chua 等，1988），但一直到2016-2017年左右，端到端学习的智能编码才进入发展的快车道（Ballé 等，2017；Chen 等，2017；Toderici 等，2016）。

1.1 核心框架

当前最流行的端到端智能图像编码方案主要基于变分自编码器结构（Variational AutoEncoder，VAE）。该结构源自更早一点的自编码器（AutoEncoder）（Chen 等，2017；Toderici 等，2016）。如图1所示，类似传统的变换编码范式，VAE架构也是由“变换”，“量化”和“熵编码”组成。不同的是，这些核心组件都是通过堆叠神经网络来实现。

对于任意一个输入图像，编码器利用分析变换将其从像素域映射为高维的潜在表征（或潜在表示、潜变量等）$y = f_\theta(x)$。量化模块则将 y 映射为离散符号 $\hat{y}$ 后输入熵编码器进一步降低统计冗余。相应地，解码器利用合成变换从 $\hat{y}$ 重建图像 $\hat{x} = g_\theta(\hat{y})$。实现中，借助反向传播（Rumelhart 等，1986）来端到端训练VAE实现边际似然下界（Evidence Lower BOund，ELBO）的最大化。最大化ELBO可以简单理解为建立替代目标函数来权衡重建失真（Distortion）与潜在表征（Latent Representation）的紧凑性（码率）。

Ballé（2018）和Duan（2023）等研究者从理论上详细证明了不论是单层VAE还是多层VAE，最大化ELBO就是面向压缩的率失真（Rate - Distortion，R - D）优化（Ortega 等，1998）。应用中，率失真优化的目标是在给定码率（bits-per-pixel，bpp）约束下最小化失真，或在给定失真下最小化码率。引入拉格朗日乘积因子 λ，优化函数表达如下：

$$\min_{\theta,\psi} \mathcal{L}_{\mathrm{RD}} = E_{x\sim p_x}\left[\lambda R(\hat{y}) + D(x,\hat{x})\right],$$
$$\hat{y} = Q\left(f_\theta(x)\right),\ \hat{x} = g_\theta(\hat{y}). \tag{1}$$

式中 f_θ, g_θ 分别为编/解码器中的变换，$Q(\cdot)$ 为量化器，$R(\cdot)$ 为码率估计模型，参数为 ψ，$D(\cdot,\cdot)$ 为失真

度量。下文分别从变换、量化、熵建模（用以码率估计）和失真度量各个方面来介绍。

1.1.1　变换

VAE中通常堆叠多层神经网络来实现编码器和解码器里配对的“分析”-“合成”变换（Analysis-Synthesis Transform），也就是公式（1）中的 f_θ, g_θ。目标是通过神经网络强大的非线性拟合能力将图像数据 x 转化为更为紧致的潜在特征（或变量）表达 y，用以替换传统范式下的DCT、DWT等。

在发展的过程，变换模块中的基础网络层大部分聚焦卷积神经网络（Ballé等，2018；Minnen等，2018），注意力网络（Chen等，2021；Cheng等，2020）以及他们的组合（Liu等，2023）。其中注意力机制的拓展也包括应用创新的架构例如Transformer（Lu等，2022；Zhu等，2022）和Mamba（Zeng等，2025；王兴刚等，2025）。部分研究也尝试了图网络（Spadaro等，2024；Tang等，2022；Yang等，2021）、可逆网络（Ho等，2021；Xie等，2021）等等。实践中，基础网络模块的选择不仅仅考量压缩性能，更重要的是其带来的复杂度、功耗以及对底层计算平台的算子支持的需求。

前述工作主要还是关注单尺度的VAE结构，近期，Duan等人（2023，2024）提出堆叠多层VAE来实现对图像从粗粒度到细粒度的渐进式层次化条件建模。该方案不仅带来性能的提升，并且支持层内并行计算，避免了熵引擎中复杂的空域自回归。Lu等人（2024）也将其推广到视频编码，通过层次化时

空信息聚合，第一次实现了帧内、帧间编码在单一模型内的统一。

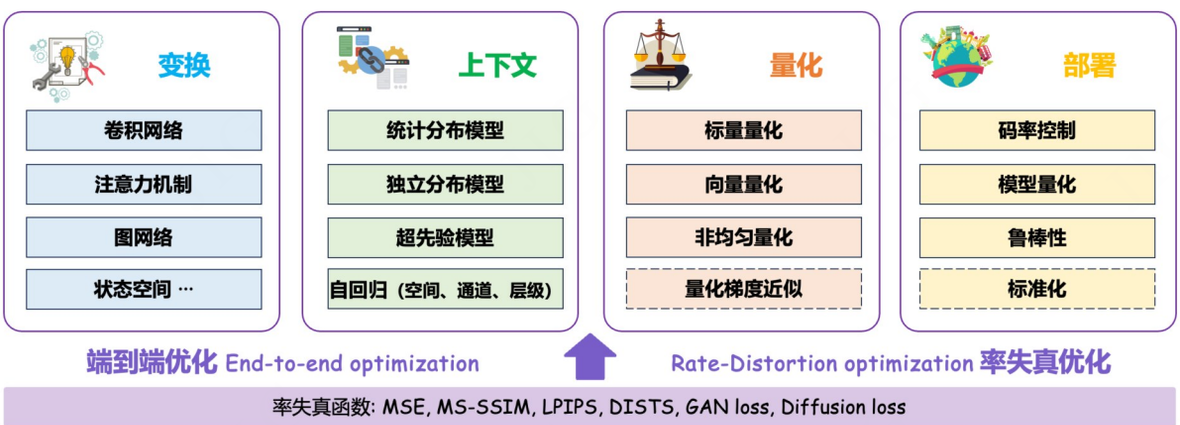

图2　端到端智能图像视频编码核心技术分解

Fig. 2　Core Techniques for End-to-End Learned Image and Video Coding

1.1.2　量化

端到端图像编码中涉及的量化主要有两个方面，包括模型量化和信号量化。用以支持跨设备跨平台一致性部署的模型量化在后续模型量化章节1.2.2中介绍。本小节探讨的量化操作是将传统编码中的变换系数（如DCT系数）、或智能编码中的潜变量从高精度浮点值映射到有限整数集合，便于后续的熵编码和整体率失真优化。

量化对于端到端学习的智能编码带来的挑战主要在于其不可微性：量化函数（如四舍五入）梯度几乎处处为零，无法通过反向传播训练；同时，还要解决如何保证可微量化的近似在训练和推理的一致性的问题。目前主流的量化方法大致可分为标量量化和矢量量化两类。

1）**标量量化（Scalar Quantization，SQ）**主要采用了如下设计

（1）加性均匀噪声近似（Additive Uniform Noise，AUN）是在训练阶段使用均匀分布噪声 $u \sim U(-0.5, 0.5)$ 来近似替代不可微的量化操作，即用 $\tilde{y} = y + u$ 代替 $\hat{y} = \mathrm{round}(y)$。该方法的核心优势在于其可微性（梯度恒为 $\frac{\delta \hat{y}}{\delta y} = 1$），从而支持端到端反向传播；同时，在假设潜在变量 y 服从连续概率分布的前提下，其熵（即码率）可解析计算，便于构建率失真优化目标。这一思想最早由Ballé等人（2016，2018）提出，成为后续大量基于网络的

端到端编码工作的基础。该方法的局限在于训练与推理过程不一致:训练时引入的是连续噪声,而实际部署时执行的是离散量化,这种不匹配可能导致模型性能下降或泛化能力受限。在此基础上,Choi等(2019)使用了一种基于universal quantization(Zamir等,1992)的优化实现方法,实验表明具有更好的收敛效果和测试性能。

(2)直通估计器(Straight-Through Estimator, STE)是另一种广泛采用的量化处理方法(Hinton, G.,2012)。其核心思想是在前向传播时使用真实的量化操作,即 $\hat{y}=\mathrm{round}(y)$,而在反向传播时则忽略量化函数的不可微性,直接令梯度穿过量化层,等效于设定 $\frac{\delta\hat{y}}{\delta y}=1$。其优点在于训练过程与实际推理完全一致,避免了因训练-推理不匹配带来的性能损失,同时实现极为简单。然而STE也存在明显局限:其梯度是粗略的近似,缺乏对量化误差的建模,容易导致训练不稳定,限制了模型性能的上限。

2)矢量量化(Vector Quantization, VQ)将高维潜向量整体映射到最近的离散码本中心(Van Den Oord等,2017),而非独立对每个标量分量进行量化。这样可以捕捉潜在特征之间的相关性,通常能更灵活地适应复杂数据分布。然而VQ需要训练可微的离散映射器,并且在高维空间中执行近邻搜索和码本学习开销很大。为兼顾训练便捷性和最终的离散表示,Agustsson等人(2017)提出的Soft-to-Hard VQ方法在训练过程中引入一个"软硬度"参数σ:训练初期使用可微的概率分配("软"分配)近似量化,然后随着σ逐渐减小,使分配趋于离散的最近邻选择。这种策略在保持端到端可微训练的同时最终实现了离散量化。后续工作(Lu等,2019; Zhang等,2023)在基于VQ的图像压缩上也进行了广泛的拓展。

上述工作主要是集中在均匀量化。非均匀量化(Non-uniform Quantization)策略可以对不同通道或不同空间位置的潜变量采用不同的量化间隔(步长),以更灵活地分配比特资源。一种称为乘积量化(Product Quantization)(El-Nouby等,2023)的标量量化变体支持对潜在变量的不同元素采用自适应的量化步长 q,其等效表示为 $q\cdot\mathrm{round}(\cdot/q)$。后续Ge等人(2024)在VQ基础上提出了上下文序列量化(Contextual Sequential Quantization, CSQ),利用内容上下文和图像纹理先验为每个潜变量向量分配自适应的量化中心,逐步离散化潜表示,从而在保持码率不变的情况下显著提升重建质量。这类非均匀量化方法根据数据分布调整量化参数,可降低统计冗余、提高率失真性能。

总的来说,加性均匀噪声、STE等简单的近似方法依然是目前智能图像编码的主流量化方法。而其它方法如矢量量化复杂度较高,优化难度大,但是能够保留更多的高维特征信息,因此在极低码率编码等特殊场景被广泛地采用(Careil等,2024;Esser等,2021;Ke等,2025)。

1.1.3 熵建模

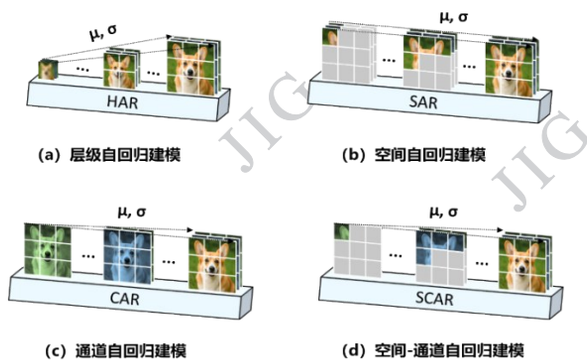

图3 上下文建模

Fig. 3 Context Modeling

熵建模(Entropy Modeling)的目标是对潜变量的概率分布进行精确建模,从而为后续的编码提供尽可能接近真实分布的概率估计。该过程直接决定了编码比特数的期望,即:

$$R=E_{y\sim q(y)}\left[-\log_2 p(y)\right] \tag{2}$$

式中 y 为编码器输出的潜变量,$p(y)$ 为估计的概率模型,$q(y)$ 为真实的潜变量分布。因此,如何构建一个高效且可泛化的概率估计器,成为智能图像压缩研究的核心问题之一。

统计建模阶段:在早期阶段,图像压缩仅仅依赖于对符号出现频率的离线统计来实现熵编码,如霍夫曼(Huffman)编码或算术编码(Arithmetic Coding)。Chen等(2017)提出的DeepCoder将Huffman编码应用于深度网络生成的特征,实现了"深度特征+传统熵模型"的结合。然而,这类方法无法自适应地建模数据分布中的上下文相关性,其压缩效率

受到明显限制。

随着深度学习的发展，熵建模逐渐从固定统计方法过渡到可学习的概率建模框架。总体上，其可分为两大组成部分：概率分布模型（distribution model）与上下文建模（context modeling），前者决定特征的概率分布特性，后者负责利用空间/通道/层次等条件信息提高概率估计的精度。

