REVIEW ARTICLE

End-to-End Learned Image Coding: Foundations, Evolution, Practical Deployment, and Outlook

Ming Lu[1], Junqi Shi[1], Zhan Ma[1*]

[1]School of Electronic Science and Engineering, Nanjing University, Nanjing, Jiangsu, China.

*Corresponding author(s). E-mail(s): mazhan@nju.edu.cn;
Contributing authors: minglu@nju.edu.cn; junqishi@smail.nju.edu.cn;

Abstract

Image coding is a fundamental technology for visual communication, storage, and intelligent sensing. Over the past decade, end-to-end learned image coding (LIC) has shifted codec design from hand-crafted transforms and fixed decoder pipelines toward data-driven optimization of nonlinear transforms, quantization, entropy models, and reconstruction objectives. This article reviews still-image LIC from its variational-autoencoder foundations to its current deployment-oriented phase. We organize the field around two linked questions: how learned codecs obtain rate–distortion gains, and what is required for reliable practical use. We systematize the literature on learned transforms, quantization, entropy modeling, and rate–distortion objectives, then review variable-rate control, integer-precision inference, cross-platform decoder consistency, robustness, lightweight implementation, and standardization. We also introduce TinyLIC as an open benchmark and reproducible case study for rate–distortion–complexity analysis; under peak signal-to-noise ratio (PSNR)-based Bjontegaard delta-rate (BD-rate) evaluation, its low-complexity variant, measured at 10 thousand multiply–accumulate operations per pixel (KMACs/pixel), requires about 15% lower bitrate than the H.265/High Efficiency Video Coding (HEVC) image-profile anchor at equivalent PSNR on Kodak, while its 50 KMACs/pixel variant matches or slightly surpasses the H.266/Versatile Video Coding (VVC) image-profile anchor on several test sets. By connecting algorithmic evolution, deployment requirements, and reproducible benchmarking, this review provides a practical reference for industrial adoption and standardization, clarifying how learned image codecs can progress from rate–distortion gains to deployable and interoperable compression systems. Materials are accessible at https://njuvision.github.io/TinyLIC/.

Keywords: learned image coding, neural compression, rate–distortion optimization, perceptual compression, deployment, visual communication

1 Introduction

Image coding has long been a cornerstone of visual communication systems [158, 319]. It supports a wide range of Internet-based multimedia services, including social photo sharing, cloud storage, online advertising, video-on-demand, live streaming, and video conferencing. For more than three decades, mainstream image coding standards have largely followed a modular and rule-driven design philosophy [28, 298, 365]. Representative codecs improve compression efficiency by refining individual components such as transform [75], prediction [191], quantization [117], and entropy coding [316], while maintaining interoperability through carefully specified bitstream syntax and standardization procedures. This paradigm has achieved remarkable success, but its components are mostly hand-crafted and optimized within a constrained design space.

Over the past decade, image compression has undergone a paradigm shift driven by deep learning. By learning compact latent representations directly from data, end-to-end learned image coding (LIC) frameworks [57, 92, 139, 245, 258] have surpassed the Joint Photographic Experts Group (JPEG) standard [365], JPEG 2000 [208], and even image-profile anchors derived from modern video coding standards, including Better Portable Graphics (BPG; H.265/High Efficiency Video Coding (HEVC) image-profile anchor) and VTM Intra (H.266/Versatile Video Coding (VVC) image-profile anchor) [28, 298], on many public benchmarks in terms of rate–distortion (R–D) performance [83, 151].

This progress marks an important shift in the history of image coding. The advantage of learned codecs is not merely the use of deeper networks. Rather, it comes from the joint optimization of representation learning, probability modeling, quantization behavior, and reconstruction quality. This optimization framework also accommodates perceptual realism and machine-task-oriented objectives, extending LIC toward ultra-low-bitrate communication [180] and machine intelligence [94].

However, as LIC matures, the field's primary bottleneck is shifting. The central question is no longer only whether learned codecs can achieve better R–D curves, but also whether they can operate as reliable, controllable, efficient, and interoperable compression subsystems in real communication environments. Practical deployment imposes constraints that are often underrepresented in academic benchmarks, including computational complexity, decoding latency, rate-control flexibility, robustness to content variation and perturbations, and compatibility with heterogeneous hardware platforms [348].