在概率分布模型方面，不同的分布假设直接影响熵模型的拟合能力与灵活性，常见选择包括：

1）均匀分布（Uniform Model）：最基础的假设，多用于可导训练的量化近似。

2）高斯分布（Gaussian Model）：最常用的形式，假设潜变量服从高斯分布，参数为均值与方差。高斯模型具有良好的可微性与训练稳定性，是绝大多数压缩框架的默认选择（如 Hyperprior2018（Ballé等，2018），JointAR2018（Minnen等，2018））. 公式如下，式中 μ 和 σ 分别代表高斯分布的均值和方差：

$$p(y|\mu,\sigma) = \mathcal{N}(y;\mu,\sigma^2) \tag{3}$$

3）拉普拉斯分布（Laplace Model）：相比上述的高斯分布尾部更尖锐（heavy-tailed），更适用于具有稀疏性或边缘结构的潜表示，在部分纹理细节较强的场景下表现更优（Zhou等，2019）. 公式如下，式中 μ 和 b 分别代表拉普拉斯分布的均值和尺度参数：

$$p(y|\mu,b) = \frac{1}{2b}\exp\left(-\frac{|y-\mu|}{b}\right) \tag{4}$$

4）混合高斯模型（Gaussian Mixture Model，GMM）：通过多分量（共K个）高斯的加权叠加刻画多模态特征分布（Cheng等，2020）。公式如下，式中 π_k 为混合权重，GMM能更准确地拟合复杂或多峰分布，在纹理丰富、非平稳图像中表现突出。

$$p(y) = \sum_{k=1}^{K}\pi_k\mathcal{N}(y;\mu_k,\sigma_k^2) \tag{5}$$

近年来，一些工作进一步探索广义高斯分布（Generalized Gaussian）（Zhang等，2025），以提高对非对称、多峰或稀疏特征分布的适应性。

在确定分布形式之后，提升建模精度的关键在于上下文依赖的建模策略。该方向的研究经历了从独立假设到条件与自回归建模的逐步演进。

1）**Factorized** 模型——独立假设下的端到端优化：随着端到端可学习压缩框架的发展，Ballé等人（2017）首次提出使用神经网络联合优化编码器、解码器与概率模型。其关键假设是潜变量各维度之间统计独立（factorized prior），即

$$p(\tilde{y}|x;\theta) = \prod_i \mathcal{U}(\tilde{y}_i;y_i,1)\ \text{with}\ y = f_\theta(x) \tag{6}$$

并将每个潜变量 y_i 建模为独立的均匀分布（通常卷积后的连续变量通过前述量化部分“量化为整数 + 均匀噪声近似”为 $\tilde{y}$ 实现可导训练）。这一模型相较于传统的固定统计编码，在端到端优化框架下显著提升了压缩性能。然而，由于完全独立的假设忽略了潜变量间的相关性，其建模能力仍然有限。

2）**Hyperprior** 模型——引入辅助潜变量的条件建模：为进一步建模潜变量之间的相关性，Ballé等人（2018）提出了著名的 Hyperprior 结构。该方法引入一个辅助潜变量 z，作为主潜变量 y 的条件变量（conditional prior），刻画其方差信息（即零均值高斯分布）：

$$p_{\tilde{y}|\tilde{z}}(\tilde{y}|\tilde{z},\theta_h) = \prod_i\left(\mathcal{N}(0,\tilde{\sigma}_i^{\,2})*\mathcal{U}\left(-\frac{1}{2},\frac{1}{2}\right)\right)(\tilde{y}_i)$$
$$\text{with}\ \tilde{\sigma} = g_h(\tilde{z};\theta_h). \tag{7}$$

式中，$\tilde{z}$ 由主潜变量 $\tilde{y}$ 经过一个独立的“超先验编解码器”g_h 生成，并通过另一个熵模型独立编码。由于 $\tilde{z}$ 的维度较小且假设服从可分解均匀分布，其额外码率开销较低，却能提供丰富的方差信息，显著提高了整体概率估计精度。这一结构奠定了现代学习式压缩系统的核心架构基础。

3）自回归建模——显式建模局部依赖：在 Hyperprior 框架中，潜变量间的关系通过全局统计建模实现，但局部依赖仍未被充分利用。为此，Minnen等人（2018）进一步将主潜变量建模为非零均值高斯分布，并首次将空间自回归机制引入熵编码概率建模中。

自此，自回归建模成为智能端到端压缩框架的标准组件。然而，自回归模型的主要瓶颈在于其严格的顺序依赖性，导致解码阶段难以实现并行计算。为缓解这一问题，研究者提出了多种分组与并行策略，如空间维度上的棋盘式（checkerboard）分组（He等，2021），通道维度上的均匀分组（Minnen等，2020）与非均匀分组（He等，2022），以及两者结合的混合方案（在每个通道组内进行局部空间预测），

以在率失真性能与解码速度之间取得平衡，这大体上可以分为四类：

1）空间自回归（Spatial Autoregressive）：沿空间邻域逐步预测潜变量（Minnen等，2018）。

2）通道自回归（Channel Autoregressive）：沿通道维度依序建模特征间依赖（Minnen等，2020）。

3）空间通道联合建模（Spatial-Channel Autoregressive）：同时建模空间与通道依赖，如基于3D卷积的联合预测（Chen等，2021），或在通道组内引入空间自回归机制（He等，2022）。

层次化自回归与多级先验：上述三类熵建模方法通常基于单层VAE结构。近年来，一类更具表达能力的层级自回归模型（Hierarchical Autoregressive Models）（Lu等，2024）被提出，其思想类似于多层级VAE（HVAE）：

$$p(\tilde{y}) = \prod_{l=1}^{L} p\left(\tilde{y}^{(l)} | \tilde{y}^{(<l)}\right) \tag{8}$$

式中，高层潜特征捕获全局上下文，低层潜特征描述细节信息，高层特征作为条件信息指导低层潜特征的概率预测。这种自上而下（top-down）的分层预测机制进一步提高了概率估计的灵活性与表达能力，在熵概率建模、解码速度、渐进式编码等方面展现出独特的优势。

总体来看，熵建模经历了从**固定统计 → 可学习分布 → 条件建模 → 自回归与层次建模**的演进过程。现代智能压缩系统往往结合多种建模策略，在压缩率、重建质量与解码速度之间实现平衡，并持续探索如何在保持率失真性能的同时，进一步提升建模泛化性与推理效率。

1.1.4 率失真损失函数

针对公式（1）定义的整体率失真优化，除去上节内容细化的码率估计，设计合适的失真损失函数也至关重要。

1） 客观保真优先

像素级均方误差MSE（Mean Squared Error）和结构相似度MS-SSIM（Wang等，2003）（Multiscale Structural Similarity）是端到端图像压缩广泛采用的失真函数。

在高斯失真假设下，MSE与经典率失真理论下与码率存在清晰的解析联系，优化稳定、实现简单，因而被广泛采用：

$$D_{\mathrm{MSE}}(x,\hat{x}) = \frac{1}{N}\left\| x - \hat{x} \right\|_2^2 \tag{9}$$

相比较MSE，结构相似度（Wang等，2004）SSIM更贴近人眼对结构的敏感性。其从亮度l、对比度c、结构s三个分量通过α,β,γ进行局部加权比较并乘性组合：

$$\mathrm{SSIM}(x,\hat{x}) = \left[l(x,\hat{x})\right]^{\alpha} \cdot \left[c(x,\hat{x})\right]^{\beta} \cdot \left[s(x,\hat{x})\right]^{\gamma} \tag{10}$$

SSIM的多尺度版本MS-SSIM在信号的金字塔分解表征的不同尺度$(j = 1, \cdots, M)$上进行聚合（通常仅最低尺度M保留亮度分量）。训练中失真项通常表示为$D_{\mathrm{MS-SSIM}} = 1 - \mathrm{MS-SSIM}$或$D_{\mathrm{MS-SSIM}} = -\log(\mathrm{MS-SSIM})$。

$$\mathrm{MS-SSIM}(x,\hat{x}) = l_M^{\alpha_M} \prod_{j=1}^{M} c_j^{\beta_j} s_j^{\gamma_j} \tag{11}$$

2）从“客观失真最小化”到“主观感知最优”随着“深特征更贴合人类主观偏好”的证据累积，Zhang等人（2018）提出了LPIPS，通过计算预训练网络的多层特征差来感知距离。与传统的像素级比较不同，LPIPS衡量的是感知空间中的表征差异，因此与人类主观偏好具有更高的一致性，并迅速成为训练与评估中的通用感知度量。基于这一方法，研究者开始探索将编码优化目标从“像素保真”转向“感知一致”。

Agustsson等（2019）和Huang等（2019）率先将感知损失以及对抗生成网络（Generative Adversarial Network，GAN）引入训练目标中，以牺牲像素保真为代价换取清晰纹理，使模型能够在仅<0.08 bpp下合成视觉上更自然的细节，显著减少块效应与过平滑伪影。但其限制也很突出：由于缺乏足够比特用于精确描述微结构，重建结果往往只保持高层语义，在细节层面与输入图像出现明显偏差。

随后，Blau和Michaeli（2019）给出了这一现象的理论化解释：存在一个三元“率-失真-感知”权衡。在固定码率下，更好的感知质量必然意味着更差的（像素/相似度意义上的）失真；反之，仅最小化失真会导致较差的感知质量。他们将“失真”形式化为两幅图像间的相似度度量，将“感知质量”形式化为真实图像分布p_x与重建分布$\hat{p}_x$的分布距离（如散度）。该结果既解释了Agustsson等的经验观察，也明确了优化目标的边界：要提升主观真实感，必须引入能缩小分布差异的机制（如GAN），从而在“牺牲部分像素保真”的前提下换取“更逼真”的观感。

在此理论指引下，HiFiC（Mentzer等，2020）将这一路线系统化：把感知损失（如LPIPS）与对抗损失（GAN）显式并入R－D目标，并与熵模型/编解码器联合训练，实现率－失真－感知三者的显式加权与可控权衡。与仅优化像素失真的方法相比，HiFiC在用户研究与感知指标（如FID（Heusel等，2017）/ KID（Bińkowski等，2018）/ LPIPS（Zhang等，2018）/ NIQE（Mittal等，2013））上系统性更优，同时比早期极低码率的生成式方案在内容忠实度上更可控、更平衡。

总体来看，率失真优化的“失真项”经历了从像素保真/结构一致再到感知/分布匹配的递进，并且失真设计不再追求单一指标最优，而是在“像素保真—感知真实”二者间做策略化取舍，并通过对抗学习提供的分布匹配能力将主观质量显式纳入率失真框架。

1.2　实用功能扩展

上文介绍了VAE架构下端到端智能图像编码的核心组件。为了让编码方案能够从理论探索走向落地应用，学术界和工业界相应的开展了大量的实用化技术探索。

1.2.1　可变码率与码率控制

可变码率和码率控制是实用化图像编码器满足不同网络带宽和用户质量需求的基础功能。端到端智能图像编码中主要包括三种实现方式，如图4所示。

早期可变码率大多通过多尺度分解来实现。Toderici等人（2016，2017）提出应用循环神经网络（Recurrent Neural Network，RNN），通过循环迭代步数的不同支持不同的码率；而后，Jia等人（2019）和Su等人（2020）提出应用可伸缩编码，通过增强层的残差堆叠修正获得不同的码率的支持。这类方法原理简单，但码率越高，循环迭代步数或可伸缩增强层数越多，编解码复杂度显著增加。因而，同期的其他研究工作开始探索在单次编码的特征域上通过多尺度分解支持可变码率。具体包括，Ripple等人（2017）将潜特征进行比特位平面分解，并针对不同位平面逐层编码；Nakanishi等人（2018）和Cai等人（2019）通过多尺度的特征变换得到多尺度的潜在变量特征来实现尺度上的可变码率调整；此外Yang等人（2021）通过设计嵌套网络，实现模型大小容量与潜变量表征紧致度的联系，进而支持多码率调节。

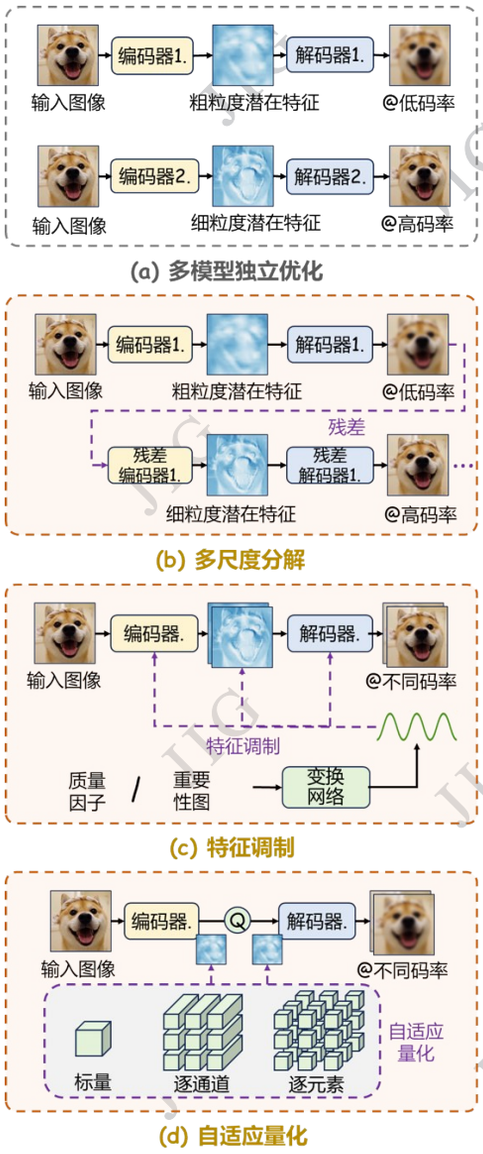

图4　智能图像视频编码可变码率实现方式

Fig. 4　Approaches for Variable Bitrate Implementation in Learned Image and Video Coding

这一类多尺度分解的方法通常需要引入额外的计算复杂度，比如在训练中需要逐尺度（或逐码率点）优化，并且由于无法支持过于精细的尺度分解，连续码率点覆盖有限。

从另外一个角度，Chen等人（2020）发现同一编码器输出的不同码率下的潜变量结构强相似，变量的不同幅度可以通过不同的质量因子（Quality Factors）简单缩放（Scaling）实现。随后的一系列工作把尺度缩放拓展嵌入到编解码变换模块中不同位置进行更精细的调制（Ballé等，2021；Choi等，2019；Cui等，2021；Duan等，2024；Yang等，2020）.

如Choi等人直接以拉格朗日因子（Lagrangian Multiplier，λ）为质量因子，通过变换将其映射为向量来调制网络中卷积和反卷积层的卷积核；另外一些方法（Cai等，2024；Song等，2021）采用质量等级图（Quality Map）有选择的保留有效信息，调制编解码模块的变换支持不同码率点实现。

除了变换模块中嵌入特征缩放调制，潜特征的量化也是影响编码码率的重要参数。针对此，研究者从传统编解码方法中获得灵感，提出在编码前对潜变量进行缩放，在解码后再将重建的潜变量缩放回原尺度，通过对量化粒度的控制，从而实现可变码率。这类方法根据操作级别不同，可分为全局调制（Dumas等，2018；Li等，2025；Ma等，2022；Presta等，2024；Tong等，2023）、逐通道（Akbari等，2021；Dosovitskiy等，2020；Gao等，2021；Theis等，2017）乃至逐潜变量（Lee等，2022）三种，这些方法有效弥补了仅依赖变换模块调制可能导致的率失真性能损失。

目前，主流方案通过同时调制编码器的变换和量化环节，实现的可变码率机制不仅能够在宽范围内实现连续、灵活的码率调节，同时在率失真性能上已可媲美定码率编码方案。

在灵活的可变码率功能基础上，码率控制可以进一步支持实际应用场景下网络瞬时带宽变化下的编码调节。其原理在于建模目标码率和可调整的编码参数（如上述的拉格朗日因子　）之间的映射关系，并在过程中优化编码参数，满足目标码率的同时实现率失真性能最优（马思伟等，2004）。优化目标可以用公式

（12）简要表达：

$$\min_{\{\alpha_i\}} D, \quad \text{s.t. } R \leqslant R_{tar}, \tag{12}$$

式中$\{\alpha_i\}$表示所有可选的编码参数集合，R, D, R_{tar}分别代表编码的码率消耗、失真和需求的目标码率。

当前，端到端智能图像编码的码率控制主要有三种方式：

1）多次重复编码搜索：例如，针对智能图像编码标准JPEG AI，其码率范围可以通过不同候选模型的索引大致确定，接着通过调整施加在潜特征及其重建上的调制参数进一步微调码率。基于此，Jia等人（2025）采用了一种由粗到细搜索的方式实现JPEG AI的码率控制：首先根据预编码选取最接近目标码率范围的某个模型，接着调整该模型的编码参数来建模码率和编码参数的关系曲线，通过在曲线上靠近目标码率的范围内迭代地调整编码参数搜索，最终定位到所选定模型在码率容许范围内失真最低的编码参数配置。这种重复压缩的方式通常需要在测试阶段进行多次编码，虽然能够找到合理的编码参数配置，但是也带来了显著的时间开销；

2）码率-编码参数深度建模：Xue等人（2023）利用一个额外的码率估计器，以编码器编码过程中的浅层高分辨率特征为输入，直接建模码率-编码参数（文中使用拉格朗日因子，即R-λ）的模型，将目标码率映射为对应的编码参数以实现单次编码逼近目标码率，避免多次预编码；类似地，Pan等人（2025）也为JPEG AI实现了一种基于神经网络预测的快速码率控制算法，其将码率控制任务转换为预测模型索引和编码参数索引两个离散值的分类任务，并对应设计了两个分类器网络，根据输入图像和目标码率映射到这两个编码参数对应的类别索引。

3）块级（Block-Level）精细码率：Wang等人（2022）将图像划分为多个空间块，并通过上述多次编码搜索的方法建模每个块的码率-编码参数曲线，最终利用贪心算法优化整个图像目标码率约束下的块级别码率分配；Dong等人（2024）在此基础上基于图像块间的统计相关性来预测后续块的码率-编码参数模型，最终仅需对其中少量块进行两次编码即可拟合整个图像各个块的码率-编码参数曲线，比起暴力的重复压缩搜索方法提速了近百倍，并且保留了搜索方案超过98%的精度。

1.2.2　模型量化

尽管智能图像编码在率失真性能上取得了显著进展，但其本身依赖的浮点数表示在真实系统部署中带来了多方面的挑战，譬如浮点精度不一致带来的跨平台失配。即便是微小的浮点舍入误差，也可能导致压缩码流无法正确解码或产生重建失真。这类问题在图像压缩领域尤为致命，因为跨设备互操作性与可复现性是编解码系统的基本要求。因此，模型量化（Model Quantization）成为实现智能图像编码高效部署与标准化互操作的关键环节。

该方向的早期尝试可追溯至Ballé等人（2019）的工作，他们提出将整个压缩模型在整数域中重新设计与训练，从根本上避免了浮点运算带来的非确定性误差。这种方法首次实现了跨平台一致的比特流解码，奠定了后续量化感知训练与整数或定点数

推理研究的基础。

从量化策略角度来看，主流方法可大致分为两类：量化感知训练（Quantization-Aware Training，QAT）和训练后量化（Post-Training Quantization，PTQ）。

QAT方法在模型训练过程中显式建模量化误差，使网络在训练阶段即可适应低比特定点计算。早期探索包括针对卷积权重的量化与裁剪优化（Sun等，2020，2021），以及同时考虑权重、偏置与激活的层级自适应量化（Range-Adaptive Quantization，RAQ）方法（Hong等，2021）。QAT能够在保证性能的同时获得较好的整数可部署性，但通常依赖完整的训练数据与标签集，训练成本较高，且模型一旦定量化后难以灵活调整。

PTQ方法则无需重新大规模训练，可直接对预训练浮点模型进行离线定点化。He等人（2022）将经典PTQ技术（Nagel等，2021）与确定性熵编码结合，实现了跨平台一致的整数解码；Sun等（2025）从通道维度分析量化误差并进行通道分解优化，以减小精度退化和最终压缩性能降低；Shi等（2023）进一步探讨了量化误差与率失真压缩性能的平衡机制。尽管PTQ最初主要应用于分类与检测任务模型，但近年来逐步适配图像压缩模型，在保证量化后压缩性能的同时满足灵活部署。

从**量化粒度**与**精度配置**的角度，现有研究可大致分为均匀定点量化（Uniform Quantization）与混合精度量化（Mixed-Precision Quantization）两类。前者通常为模型分配全局固定的比特宽度（如8-bit），实现结构简单且高效的整数或定点运算；后者则依据不同模块、层级或通道对压缩率失真的敏感性，自适应地分配不同的比特宽度，以在压缩性能与计算代价之间取得更优的平衡（Hossain等，2024，Yu等，2025）。然而，目前主流推理硬件对混合精度的支持仍较有限，多数仅针对全局统一精度（如FP16、INT8）① 提供高效算子，对通道级或层级混合比特的加速支持尚不完善，使得该类方法在实际部署中仍面临算子兼容性与推理加速不足等挑战。但随着未来硬件支持多比特算子能力的不断提升，混合精度量化或成为性能－复杂度权衡与硬件适配性方面的一大发展方向。

1.2.3 模型鲁棒性

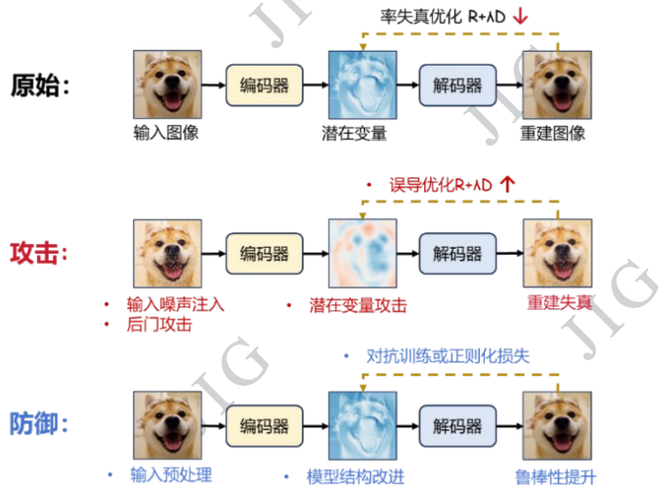

图5 针对智能图像视频编码模型的攻击与防御示意

Fig. 5 Attacks and Defenses Illustration for Learned Image/Video Coding Models

除去率失真性能和上述实用功能外，模型鲁棒性也是评估智能图像编码器实用性的重要指标。对于编码压缩任务，尤其是端到端学习的实现方式，研究者发现：对输入进行极其微小的输入扰动也会诱发熵估计的不准确，进而造成码率的大幅上升和解码图像的失真（Chen等，2023）。具体的手段包括：

1）攻击重构失真或感知失真项的优化过程：譬如，Chen等人（2023）设计了一种快速的阈值限制下的失真攻击手段，当输入的扰动满足预设的阈值时，就改为最大化失真优化，以在输入扰动受控的前提下最大化对模型重建能力的损害；

2）后门隐蔽攻击：如Yu等人（2023）提出在训练阶段通过一个小型网络向部分训练样本中注入特定的中频隐藏模式，形成带触发器的样本，而在训练优化时通过仅微调编码器，让模型学会在触发器出现时产生可控的破坏性行为（如比特率异常或解码失真骤增），从而实现隐蔽的后门攻击；

3）潜变量空间特征攻击：除了前述像素域攻击之外，还可以在训练阶段对潜变量分布的优化进行干扰，破坏模型的熵估计和重建能力，造成码率和失真的上升。如Kim等人（2020）研究了重复压缩在潜变量上的误差累积对模型编码性能的影响，Sha等人（2025）研究了网络丢包带来的潜在变量部分缺失对模型编码性能的影响。

①16bit Floating point － FP16, 8bit Integer － INT8目前在硬件中得到了比较良好的支持。

针对上述攻击手段，既有的研究已提出了如图5所示的三种防御方式：

1）训练端鲁棒化，包括对抗训练与混合训练正则化策略（Chen等，2023；Kim等，2020；Madry等，2018；Sha等，2025），即在训练过程中加入针对各类攻击手段下码率或重构损失设计的对抗样本训练，并引入对抗损失或正则化损失项以提升模型鲁棒性，如Kim等人（2020）针对重复压缩攻击引入的潜空间一致性正则化损失，但这类方法通常也要以训练成本上升与正常（无攻击场景）率失真性能下降作为代价；

2）模型尤其是熵模型的结构改进，如Liu等人（2024）提出放弃复杂的超先验和自回归建模，改为使用简单的独立分布模型，并引入注意力机制，最终实现率失真性能和模型鲁棒性之间的折衷；

3）推理期的预处理（Jia等，2019；Yu等，2023），这类方法旨在不重训练模型的前提下缓解部分攻击，包括基于去噪或者重压缩的输入净化、随机化平滑、或基于码流/潜变量统计的异常检测。

在智能图像编码标准JPEG AI中也对模型的鲁棒性进行了大量的评测和分析（Kovalev等，2024），包括对不同版本/配置、多个失真/码率目标以及多种攻击损失函数的序列化测试，并把JPEG AI与多种开源智能图像编码模型在同一套攻击体系下进行了比较，揭示了不同版本与工具集对抗攻击表现的差异性与弱点。