This shift has redirected attention from performance-driven model design to system-oriented codec design. Recent studies have explored lightweight architectures [26], low-complexity entropy models [137], variable-rate and multi-rate coding [54], robustness under distribution shifts and adversarial perturbations [55], model quantization and integer inference [346], and hardware-friendly implementations [49]. In parallel, the standardization of JPEG Artificial Intelligence (JPEG AI)[1] [12] and IEEE 1857.11[2] [338] indicates that LIC is moving from academic prototypes toward industrial validation. These developments motivate a survey that treats compression efficiency, deployability, and interoperability as coupled design objectives.

[1] https://jpeg.org/jpegai/
[2] https://standards.ieee.org/ieee/1857.11/10691/

Table 1 Comparison of relevant learned image compression surveys.

Survey	Year	Core architecture	Perceptual coding	Practical functions	Performance evaluation	Reproducible benchmark
Ma *et al.* [267]	2019	✓	✗	✗	✗	✗
Liu *et al.* [238]	2020	✓	✓	✗	✓	✗
Ding *et al.* [83]	2021	✓	✗	✗	✓	✗
Hu *et al.* [151]	2021	✓	△	✗	✓	✓
Mishra *et al.* [277]	2022	✓	△	✗	✓	✗
Jamil *et al.* [160]	2023	✓	✗	✗	✓	✗
Zhang *et al.* [436]	2023	✓	✓	✗	✓	✗
Chen *et al.* [64]	2024	✓	✓	✗	✓	✗
Chen *et al.* [48]	2024	✓	✓	✗	✗	✗
Huang *et al.* [153]	2024	✓	✓	✗	✓	✗
Ours	**2026**	✓	✓	✓	✓	✓

Note: ✓ : clear coverage; ✗ : no coverage; △ : partial or limited coverage. A topic is marked as clear coverage when a survey contains a dedicated section, table, or repeated discussion; partial coverage denotes brief or indirect treatment.

This article is therefore written as a survey and tutorial on end-to-end LIC for still images. Its scope is the line of codecs that optimize analysis–synthesis transforms, quantization strategies, entropy models, and reconstruction objectives in an end-to-end manner, mainly from the resurgence of deep compression around 2015–2017 to the current standardization and deployment phase. Foundation-model-assisted compression, implicit representations for compression, as well as the extension to learned video coding are not surveyed exhaustively. They are discussed only when they clarify the boundaries and future trajectory of learned image coding.

1.1 Survey Criteria

This survey was assembled through a structured narrative review process covering literature from 2015 to June 2026, with emphasis on the period after end-to-end learned image compression became a reproducible research topic. We screened papers from major machine learning, computer vision, multimedia, signal processing, and communications venues, including NeurIPS, ICML, ICLR, CVPR, ICCV, ECCV, ACM Multimedia (ACM MM), DCC, PCS, ICIP, TIP, TMM, TCSVT, and related IEEE venues. We also considered arXiv preprints when they showed clear field influence through open-source adoption, repeated benchmark use, released implementations, or standardization-relevant evidence. Backward and forward citation tracing from representative codecs, standards documents, benchmark reports, and prior surveys was then used to refine both coverage and taxonomy.

Papers were included when they met at least one of the following criteria: (i) proposing a core LIC component, such as a transform, quantizer, entropy model, variable-rate mechanism, or optimization objective; (ii) reporting reproducible evaluation on public image-compression benchmarks, such as Kodak, Tecnick, Challenge on Learned Image Compression (CLIC), JPEG AI test material, or related common test sets; (iii) addressing deployment constraints, including complexity, latency, quantization, robustness, rate control, network resilience, or cross-platform consistency; or (iv) contributing to standards, common test conditions, or reference software, including JPEG AI and IEEE 1857.11.

Given this context, this survey centers on end-to-end LIC as a representative and relatively mature branch of learned visual compression. Compared with video coding, still-image coding currently benefits from more stable benchmarks, lower system complexity, more concrete standardization progress to date, and stronger near-term deployment potential. It therefore provides a tractable entry point for understanding how learned compression is evolving from algorithmic innovation to practical codec design. Rather than listing R–D gains alone, this survey emphasizes the transition from neural compression models to deployable compression systems.

To provide a structured overview, we organize end-to-end LIC from two complementary perspectives. The *architectural perspective* explains how learned transforms, signal quantization, entropy models, and optimization objectives interact to determine coding efficiency. The *system perspective* explains how variable-rate control, cross-platform consistency, robustness, lightweight deployment, and standardization determine whether a learned codec can function in real communication pipelines. This two-level taxonomy connects the historical evolution of neural codec architectures with the emerging requirements of practical deployment. Accordingly, Sec. 2 focuses on architectural foundations, Sec. 3 focuses on system-level deployment issues, and Sec. 4 connects the two through a reproducible deployment-oriented case study.