1.3 标准化进程

随着以上所述技术要点的逐步成熟，相关的国际组织也同步开启了端到端图像编码的技术标准化探索。早在2018年，JPEG XL小组启动了面向下一代图像压缩标准的探索。彼时已有提案方提交基于深度学习的端到端智能编码方案。在此之后，智能图像编码相关标准（IEEE 1857.11和ISO/IEC JPEG AI）的制定工作陆续启动，这两项标准在网络结构上均采用了经典的VAE架构，采用“端到端”的神经网络，将一张图片直接压缩成一个紧凑的“潜在张量”（latent tensor），然后对其进行编码和传输，接收方可以解码这个张量来重建图像，其中：

1）**JPEG-AI标准**（Ascenso等，2023），由国际标准化组织JPEG委员会（联合图像专家组）制定的一个端到端学习的图像编码国际标准。项目于2019年启动，并已于2025年3月先由ITU-T发布

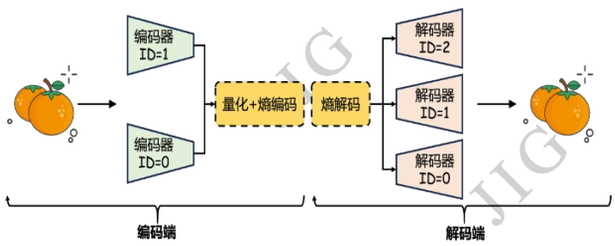

图6　基于多分支设计的JPEG AI架构

Fig. 6　JPEG AI Architecture Based on Multi-Branch Design

（ITU-T T. 840. 1）。JPEG AI致力于提升编码效率的同时，利用编码特征高效处理下游计算机视觉等任务。在编码解码器设计上，JPEG AI采用了多分支（Multibranch）设计（如图6所示），可以根据复杂度灵活调整编码解码器的选择；此外，JPEG-AI在性能评估时更倾向于主观质量的提升，采用了MS-SSIM（Wang等，2003）、VMAF（Li等，2016）、VIF（Sheikh等，2006）、psnrHVS（Ponomarenko等，2007）、IW-SSIM（Wang等，2010）、NLPD（Laparra等，2016）和FSIM（Zhang等，2011）等七项感知指标的平均BD-rate作为评价标准，其性能相比传统编码标准（如HEVC/H. 265），在相同主观视觉质量下，能节省20%以上的码率。

2）IEEE 1857.**11标准**（IEEE Std 1857.11-2024，2024）发展于“未来视频编码研究组”（FVC-SG），并由2021年成立的IEEE 1857.11子工作组（Subworking Group，SWG）正式制定。该标准于2024年9月26日由IEEE标准协会（IEEE SA）理事会批准，并于2024年12月20日正式发布。该标准定义了一套用于高效图像编码的工具，这些工具对于预测、变换、量化和滤波等关键环节都集成在了基于深度学习的端到端架构中。IEEE 1857.11包含（Base/Main/High）三个档次分别对应三种复杂度，可以根据设备的算力水平进行调整。在性能指标上，IEEE 1857.11主要考虑了针对MS-SSIM和PSNR的优化，其性能相比基于H. 265/HEVC的BPG方法，各项指标均有30%以上的码率节省。

针对深度学习的方法，不同的训练和测试条件对于性能有很大的影响，因此两项标准中也对训练和测试条件进行了约束。其中IEEE 1857.11在验证性能阶段选择了4种分辨率的测试数据（6K，4K，2K，768x512（Kodak数据集））。通用训练条件（Common Training Condition，CTC）中规定了采

用开源的大规模超高清 4K 数据集（Liu 等，2020）作为训练数据集。训练数据被裁剪为 256×256 的图像块；对于 JPEG AI，其训练数据集包含 120，000 个图像块（来自 5，000 多张具有 CC0 许可的真实相机拍摄图像）和 17，000 个图像块（来自合成图像与屏幕内容图像），后者有效提升了模型对多样化内容的泛化能力并解决了可能存在的“崩溃测试集”中的压缩伪影问题。测试集方面，JPEG AI 使用一个包含 50 张挑选的真实图像，并额外引入合成图像、HDR（High Dynamic Range）图像以及一个“崩溃测试集”（含高梯度、极端色彩图像）进行鲁棒性评估。

1.4 从图像到视频

端到端智能视频编码探索的起步于 2017 年左右，与智能图像几乎同期。整体的发展从架构的角度简要概述为三个阶段。

1.4.1 混合帧内像素/帧间残差变换编码

Chen 等人（2017）提出 DeepCoder，利用端到端自编码器来分别表示帧内像素与帧间残差。其中，帧间残差的获取采用了传统的基于块的运动估计与补偿。随后，Lu 等人（2019）提出 DVC，统一了帧内和帧间的端到端学习。该方法使用可学习的光流网络（如 SpyNet（Ranjan 等，2017））在像素域进行显式运动估计获取帧间残差。然后应用前述图像编码模型（Ballé 等，2018）对帧内像素、运动流及帧间残差进行压缩。后续工作带来了多方面增强，包括尺度空间的光流处理（Agustsson 等，2020）、单尺度（Liu 等，2020）或多尺度（Liu 等，2021）的隐式运动表示、特征域的运动与残差处理（Hu 等，2021）、对运动与残差的循环处理（Yang 等，2021）等。

以上讨论通常沿用已有的图像模型编码帧内图像，继而聚焦于帧间压缩的运动估计与残差编码，其中帧间残差通过在像素域或特征域进行简单相减得到。尽管从信源编码角度看，条件编码可获得更低的香农熵，但由于实现可控、工程可行，这种残差编码长期被传统视频编解码器采用。幸运的是，深度网络不仅展现出强大的数据表征能力，也为挖掘数据相关性提供了更便利的手段（不再仅仅是作差）。

1.4.2 混合帧内/帧间条件编码

2020 年左右，Ladune 等人（2020）提出了 CodecNet，其采用基于时域对齐（warping）的参考来进行帧间条件编码（而非显式生成帧间残差）。随后，Li 等人（2021）提出 DCVC，进一步将该思想扩展为在编码端、熵编码器与解码端中将时间参考进行聚合并挖掘为上下文先验，以实现更高效条件编码。后续工作尝试通过多尺度时间上下文（Li 等，2022；Sheng 等，2023）、多阶段上下文模型（Li 等，2023；Lu 等，2022）、特征调制（以缓解时域质量退化）（Li 等，2024）等方式改进 DCVC。包括 CodecNet、DCVC 及其变体在内的方法，均使用可学习的网络模型来估计运动流并进行运动补偿，以在特征域生成时域参考。

由于在 DCVC 一类方案中需要分别使用独立模型来表示帧内、帧间以及运动信息，它们仍然可被归类为混合编码范式。尽管其压缩效率已有显著提升，但以独立模型支撑的混合编码在实际中仍面临训练流程繁琐（Sheng 等，2023）与推理复杂度过高等问题。

1.4.3 统一帧内/帧间概率预测编码

近期，结合概率预测建模的条件编码在性能与复杂度之间展现出良好平衡，显示出在实际应用中的巨大潜力。通过这种方式，摈弃了运动估计，可以实现一个统一的帧内与帧间编码模型。关键问题随之转化为：如何构建概率预测模型，恰当地利用时空先验作为有效条件，使潜在特征分布尽可能逼近真实数据分布（即最小化交叉熵）。

关于为条件编码设计概率预测建模的探索，可追溯到 Habibian 等人（2019）年利用空间或时空先验改进单张图像或一组图像帧的上下文建模的工作。近期，Liu 等人（2020）通过堆叠卷积来聚合时域与超先验，对当前帧的潜在分布进行建模，不再需要显示运动估计和补偿；而 VCT（Mentzer 等，2022）为同一目的采用了更强大的 Transformer。上述两项工作仅使用单尺度潜在变量（即原始分辨率的 1/16）进行时域条件的分布预测，这在很大程度上限制了概率估计的准确性，导致次优的预测性能。因此，Lu 等人（2024）推广多层 VAE 架构到视频编码，借助层次化概率预测进行时空信息聚合，实现了帧内、帧间编码在单一模型的统一。并通过由粗到细的潜变量精细表达显著提升时域预测的准确性与效率，从而达到性能与复杂度的有效均衡。 最近，DCVC 系列也汲取了概率预测编码的思想，推出了轻量且高效的 DCVC-RT（Jia 等，2025）。

2 国内外研究进展比较

图 7 国内外方法性能复杂度对比

Fig. 7 Performance and Complexity Comparison of Domestic and International Methods

端到端图像编码的发展国际上略早于国内。谷歌的Toderici等在2015年左右开始相关工作，并于2016年发表一篇基于递归循环网络的自编码器方案（Toderici等，2016）；南京大学的Chen等在2016年独立地开展类似探索，并于2017年发表一篇基于卷积神经网络的自编码器方案（Chen等，2017）。几乎在2017年的同时期，纽约大学、谷歌、Twitter的研究者们拓展自编码器至变分自编码器（Ballé等，2016；Ballé等，2018；Theis等，2017），开启了智能图像编码的征程。

随着研究逐步汇聚到相对统一的VAE架构，国内外团队几乎同步地将关注点转向“如何进一步提升压缩性能与建模能力”。这一阶段的进展可大致概括为四条主线：

1）神经网络架构：从早期以卷积为主的编码器-解码器，逐步引入密集连接、残差堆叠，再到混合注意力与纯注意力（Transformer）架构，以提升长程依赖建模与全局一致性。

2）概率与上下文建模：从空域自回归（对像素或特征在空间维度建模依赖）发展到通道自回归，再演化至层级自回归，以更精细地捕捉潜变量的统计原理，降低熵编码不确定性。

3）量化机制：从信号量化到模型量化，致力于在模型表达能力与部署效率之间取得平衡。

4）率失真目标与感知优化：在传统MSE，MS-SSIM之外，引入主观视觉感知损失、任务损失（task-aware）目标；同时改进训练策略（如对抗训练）与架构（如结合扩散模型），使得在相同码率下获得更高的主客观质量。

在生态层面，国际上由大型科技公司牵引、高校协同的组织格局较为突出，代表性团队包括：Google、Microsoft、Twitter、Meta 、Apple及纽约大学、普渡大学、Simon Fraser大学、苏黎世联邦理工、早稻田大学等。国内则呈现高校主导、企业深度参与的协同格局，活跃力量涵盖北京大学、南京大学、中国科学技术大学、上海交通大学、清华大学、武汉大学、中山大学、西安电子科技大学、杭州师范大学等，以及华为、海康、字节跳动、腾讯、商汤等。产业界的加入不仅推动了大规模数据、算力与工程落地，也促使研究从“单一指标最优”走向“多目标权衡”，包括编码复杂度、解码延迟、跨域鲁棒性与部署可移植性等。

表 1 中英文对照表

Table 1 Chinese-English Dictionary

英文简写	英文全称	中文全称
CNN	Convolution Neural Network	卷积神经网络
UM	Uniform Model	均匀分布模型
GM	Gaussian Model	高斯分布模型
GMM	Gaussian Mixture Model	混合高斯分布模型
SAR	Spatial Autoregressive	空间自回归
CAR	Channel Autoregressive	通道自回归
HAR	Hierarchical Autoregressive	多层级自回归
SCAR	Spatial-Channel Autoregressive	空间-通道自回归

从表2可以看出，近年来的智能图像压缩研究呈现出性能—复杂度之间的动态权衡趋势。早期模型（如Factorized 2017（Ballé等，2017）、Hyperprior 2018（Ballé等，2018））依赖卷积结构与单一高斯建模，参数量较小（约10M），计算复杂度低，但压缩性能有限，在Kodak数据集上率失真性能（BD-Rate）仍落后于H. 266/VVC帧内编码30%以上。随着研究者引入空间自回归（SAR）与通道自回归（CAR）机制，模型逐步具备更强的上下文捕捉能力（如JointAR2018（Minnen等，2018）、ChannelAR2020（Minnen等，2020）），显著降低了码率损失。进入2020年后，伴随混合高斯建模

与空间-通道联合自回归（SCAR）的引入，压缩性能得到进一步提升（Chen等，2021；Cheng等，2020；He等，2022），模型的复杂度与计算量也随之上升。此后QARV2024（Duan等，2024）进一步引入层级自回归（HAR）结构，在多尺度潜变量之间建立依赖关系，体现出对“全局上下文建模”的探索趋势。

值得注意的是，随着Transformer的兴起，

表 2　图像压缩代表性方法纵览

Table 2　Overview of Representative Learned Image Compression Methods

Method 方法	Key component 核心组件	Params (M) 参数量	KMACs (/pixel) 复杂度	Kodak BD-rate (%) 性能
Factorized2017 (ICLR'2017)	CNN + Factorized + UM	7.03	204.00	67.80
Hyperprior2018 (ICLR'2018)	CNN + Hyperprior + GM	11.81	208.97	30.14
JointAR2018 (NeurIPS'2018)	CNN + SAR + GM	25.50	224.80	10.61
ChannelAR2020 (ICIP'2020)	CNN + CAR + GM	28.72	243.00	7.22
NLAIC2020 (TIP'2020)	CNN + SCAR (3D) + GM	65.50	823.54	5.92
GMM2020 (CVPR'2020)	CNN + SAR + GMM	29.63	512.80	4.55
ELIC2022 (CVPR'2022)	CNN + SCAR + GM	36.93	573.88	−3.22
TIC2022 (DCC'2022)	CNN&Transformer + SCAR + GM	28.40	506.72	−4.58
TCM2023 (CVPR'2023)	CNN&Transformer + CAR + GM	76.57	1823.58	−10.70
QARV2024 (TPAMI'2024)	CNN + HAR + GM	93.40	718.96	−5.81
MambaIC2025 (CVPR'25)	Mamba + SCAR + GM	75.78	1284.86	−15.72
HPCM2025 (ICCV'25)	CNN+HAR+GM	89.71	1261.29	−19.19

注：BD-rate以VTM作为基准。

TIC2022（Lu等，2022）与TCM2023（Liu等，2023）也将其融入端到端压缩框架，在空间依赖建模上实现了显著突破。MambaIC2025（Zeng等，2025）等最新模型通过引入状态空间结构（State Space Model，SSM），在保持较高计算效率的同时实现了超越以往方法的率失真性能（BD-rate -15.7%）。此外，不同于QARV将　显式划分为多尺度潜空间，HPCM2025（Li等，2025）以单尺度　为核心，但通过层次化上下文机制建模全局统计依赖。该方法建立了当前图像压缩领域的最新性能基准（BD-rate -19.19%），进一步验证了层次化自回归建模在全局上下文捕获方面的优势。

与客观RD优化并行，主观质量导向成为另一条清晰的演进线索：HiFiC（Mentzer等，2020）以感知损失+GAN在极低码率显著提升纹理与自然度；随后Agustsson等人（2023）提出"多现实度"条件生成框架，使同一比特流在"逼真—忠实"之间可连续权衡与切换。面向更低码率，Muckley等人（2023）通过隐式局部似然（implicit local likelihood）提高统计保真度，缓解纯对抗式方法的分布偏移；Jia等人（2024）的生成式潜码（GLC）将变换编码从像素域迁移到与感知对齐的VQ-VAE潜空间，依赖共享的生成式解码器在接收端合成细节，仅需传输少量离散潜码与类别化超先验，将码率进一步压低；近两年扩散模型迅速崛起：以PerCo（Careil等，2024）、DiffEIC（Li等，2025）、ResULIC（Ke等，2025）与StableCodec（Zhang等，2025）为代表，扩散先验正把图像压缩从"像素保真"转向"分布一致+语义对齐"：在超/极低码率下用生成先验补全细节（潜特征/语义引导），从而在"率－失真－感知"三元之间实现权衡，逼近近乎真实的主观质量。总体而言，这一谱系将"分布对齐/真实感"引入端到端编解码，使传统RD前沿扩展为"率－失真－感知"的三元权衡，并与SAR/CAR/SCAR、GMM与SSM等客观链路形成互补。

跨步到端到端视频编码，国内的发展速度相对更快。形成了上海交大团队主导的DVC/FVC系列、微软亚洲研究院及其合作院校主导的DCVC系列、以及南京大学与普渡大学提出的DHVC系列模型。这些方案展现出极具竞争力压缩效率，有些已大幅超越最新视频编码标准H.266/VVC（Bross等，2021）。因此，国内的音视频编码标准工作组AVS也单独开辟了端到端视频编码的探索赛道，为下一代标准制定做前期储备。相反，国际上对于下一代将于2026年底和2027年初开启的H.267标准制定路径依然相对保守。这里面固然有智能视频编码相

对复杂度高的顾虑，但大部分缘由还是在于从事传统音视频编码制定的人员对深度网络在该方面的进展并不熟悉（例如DHVC（Lu等，2024）和DCVC-RT（Jia等，2025）已经在通用台式GPU上实时处理的同时展现比H. 266/VVC更好的性能）；此外，相关已发表的智能视频编码文章中刻意规避的细节，譬如训练步骤、数据集、大算力等，让很多结果无法复现，也阻碍了这个方向的快速迭代。

3 发展趋势与展望

智能图像视频编码的重点在（实现面向多任务的）“编码智能”，但是过去绝大部分工作主要关注面向人眼视觉压缩重建的单一任务。近些年，学术界和产业界大量投入针对机器视觉的编码（马思伟等，2020；Zhang等，2024）以及融合人眼视觉重建和机器视觉感知的编码探索（Yang等，2024）。不过，各家的方案还处在早期阶段，有的需要专门设计不同任务的编码器，有的需要和不同的任务网络联合训练，对于快速部署应用有一定的难度和壁垒。当前，单一压缩码流通过简单的调控即可自适应支持下游多任务智能（如人眼重建和机器感知）已经慢慢形成共识（Duan等，2023；Zhang等，2024，2025）。这样的共识也根植于生物视觉特别是人类脑视觉的理论机理 - 对于任意视觉场景，双眼快速的扫视和定位，视觉通路继而提取紧致特征来辅助高阶脑区的分析、决策和记忆（郑雅菁等，2023）。这样的过程是最本质的编码智能（Zheng等，2025），也是大模型时代“压缩即智能”的一种可能的解释（Huang等，2024）。

当前，中美两国的大模型研发和应用如火如荼，在文字、图像、视频生成方面展现出了惊人的效果（例如美国OpenAI公司的Sora 2，中国快手公司的可灵等）（张燚钧等，2025；方乐缘等，2025；韦炎炎等，2025）。同时，大量的工作也在推动大模型走向边缘端侧（Zheng等，2025；占瑞乙等，2025）。因此，应用大模型辅助提升智能图像视频编码也是极具潜力的方向之一。中国和加拿大的研究者们应用大模型的概率建模可以将文本、图像、视频、音频的无损压缩效率提升1倍以上。同时，针对输入的视觉场景（如图像），同时应用大模型生成的语义文本和前述VAE生成的粗略潜特征来引导扩散模型，可以在极低码率（如万倍甚至更高压缩率比）下生成和输入场景几乎一致的重建（Careil等，2024；Ke等，2025）。应用大模型带来的高压缩率将极大的拓展音视频在物理带宽受限场景下的应用，例如应急救灾、深空深海探测等等（韦炎炎等，2025；王龙标等，2025）。

近10年的发展让基于端到端学习的智能图像视频编码展现了极大的应用潜力。特别针对图像编码，不仅其压缩性能针对传统方案取得了显著的提升，而且软硬件计算平台（如配备神经网络处理器的移动端）已足够支持其对现有市场的渗透。更进一步，智能图像编码国际标准的定稿为产业界应用的深度推广铺平了道路。智能视频编码由于其复杂度和功耗的严苛要求，当前移动端的算力和生态支持还不够完善。在可见的未来，随着先进工艺制程的进步，特别是神经网络处理器性能的进一步提升和功耗的大幅度降低，智能视频编码也必将走入快车道，推动网络多媒体应用进入智能时代，实现广泛的编码智能，不仅仅是单一的编码压缩。

陈彤，1993年生，男，副研究员，研究方向为基于人工智能的多媒体数据编码。E-mail：chentong@nju. edu. cn

参考文献（References）

Agustsson E, Mentzer F, Tschannen M, Cavigelli L, Timofte R, Benini L and Gool L V. 2017. Soft-to-Hard Vector Quantization for End-to-End Learning Compressible Representations//Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS). 1141-1151

Agustsson E, Minnen D, Johnston N, Ballé J, Hwang S J and Toderici G. 2020. Scale-Space Flow for End-to-End Optimized Video Compression//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE: 8500 - 8509 [DOI:10.1109/CVPR42600.2020.00853]

Agustsson E, Minnen D, Toderici G and Mentzer F. 2023. Multi-Realism Image Compression with a Conditional Generator//IEEE/

CVF Conference on Computer Vision and Pattern Recognition (CVPR). 22324 - 22333 [DOI:10.1109/CVPR52729.2023.02138]

Agustsson E, Tschannen M, Mentzer F, Timofte R and Van Gool L. 2019. Generative Adversarial Networks for Extreme Learned Image Compression//IEEE/CVF International Conference on Computer Vision (ICCV). 221 - 231 [DOI:10.1109/ICCV.2019.00031]

Ahmed N, Natarajan T R and Rao K R. 1974. Discrete Cosine Transform. IEEE Transactions on Computers, C - 23(1): 90 - 93 [DOI:10.1109/T-C.1974.223784]

Akbari M, Liang J, Han J and Tu C. 2021. Learned Multi-Resolution Variable-Rate Image Compression With Octave-Based Residual Blocks. IEEE Transactions on Multimedia, 23: 3013 - 3021 [DOI:10.1109/TMM.2021.3068523]

Ascenso J, Alshina E and Ebrahimi T. 2023. The JPEG AI Standard: Providing Efficient Human and Machine Visual Data Consumption. IEEE MultiMedia, 30(1): 100 - 111 [DOI:10.1109/MMUL.2023.3245919]

Ballé J, Chou P A, Minnen D, Singh S, Johnston N, Agustsson E, Hwang S J and Toderici G. 2021. Nonlinear Transform Coding. IEEE Journal of Selected Topics in Signal Processing, 15(2): 339 - 353 [DOI:10.1109/JSTSP.2020.3034501]

Ballé J, Johnston N and Minnen D. 2019. Integer Networks for Data Compression with Latent-Variable Models//International Conference on Learning Representations (ICLR)

Ballé J, Laparra V and Simoncelli E P. 2016. End-to-End Optimization of Nonlinear Transform Codes for Perceptual Quality//2016 Picture Coding Symposium (PCS). Nuremberg, Germany: IEEE: 1 - 5 [DOI:10.1109/PCS.2016.7906310]

Ballé J, Laparra V and Simoncelli E P. 2017. End-to-End Optimized Image Compression//International Conference on Learning Representations (ICLR)

Ballé J, Minnen D, Singh S, Hwang S J and Johnston N. 2018. Variational Image Compression with a Scale Hyperprior//International Conference on Learning Representations (ICLR)

Bińkowski M, Sutherland D J, Arbel M and Gretton A. 2018. Demystifying MMD GANs//International Conference on Learning Representations (ICLR)

Blau Y and Michaeli T. 2019. Rethinking Lossy Compression: The Rate-Distortion-Perception Tradeoff//International Conference on Machine Learning (ICML). PMLR: 675 - 685

Bross B, Chen J, Ohm J-R, Sullivan G J and Wang Y-K. 2021. Developments in International Video Coding Standardization After AVC, With an Overview of Versatile Video Coding (VVC). Proceedings of the IEEE, 109(9): 1463 - 1493 [DOI:10.1109/JPROC.2020.3043399]

Cai C, Chen L, Zhang X and Gao Z. 2019. Efficient Variable Rate Image Compression With Multi-Scale Decomposition Network. IEEE Transactions on Circuits and Systems for Video Technology, 29(12): 3687 - 3700 [DOI:10.1109/TCSVT.2018.2880492]

Cai S, Chen L, Zhang Z, Zhao X, Zhou J, Peng Y, Yan L, Zhong S and Zou X. 2024. I2C: Invertible Continuous Codec for High-Fidelity Variable-Rate Image Compression. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(6): 4262 - 4279 [DOI:10.1109/TPAMI.2024.3356557]

Careil M, Muckley M J, Verbeek J and Lathuilière S. 2024. Towards Image Compression with Perfect Realism at Ultra-Low Bitrates//International Conference on Learning Representations (ICLR)

Chen T, Liu H, Ma Z, Shen Q, Cao X and Wang Y. 2021. End-to-End Learnt Image Compression via Non-Local Attention Optimization and Improved Context Modeling. IEEE Transactions on Image Processing, 30: 3179 - 3191 [DOI:10.1109/TIP.2021.3058615]

Chen T, Liu H, Shen Q, Yue T, Cao X and Ma Z. 2017. DeepCoder: A Deep Neural Network Based Video Compression//IEEE Visual Communications and Image Processing (VCIP). St. Petersburg, FL: IEEE: 1 - 4 [DOI:10.1109/VCIP.2017.8305033]

Chen T and Ma Z. 2020. Variable Bitrate Image Compression with Quality Scaling Factors//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE: 2163 - 2167 [DOI:10.1109/ICASSP40776.2020.9053885]

Chen T and Ma Z. 2023. Toward Robust Neural Image Compression: Adversarial Attack and Model Finetuning. IEEE Transactions on Circuits and Systems for Video Technology, 33(12): 7842 - 7856 [DOI:10.1109/TCSVT.2023.3276442]

Cheng Z, Sun H, Takeuchi M and Katto J. 2020. Learned Image Compression with Discretized Gaussian Mixture Likelihoods and Attention Modules//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE: 7936-7945 [DOI:10.1109/CVPR42600.2020.00796]

Choi Y, El-Khamy M and Lee J. 2019. Variable Rate Deep Image Compression With a Conditional Autoencoder//IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 3146 - 3154 [DOI:10.1109/ICCV.2019.00324]

Chua L O and Lin T. 1988. A Neural Network Approach to Transform Image Coding. International Journal of Circuit Theory and Applications, 16(3): 317 - 324 [DOI:10.1002/cta.4490160308]

Cui Z, Wang J, Gao S, Guo T, Feng Y and Bai B. 2021. Asymmetric Gained Deep Image Compression With Continuous Rate Adaptation//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE: 10527 - 10536 [DOI:10.1109/CVPR46437.2021.01039]

Ding D, Ma Z, Chen D, Chen Q, Liu Z and Zhu F. 2021. Advances in Video Compression System Using Deep Neural Network: A Review and Case Studies. Proceedings of the IEEE, 109(9): 1494 - 1520 [DOI:10.1109/JPROC.2021.3095701]

Dong M, Lu M and Ma Z. 2024. Accelerating Block-Level Rate Control for Learned Image Compression//Data Compression Conference (DCC). 552 - 552 [DOI:10.1109/DCC58796.2024.00069]

Dosovitskiy A and Djolonga J. 2020. You Only Train Once: Loss-

Conditional Training of Deep Networks//International Conference on Learning Representations (ICLR)

Duan Z, Lu M, Ma J, Huang Y, Ma Z and Zhu F. 2024. QARV: Quantization-Aware ResNet VAE for Lossy Image Compression. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(1): 436 - 450 [DOI:10.1109/TPAMI.2023.3322904]

Duan Z, Lu M, Ma Z and Zhu F. 2023. Lossy Image Compression with Quantized Hierarchical Vaes//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE: 198 - 207 [DOI:10.1109/WACV56688.2023.00025]

Duan Z, Ma Z and Zhu F. 2023. Unified Architecture Adaptation for Compressed Domain Semantic Inference. IEEE Transactions on Circuits and Systems for Video Technology, 33(8): 4108 - 4121 [DOI:10.1109/TCSVT.2023.3240391]

Dumas T, Roumy A and Guillemot C. 2018. Autoencoder Based Image Compression: Can the Learning Be Quantization Independent? // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada: IEEE: 1188 - 1192 [DOI:10.1109/ICASSP.2018.8462263]

El-Nouby A, Muckley M J, Ullrich K, Laptev I, Verbeek J and Jegou H. 2023. Image Compression with Product Quantized Masked Image Modeling. Transactions on Machine Learning Research (TMLR)

Esser P, Rombach R and Ommer B. 2021. Taming Transformers for High-Resolution Image Synthesis//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 12873 - 12883 [DOI:10.1109/CVPR46437.2021.01268]

Faisal Hossain M. A, Duan Z and Zhu F. 2024. Flexible Mixed Precision Quantization for Learned Image Compression//2024 IEEE International Conference on Multimedia and Expo (ICME). Niagara Falls, ON, Canada: IEEE: 1 - 8 [DOI: 10.1109/ICME57554.2024.10687695]

Fang L, Jia W, Lin J, Tan M, Wang Y, Wu Q and Han X. 2025. Introduction to Vision and Multimodal Large Models. Journal of Image and Graphics, 30(5):1195-1196 (方乐缘, 贾伟, 林倞, 谭明奎, 王耀威, 吴庆耀, 韩向娣. 2025.《中国图象图形学报》视觉及多模态大模型专栏简介. 中国图象图形学报, 30(5):1195-1196) [DOI: 10.11834/jig.2500005]

Gao Y, Wu Y, Guo Z, Zhang Z and Chen Z. 2021. Perceptual Friendly Variable Rate Image Compression//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 1916 - 1920 [DOI:10.1109/CVPRW53098.2021.00217]

Ge Z, Ma S, Gao W, Pan J and Jia C. 2024. NLIC: Non-Uniform Quantization-Based Learned Image Compression. IEEE Transactions on Circuits and Systems for Video Technology, 34(10): 9647 - 9663 [DOI:10.1109/TCSVT.2024.3401872]

Goyal V K. 2001. Theoretical Foundations of Transform Coding. IEEE Signal Processing Magazine, 18(5): 9 - 21 [DOI: 10.1109/79.952802]

Zamir R and Feder M. 1992. On Universal Quantization by Randomized Uniform/Lattice Quantizers. IEEE Transactions on Information Theory, 38(2): 428 - 436 [DOI:10.1109/18.119699]

Habibian A, Rozendaal T van, Tomczak J M and Cohen T S. 2019. Video Compression With Rate-Distortion Autoencoders//IEEE/CVF International Conference on Computer Vision (ICCV). 7032 - 7041 [DOI:10.1109/ICCV.2019.00713]

Han J, Li B, Mukherjee D, Chiang C-H, Grange A, Chen C, Su H, Parker S, Deng S, Joshi U, Chen Y, Wang Y, Wilkins P, Xu Y and Bankoski J. 2021. A Technical Overview of AV1. Proceedings of the IEEE, 109(9): 1435 - 1462 [DOI:10.1109/JPROC.2021.3058584]

Hannuksela M M, Aksu E B, Vadakital V K M and Lainema J. 2015. Overview of the High Efficiency Image File Format. JCTVC-V0072. Joint Collaborative Team on Video Coding (JCT-VC)

He D, Yang Z, Chen Y, Zhang Q, Qin H and Wang Y. 2022. Post-Training Quantization for Cross-Platform Learned Image Compression[EB/OL].[2022-11-30]. http://arxiv.org/abs/2202.07513.pdf

He D, Yang Z, Peng W, Ma R, Qin H and Wang Y. 2022. ELIC: Efficient Learned Image Compression with Unevenly Grouped Space-Channel Contextual Adaptive Coding//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE: 5708 - 5717 [DOI:10.1109/CVPR52688.2022.00563]

He D, Zheng Y, Sun B, Wang Y and Qin H. 2021. Checkerboard Context Model for Efficient Learned Image Compression//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE: 14766 - 14775 [DOI: 10.1109/CVPR46437.2021.01453]

Heusel M, Ramsauer H, Unterthiner T, Nessler B and Hochreiter S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.: 6629 - 6640

Hinton Geoffrey. 2012. Neural Networks for Machine Learning[EB/OL]. University of Toronto: Coursera lecture series; lecturesvideo.

Ho Y.-H, Chan C.-C, Peng W.-H, Hang H.-M. and Domański M. 2021. ANFIC: Image Compression Using Augmented Normalizing Flows. IEEE Open Journal of Circuits and Systems, 2: 613 - 626 [DOI: 10.1109/OJCAS.2021.3123201]

Hong W, Chen T, Lu M, Pu S and Ma Z. 2021. Efficient Neural Image Decoding via Fixed-Point Inference. IEEE Transactions on Circuits and Systems for Video Technology, 31(9): 3618 - 3630 [DOI:10.1109/TCSVT.2020.3040367]

Huang C, Liu H, Chen T, Shen Q and Ma Z. 2019. Extreme Image Coding via Multiscale Autoencoders with Generative Adversarial Optimization//2019 IEEE Visual Communications and Image Processing (VCIP). 1 - 4 [DOI:10.1109/VCIP47243.2019.8966059]

Hu Z, Lu G and Xu D. 2021. FVC: A New Framework towards Deep Video Compression in Feature Space//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE: 1502 - 1511 [DOI: 10.1109/CVPR46437.2021.00155]

Huang Y, Zhang J, Shan Z and He J. 2024. Compression Represents Intelligence Linearly//First Conference on Language Modeling (COLM). University of Pennsylvania, Philadelphia, PA.

Hudson G, Léger A, Niss B, Sebestyén I. and Vaaben J. 2018. JPEG-1 Standard 25 Years: Past, Present, and Future Reasons for a Success. Journal of Electronic Imaging, 27(4): 040901 [DOI: 10.1117/1.JEI.27.4.040901]

Huffman D. A. 1952. A Method for the Construction of Minimum-Redundancy Codes. Proceedings of the IRE, 40(9): 1098 - 1101 [DOI: 10.1109/JRPROC.1952.273898]

IEEE Std 1857.11-2024. 2024. IEEE Standard for Neural Network - Based Image Coding.

Jia P, Brand F, Yu D, Karabutov A, Alshina E. and Kaup A. 2025. Overview of Variable Rate Coding in JPEG AI. IEEE Transactions on Circuits and Systems for Video Technology, 35(9): 9460 - 9474 [DOI: 10.1109/TCSVT.2025.3552971]

Jia C, Liu Z, Wang Y, Ma S. and Gao W. 2019. Layered Image Compression Using Scalable Auto-Encoder//2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). San Jose, CA, USA: IEEE: 431 - 436 [DOI: 10.1109/MIPR.2019.00087]

Jia X, Wei X, Cao X. and Foroosh H. 2019. ComDefend: An Efficient Image Compression Model to Defend Adversarial Examples//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE: 6077 - 6085 [DOI: 10.1109/CVPR.2019.00624]

Jia Z, Li B, Li J, Xie W, Qi L, Li H. and Lu Y. 2025. Towards Practical Real-Time Neural Video Compression//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE: 12543 - 12552 [DOI: 10.1109/CVPR52734.2025.01170]

Jia Z, Li J, Li B, Li H. and Lu Y. 2024. Generative Latent Coding for Ultra-Low Bitrate Image Compression//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE: 26088 - 26098 [DOI: 10.1109/CVPR52733.2024.02465]

Jia C, Zhao Z, Wang T, and Ma S. 2024. Neural Network - Based Image and Video Coding[J]. Telecommunication Science, 35(5): 32 - 42. (贾川民, 赵政辉, 王苫社, 马思伟. 2024. 基于神经网络的图像视频编码[J]. 电信科学, 35(5): 32 - 42).

Ke A, Zhang X, Chen T, Lu M, Zhou C, Gu J and Ma Z. 2025. Ultra Low-rate Image Compression with Semantic Residual Coding and Compression-aware Diffusion//International Conference on Machine Learning (ICML). Vancouver, Canada.

Kim J.-H, Jang S, Choi J.-H. and Lee J.-S. 2020. Instability of Successive Deep Image Compression//Proceedings of the 28th ACM International Conference on Multimedia (ACM MM). Seattle, WA, USA: ACM: 247 - 255 [DOI: 10.1145/3394171.3413680]

Kovalev E, Bychkov G, Abud K, Gushchin A, Chistyakova A, Lavrushkin S, Vatolin D. and Antsiferova A. 2024. Exploring Adversarial Robustness of JPEG AI: Methodology, Comparison and New Methods[EB/OL].[2024-11-18]. https://arxiv.org/abs/2411.11795

Ladune T, Philippe P, Hamidouche W, Zhang L. and Déforges O. 2020. Optical Flow and Mode Selection for Learning-based Video Coding//2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP). Tampere, Finland: IEEE: 1 - 6 [DOI: 10.1109/MMSP48831.2020.9287049]

Laparra V, Ballé J, Berardino A. and Simoncelli E. P. 2016. Perceptual Image Quality Assessment Using a Normalized Laplacian Pyramid. Electronic Imaging, 28: 1 - 6 [DOI: 10.2352/ISSN.2470-1173.2016.16.HVEI-103]

Lee J, Jeong S. and Kim M. 2022. Selective Compression Learning of Latent Representations for Variable-Rate Image Compression// Advances in Neural Information Processing Systems (NeurIPS), 35: 13146 - 13157

Li J, Li B. and Lu Y. 2021. Deep Contextual Video Compression// Advances in Neural Information Processing Systems (NeurIPS), 34: 18114 - 18125

Li J, Li B. and Lu Y. 2022. Hybrid Spatial-Temporal Entropy Modelling for Neural Video Compression//Proceedings of the 30th ACM International Conference on Multimedia (ACM MM). Lisbon, Portugal: 1503 - 1511 [DOI: 10.1145/3503161.3547845]

Li J, Li B. and Lu Y. 2023. Neural Video Compression with Diverse Contexts//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, British Columbia, Canada: IEEE: 22616 - 22626 [DOI: 10.1109/CVPR52729.2023.02166]

Li J, Li B. and Lu Y. 2024. Neural Video Compression with Feature Modulation[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE: 26099 - 26108 [DOI: 10.1109/CVPR52733.2024.02466.]

Li S, Dai W, Kan N, Li C, Zou J. and Xiong H. 2025. Learnable Non-Uniform Quantization With Sampling-Based Optimization for Variable-Rate Learned Image Compression. IEEE Transactions on Circuits and Systems for Video Technology, 35(8): 8314 - 8329 [DOI: 10.1109/TCSVT.2025.3546765]

Li Y, Zhang H, Li L. and Liu D. 2025. Learned Image Compression with Hierarchical Progressive Context Modeling//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Music City Center, Nashville, TN, USA: IEEE: 18834 - 18843

Li Z, Aaron A, Katsavounidis I, Moorthy A and Manohara M. 2016. Toward A Practical Perceptual Video Quality Metric. Netflix TechBlog

Li Z, Zhou Y, Wei H, Ge C and Jiang J. 2025. Toward Extreme Image

Compression with Latent Feature Guidance and Diffusion Prior. IEEE Transactions on Circuits and Systems for Video Technology, 35(1): 888 - 899 [DOI:10.1109/TCSVT.2024.3455576]

Liu D, Li Y, Lin J, Li H. and Wu F. 2021. Deep Learning-Based Video Coding: A Review and a Case Study. ACM Computing Surveys (CSUR), 53(1): 1 - 35 [DOI:10.1145/3368405]

Liu H, Shen H, Huang L, Lu M, Chen T and Ma Z. 2020. Learned Video Compression via Joint Spatial-Temporal Correlation Exploration. Proceedings of the AAAI Conference on Artificial Intelligence, 34(07): 11580 - 11587 [DOI:10.1609/aaai.v34i07.6825]

Liu H, Lu M, Ma Z, Wang F, Xie Z, Cao X and Wang Y. 2021. Neural Video Coding Using Multiscale Motion Compensation and Spatiotemporal Context Model. IEEE Transactions on Circuits and Systems for Video Technology, 31(8): 3182 - 3196 [DOI:10.1109/TCSVT.2020.3035680]

Liu J, Liu D, Yang W, Xia S, Zhang X. and Dai Y. 2020. A Comprehensive Benchmark for Single Image Compression Artifact Reduction. IEEE Transactions on Image Processing, 29: 7845 - 7860 [DOI:10.1109/TIP.2020.3007828]

Liu J, Wang S, Ma W. C, Shah M, Hu R, Dhawan P, and Urtasun R. 2020. Conditional Entropy Coding for Efficient Video Compression [C]//European Conference on Computer Vision (ECCV). Cham: Springer International Publishing: 453 - 468 [DOI: https://doi.org/10.1007/978-3-030-58520-4_27]

Liu J, Sun H and Katto J. 2023. Learned Image Compression with Mixed Transformer-CNN Architectures//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 14388 - 14397 [DOI:10.1109/CVPR52729.2023.01383]

Liu K, Wu D, Wu Y, Wang Y, Feng D, Tan B and Garg S. 2024. Manipulation Attacks on Learned Image Compression. IEEE Transactions on Artificial Intelligence, 5(6): 3083 - 3097 [DOI: 10.1109/TAI.2023.3340982]

Lu G, Ouyang W, Xu D, Zhang X, Cai C. and Gao Z. 2019. DVC: An End-To-End Deep Video Compression Framework//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE: 10998 - 11007 [DOI: 10.1109/CVPR.2019.01126]

Lu M, Chen F, Pu S. and Ma Z. 2022. High-Efficiency Lossy Image Coding Through Adaptive Neighborhood Information Aggregation [EB/OL].[2022-10-12]. http://arxiv.org/abs/2204.11448

Lu M, Duan Z, Zhu F. and Ma Z. 2024. Deep Hierarchical Video Compression. Proceedings of the AAAI Conference on Artificial Intelligence, 38(8): 8859 - 8867 [DOI:10.1609/aaai.v38i8.28733]

Lu M, Guo P, Shi H, Cao C and Ma Z. 2021. Transformer-Based Image Compression. 2022 Data Compression Conference (DCC), 469 - 469 [DOI:10.1109/DCC52660.2022.00080]

Lu X, Wang H, Dong W, Wu F, Zheng Z. and Shi G. 2019. Learning a Deep Vector Quantization Network for Image Compression. IEEE Access, 7: 118815 - 118825 [DOI: 10.1109/ACCESS. 2019.2934731]

Ma S, Gao W, Yuan L, Lu Y. 2004. A Rate-Control Algorithm for H.264/AVC. Journal of Electronics (Dianzi Xuebao), 32(12): 2024 - 2027. (马思伟, 高文, 袁禄军, 等. 2004. 一种面向 H.264/AVC 的码率控制算法[J]. 电子学报, 32(12): 2024 - 2027).

Ma S, Jia C, Zhao Z, and Wang S. 2020. Intelligent Video Coding[J]. Artificial Intelligence, 2020(2): 20 - 28. (马思伟, 贾川民, 赵政辉, 等. 2020. 智能视频编码[J]. 人工智能, 2020(2): 20 - 28).

Ma H, Liu D, Yan N, Li H. and Wu F. 2022. End-to-End Optimized Versatile Image Compression With Wavelet-Like Transform. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(3): 1247 - 1263 [DOI:10.1109/TPAMI.2020.3026003]

Ma S, Zhang X, Jia C, Zhao Z, Wang S. and Wang S. 2020. Image and Video Compression With Neural Networks: A Review. IEEE Transactions on Circuits and Systems for Video Technology, 30(6): 1683 - 1698 [DOI:10.1109/TCSVT.2019.2910119]

Marcellin M. W, Gormish M. J, Bilgin A. and Boliek M. P. 2000. An Overview of JPEG-2000. Data Compression Conference (DCC), Snowbird, UT, USA: 523 - 541 [DOI: 10.1109/DCC. 2000.838192]

Madry A, Makelov A, Schmidt L, Tsipras D and Vladu A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks// International Conference on Learning Representations (ICLR)

Mentzer F, Toderici G, Minnen D, Hwang S-J, Caelles S, Lucic M and Agustsson E. 2022. VCT: A Video Compression Transformer// Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS), 35: 13091 - 13103

Mentzer F, Toderici G, Tschannen M and Agustsson E. 2020. High-fidelity Generative Image Compression//Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS), 33: 11913 - 11924

Minnen D, Ballé J and Toderici G. 2018. Joint Autoregressive and Hierarchical Priors for Learned Image Compression//Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS). Red Hook, NY, USA: Curran Associates Inc.: 10794 - 10803

Minnen D. and Singh S. 2020. Channel-Wise Autoregressive Entropy Models for Learned Image Compression. 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates: 3339 - 3343 [DOI:10.1109/ICIP40778.2020.9190935]

Mittal A, Soundararajan R and Bovik A C. 2013. Making a "Completely Blind" Image Quality Analyzer. IEEE Signal Processing Letters, 20(3): 209 - 212 [DOI:10.1109/LSP.2012.2227726]

Muckley M. J, El-Nouby A, Ullrich K, Jégou H. and Verbeek J. 2023. Improving Statistical Fidelity for Neural Image Compression with Implicit Local Likelihood Models//International Conference on Machine Learning (ICML). PMLR: 25426 - 25443

Nagel M, Fournarakis M, Amjad R, Bondarenko Y, Baalen M and Blankevoort T. 2021. A White Paper on Neural Network Quantization[EB/OL].[2021-06-15]. http://arxiv.org/abs/2106.08295

Nakanishi K M, Maeda S, Miyato T and Okanohara D. 2019. Neural Multi-Scale Image Compression//Asian Conference on Computer Vision (ACCV). Cham: Springer International Publishing: 718 - 732 [DOI:10.1007/978-3-030-20876-9_45]

Ortega A. and Ramchandran K. 1998. Rate-Distortion Methods for Image and Video Compression. IEEE Signal Processing Magazine, 15(6): 23 - 50 [DOI:10.1109/79.733495]

Watson A. B. 1998. Toward a Perceptual Video-Quality Metric [C]// Human Vision and Electronic Imaging III. SPIE: 139 - 147 [DOI: 10.1117/12.320105.]

Pan X, Ding G, Chen Z and Chen C. 2025. An Efficient Neural Rate Control for JPEG-AI. IEEE Transactions on Circuits and Systems for Video Technology, 1 - 1 [DOI: 10.1109/TCSVT. 2025. 3614007]

Ponomarenko N, Silvestri F, Egiazarian K, Carli M, Astola J and Lukin V. 2007. On Between-Coefficient Contrast Masking of DCT Basis Functions//Proceedings of the Third International Workshop on Video Processing and Quality Metrics. Scottsdale USA

Presta A, Tartaglione E, Fiandrotti A and Grangetto M. 2024. STanH : Parametric Quantization for Variable Rate Learned Image Compression. IEEE Transactions on Image Processing, 34: 639-651 [DOI: 10.1109/TIP.2025.3527883]

Ranjan A and Black M J. 2017. Optical Flow Estimation Using a Spatial Pyramid Network//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE: 2720-2729 [DOI:10.1109/CVPR.2017.291]

Rippel O and Bourdev L. 2017. Real-Time Adaptive Image Compression//Proceedings of the 34th International Conference on Machine Learning. Sydney, NSW, Australia: PMLR: 70:2922-2930

Rumelhart D E, Hinton G E and Williams R J. 1986. Learning Representations by Back-Propagating Errors. Nature, 323(6088): 533 - 536 [DOI:10.1038/323533a0]

Sha H, Dong M, Luo Q, Lu M, Chen H and Ma Z. 2025. Towards Loss-Resilient Image Coding for Unstable Satellite Networks. Proceedings of the AAAI Conference on Artificial Intelligence, 39(12): 12506-12514 [DOI:10.1609/aaai.v39i12.33363]

Sheikh H R and Bovik A C. 2006. Image Information and Visual Quality. IEEE Transactions on Image Processing, 15(2): 430 - 444 [DOI: 10.1109/TIP.2005.859378]

Sheng X, Li J, Li B, Li L, Liu D and Lu Y. 2023. Temporal Context Mining for Learned Video Compression. IEEE Transactions on Multimedia, 25: 7311 - 7322 [DOI:10.1109/TMM.2022.3220421]

Shi J, Lu M and Ma Z. 2023. Rate-Distortion Optimized Post-Training Quantization for Learned Image Compression. IEEE Transactions on Circuits and Systems for Video Technology, 34(5): 3082-3095 [DOI: 10.1109/TCSVT.2023.3323015]

Song M, Choi J and Han B. 2021. Variable-Rate Deep Image Compression through Spatially-Adaptive Feature Transform//IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE: 2360 - 2369 [DOI: 10.1109/ICCV48922. 2021.00238]

Spadaro G, Presta A, Tartaglione E, Giraldo J H, Grangetto M and Fiandrotti A. 2024. Gabic: Graph-Based Attention Block for Image Compression//IEEE International Conference on Image Processing (ICIP). IEEE: 1802 - 1808 [DOI: 10.1109/ICIP51287.2024. 10647413]

Su R, Cheng Z, Sun H and Katto J. 2020. Scalable Learned Image Compression With A Recurrent Neural Networks-Based Hyperprior// 2020 IEEE International Conference on Image Processing (ICIP). Abu Dhabi, United Arab Emirates: IEEE: 3369 - 3373 [DOI:10. 1109/ICIP40778.2020.9190704]

Sullivan G J and Wiegand T. 2005. Video Compression—From Concepts to the H.264/AVC Standard. Proceedings of the IEEE, 93(1): 18 - 31 [DOI:10.1109/JPROC.2004.839612]

Sun H, Cheng Z, Takeuchi M and Katto J. 2020. End-To-End Learned Image Compression With Fixed Point Weight Quantization//2020 IEEE International Conference on Image Processing (ICIP). Abu Dhabi, United Arab Emirates: IEEE: 3359 - 3363 [DOI:10.1109/ ICIP40778.2020.9190805]

Sun H, Yu L and Katto J. 2021. Learned Image Compression with Fixed-Point Arithmetic//2021 Picture Coding Symposium (PCS). Bristol, United Kingdom: IEEE: 1 - 5 [DOI: 10.1109/PCS50896.2021. 9477496]

Sun H, Yu L and Katto J. 2025. Q-LIC: Quantizing Learned Image Compression With Channel Splitting. IEEE Transactions on Circuits and Systems for Video Technology, 35(4): 3798 - 3811 [DOI: 10. 1109/TCSVT.2022.3231789]

Tang Z, Wang H, Yi X, Zhang Y, Kwong S and Kuo C-C J. 2022. Joint Graph Attention and Asymmetric Convolutional Neural Network for Deep Image Compression. IEEE Transactions on Circuits and Systems for Video Technology, 33(1): 421 - 433 [DOI: 10.1109/ TCSVT.2022.3187654]

Theis L, Shi W, Cunningham A and Huszár F. 2017. Lossy Image Compression with Compressive Autoencoders//5th International Conference on Learning Representations. Toulon, France

Toderici G, O'Malley S M, Hwang S J, Vincent D, Minnen D, Baluja S, Covell M and Sukthankar R. 2016. Variable Rate Image Compression with Recurrent Neural Networks//4th International Conference on Learning Representations. San Juan, Puerto Rico

Toderici G, Vincent D, Johnston N, Hwang S J, Minnen D, Shor J and Covell M. 2017. Full Resolution Image Compression with Recurrent Neural Networks//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI: IEEE: 5435 - 5443 [DOI:10.1109/CVPR.2017.577]

Tong K, Wu Y, Li Y, Zhang K, Zhang L and Jin X. 2023. QVRF: A Quantization-Error-Aware Variable Rate Framework for Learned Image Compression//2023 IEEE International Conference on Image Processing (ICIP). Kuala Lumpur, Malaysia: IEEE: 1310 - 1314 [DOI:10.1109/ICIP49359.2023.10222717]

Van Den Oord A and Vinyals O. 2017. Neural Discrete Representation Learning//Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS). Red Hook, NY, USA: Curran Associates Inc: 6309 - 6318

Wang L, Jiang Y, Wang T, Wang X and Dang J. 2025. Information Disentanglement-Based Self-Supervised Learning Speech Pretrained Large Model. Journal of Image and Graphics, 30(5):1272-1285 (王龙标, 江宇, 王天锐, 王晓宝, 党建武. 2025. 信息解耦式自监督预训练语音大模型. 中国图象图形学报, 30(5):1272-1285) [DOI: 10.11834/jig.240607]

Wang X, Zhang C, Ren W, Fu X, Zhou T, Zhao F and Shi Z. 2025. Introduction to Visual State Space Models and Applications. Journal of Image and Graphics, 30(10):3171-3172 (王兴刚, 张长青, 任文琦, 傅雪阳, 周涛, 赵峰, 石争浩, 陈秀妍. 2025.《中国图象图形学报》视觉状态空间模型及应用专栏简介. 中国图象图形学报, 30(10):3171-3172) [DOI: 10.11834/jig.2500010]

Wang X, Lu M and Ma Z. 2022. Block-Level Rate Control for Learnt Image Coding//2022 Picture Coding Symposium (PCS). San Jose, CA, USA: IEEE: 157 - 161 [DOI: 10.1109/PCS56426.2022.10018043]

Wang Z, Bovik A C, Sheikh H R and Simoncelli E P. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Transactions on Image Processing, 13(4): 600 - 612 [DOI: 10.1109/TIP.2003.819861]

Wang Z and Li Q. 2010. Information Content Weighting for Perceptual Image Quality Assessment. IEEE Transactions on Image Processing, 20(5): 1185 - 1198 [DOI: 10.1109/TIP.2010.2092435]

Wang Z, Simoncelli E P and Bovik A C. 2003. Multiscale Structural Similarity for Image Quality Assessment//The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. Pacific Grove, CA, USA: IEEE: 1398 - 1402 [DOI: 10.1109/ACSSC.2003.1292216]

Wei Y, Mao T, Li B, Wang F, Li F, Zhang Z and Zhao Y. 2025. Visual and Large Multimodal Models Promote Image Restoration and Enhancement: Research Progress. Journal of Image and Graphics, 30(5):1197-1219 (韦炎炎, 毛天一, 李柏昂, 王飞, 李锋, 张召, 赵洋. 2025. 视觉模型及多模态大模型推进图像复原增强研究进展. 中国图象图形学报, 30(5):1197-1219) [DOI: 10.11834/jig.240436]

Xie Y, Cheng K L and Chen Q. 2021. Enhanced Invertible Encoding for Learned Image Compression//Proceedings of the 29th ACM International Conference on Multimedia (MM '21). ACM: 162 - 170 [DOI:10.1145/3474085.3475213]

Xue N and Zhang Y. 2023. Lambda-Domain Rate Control for Neural Image Compression//Proceedings of the 5th ACM International Conference on Multimedia in Asia. Tainan Taiwan: ACM: 1 - 7 [DOI: 10.1145/3595916.3626372]

Yang C, Ma Y, Yang J, Liu S and Wang R. 2021. Graph-Convolution Network for Image Compression//IEEE International Conference on Image Processing (ICIP). Anchorage, AK, USA: IEEE: 2094 - 2098 [DOI: 10.1109/ICIP42928.2021.9506704]

Yang F, Herranz L, Cheng Y and Mozerov M G. 2021. Slimmable Compressive Autoencoders for Practical Neural Image Compression// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE: 4996 - 5005 [DOI:10.1109/CVPR46437.2021.00496]

Yang F, Herranz L, Weijer J V D, Guitian J A I, Lopez A M and Mozerov M G. 2020. Variable Rate Deep Image Compression With Modulated Autoencoder. IEEE Signal Processing Letters, 27: 331 - 335 [DOI:10.1109/LSP.2020.2970539]

Yang R, Mentzer F, Gool L V and Timofte R. 2021. Learning for Video Compression with Recurrent Auto-Encoder and Recurrent Probability Model. IEEE Journal of Selected Topics in Signal Processing, 15(2): 388 - 401 [DOI:10.1109/JSTSP.2020.3043590]

Yang W, Huang H, Hu Y, Duan L-Y and Liu J. 2024. Video Coding for Machines: Compact Visual Representation Compression for Intelligent Collaborative Analytics. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(7): 5174 - 5191 [DOI:10.1109/TPAMI.2024.3367293]

Yu J, Mai S, Zhang P, Jiang Y and Cheng J. 2025. Mixed-Precision Post-Training Quantization for Learned Image Compression. IEEE Internet of Things Journal, 12(16): 34392 - 34405 [DOI: 10.1109/JIOT.2025.3578318]

Yu Y, Wang Y, Yang W, Lu S, Tan Y-P and Kot A C. 2023. Backdoor Attacks Against Deep Image Compression via Adaptive Frequency Trigger// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: IEEE: 12250 - 12259 [DOI:10.1109/CVPR52729.2023.01179]

Zeng F, Tang H, Shao Y, Chen S, Shao L and Wang Y. 2025. MambaIC: State Space Models for High-Performance Learned Image Compression//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville TN, USA

Zhang H, Li L and Liu D. 2025. Generalized Gaussian Model for Learned Image Compression. IEEE Transactions on Image Processing, 34: 1950 - 1965 [DOI:10.1109/TIP.2025.3550013]

Zhang J, Jia C, Lei M, Wang S, Ma S and Gao W. 2019. Recent Development of AVS Video Coding Standard: AVS3//2019 Picture Coding Symposium (PCS). Ningbo, China: IEEE: 1 - 5 [DOI: 10.1109/PCS48529.2019.8954589]

Zhang L, Zhang L, Mou X and Zhang D. 2011. FSIM: A Feature Similarity Index for Image Quality Assessment. IEEE Transactions on Image Processing, 20(8): 2378 - 2386 [DOI: 10.1109/TIP.2011.

2109730]

Zhang Q, Mei J, Guan T, Sun Z, Zhang Z and Yu L. 2024. Recent Advances in Video Coding for Machines Standard and Technologies. ZTE Communications, 22(1): 62-76.

Zhang R, Isola P, Efros A A, Shechtman E and Wang O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lack City, USA: IEEE 586 - 595 [DOI: 10.1109/CVPR.2018.00068]

Zhang T, Luo X, Li L and Liu D. 2025. StableCodec: Taming One-Step Diffusion for Extreme Image Compression//Proceedings of the IEEE/CVF International Conference on Computer Vision.

Zhang X, Guo P, Lu M and Ma Z. 2024. All-in-One Image Coding for Joint Human-Machine Vision with Multi-Path Aggregation// Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, BC, Canada: Curran Associates Inc. :71465 - 71503 [DOI: 10.5555/3737916.3740199]

Zhang X, Lu M, Chen Y and Ma Z. 2025. Perception-Oriented Latent Coding for High-Performance Compressed Domain Semantic Inference//IEEE International Conference on Multimedia and Expo (ICME). Nantes, France: IEEE 1-6 [DOI: 10.1109/ICME59968.2025.11209906]

Zhang X and Wu X. 2023. LVQAC: Lattice Vector Quantization Coupled with Spatially Adaptive Companding for Efficient Learned Image Compression// 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: IEEE 10239-10248 [DOI: 10.1109/CVPR52729.2023.00987]

Zhang Y, Zhang R, Zhou H, Ji Q, Yu Z and Huang T. 2025. Research Status and Development Trends of Vision Foundation Models. Journal of Image and Graphics, 30(01): 0001-0024 (张燚钧, 张润清, 周华健, 齐骥, 余肇飞, 黄铁军. 2025. 视觉基础模型研究现状与发展趋势. 中国图象图形学报, 30(01): 0001-0024) [DOI: 10.11834/jig.230911]

Zhan R, Fan Y, Zhou L, Xie Y, Chen J, Yang H, Huang D and Wang Y. Dynamically Distribution-Aware Quantization for Diffusion Models[J/OL]. Journal of Image and Graphics, 2025, 1-10 (占瑞乙, 樊轶, 周丽娜, 谢宇宝, 陈佳鑫, 杨鸿宇, 黄迪, 王蕴红. 分布范围动态感知的扩散模型量化[J/OL]. 中国图象图形学报, 2025, 1-10 [DOI: 10.11834/jig.250319]

Zhen Y, Yu Z and Huang T. A Literature Review for Neural Networks-Based Encoding Models of Biological Visual System[J]. Journal of Image and Graphics, 2023, 28(2): 335-357 (郑雅菁, 余肇飞, 黄铁军. 生物视觉系统的神经网络编码模型综述[J]. 中国图象图形学报, 2023, 28(2): 335-357) [DOI: 10.11834/jig.220461]

Zheng J and Meister M. 2025. The Unbearable Slowness of Being: Why Do We Live at 10 Bits/s? Neuron, 113(2): 192 - 204 [DOI: 10.1016/j.neuron.2024.11.008]

Zheng Y, Chen Y, Qian B, Shi X, Shu Y and Chen J. 2025. A Review on Edge Large Language Models: Design, Execution, and Applications. ACM Computing Surveys, 57(8): 1 - 35 [DOI: 10.1145/3719664]

Zhou L, Sun Z, Wu X and Wu J. 2019. End-to-End Optimized Image Compression with Attention Mechanism//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. California, USA

Zhu Y, Yang Y and Cohen T. 2022. Transformer-Based Transform Coding//The Tenth International Conference on Learning Representations (Virtual)

Zhu W, Wang X, Tian Y and Go W. 2022. Multimedia Intelligence: When Multimedia Meets Artificial Intelligence. Chinese Journal of Image and Graphics, 27(9): 2551-2573 (朱文武, 王鑫, 田永鸿, 高文. 2022. 多媒体智能: 当多媒体遇到人工智能[J]. 中国图象图形学报, 27(9): 2551-2573) [DOI: 10.11834/jig.220086]

作者简介

陆明，男，副研究员，主要研究方向为智能视频编码。Email: minglu@nju.edu.cn

石峻奇，男，博士研究生，主要研究方向为信号表征。Email: junqishi@smail.nju.edu.cn

丛吾洋，男，博士研究生，主要研究方向为视频编码。Email: congwuyang@smail.nju.edu.cn

丁丹丹，女，教授，主要研究方向为多媒体通信。Email: DandanDing@hznu.edu.cn

贾川民，男，助理教授，主要研究方向为多媒体信号处理。Email: cmjia@pku.edu.cn

刘家瑛，女，副教授，主要研究方向为智能媒体计算与理解。Email: liujiaying@pku.edu.cn

刘东，男，教授，主要研究方向为图像视频编码。Email: dongeliu@ustc.edu.cn

宋利，男，副教授，主要研究方向为视频编码与数据压缩。Email: song_li@sjtu.edu.cn

马思伟，男，教授，主要研究方向为视频处理与编码。Email: swma@pku.edu.cn

杨铀，男，教授，主要研究方向为计算机视觉。Email: yangyou@hust.edu.cn

刘文予，男，教授，主要研究方向为机器学习与计算机视觉。Email: liuwy@hust.edu.cn

曹汛，男，教授，主要研究方向为图像视频处理与计算摄像学。Email: caoxun@nju.edu.cn