The main scope and contributions of this survey are summarized as follows.

- **Architectural foundations and taxonomy** (Sec. 2): We review the evolution of end-to-end LIC through its core components and explain how analysis–synthesis transforms, signal quantization, entropy modeling, and R–D objectives jointly produce coding gains.
- **System-level deployment analysis** (Sec. 3): We examine system-level issues that determine practical viability, including variable-rate coding, rate control, model quantization, cross-platform consistency, robustness, lightweight deployment, and standardization progress.
- **Reproducible case study and open benchmark** (Sec. 4): We introduce TinyLIC as a lightweight reference platform for studying practical LIC design under controlled conditions. The benchmark is used to analyze rate–distortion–complexity trade-offs, controllability, robustness, progressive transmission, loss resilience, and deployability.
- **Outlook and open problems**: We identify and discuss open challenges in coding for intelligence, neural scene representations, foundation-model-assisted compression, complexity, and interoperable standards.

Existing reviews have played important roles in mapping the early development of neural visual coding. Ma *et al.* [267] reviewed neural-network-based image and video compression at a stage when many methods were still connected to traditional coding modules, while Liu *et al.* [238] and Ding *et al.* [83] emphasized deep tools and case studies for video coding systems. These works clarified how neural networks can improve prediction, transform, filtering, post-processing, and early end-to-end coding pipelines. However, their main focus is the transition from hand-crafted or hybrid

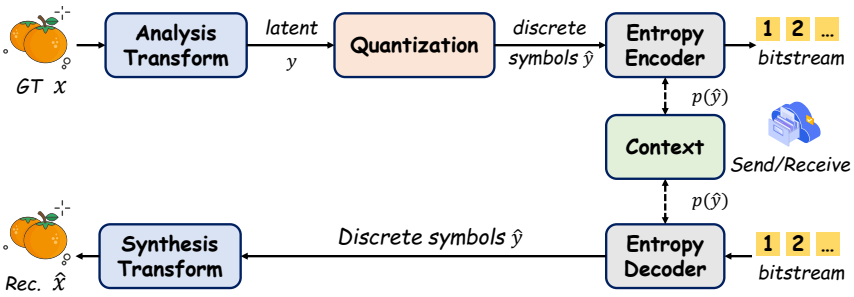


Fig. 1 A variational-autoencoder (VAE)-based LIC. The analysis transform maps the input image into latents that are then quantized and entropy-coded. The decoder reconstructs the image using the entropy-decoded latents and the synthesis transform.

codecs to learning-based coding tools, rather than the deployment requirements of mature end-to-end still-image codecs.

More recent image-oriented surveys provide broader coverage of learned image compression architectures and benchmarks. Hu *et al.* [151] offered an influential benchmark and organized end-to-end LIC around network architectures, entropy models, and rate control; Mishra *et al.* [277] and Jamil *et al.* [160] further summarized deep architectures and learning-driven lossy image compression methods. In parallel, perceptual and generative coding surveys [48, 64, 436] focus on visual quality, generative priors, and perceptual optimization, while Huang *et al.* [153] connects LIC with human-machine coding and emerging standards. These reviews provide valuable algorithmic and conceptual context, but deployment-oriented issues such as integer inference, decoder consistency, entropy-coder runtime, robustness, rate-control guarantees, lightweight implementation, and standard conformance are usually discussed separately or only as future challenges.

This survey therefore extends prior work by treating deployability as a central organizing dimension rather than an afterthought. We first review the architectural mechanisms that produce R–D gains, and then examine how these mechanisms interact with practical constraints required by real communication, storage, and edge-intelligence systems. The inclusion of TinyLIC further turns this discussion into a reproducible case study for rate–distortion–complexity, controllability, robustness, progressive transmission, and loss resilience. Table 1 summarizes these differences from representative prior surveys.

2 Core Framework

For clarity, this survey uses *learned image coding* (LIC) to denote still-image compression systems whose transform, entropy model, and reconstruction objective are optimized from data. We use *end-to-end LIC* when the analysis transform, differentiable quantization relaxation, entropy model, and synthesis transform are trained jointly under a coding objective. The term *neural image coding* is used only as a broader synonym when discussing the literature. Unless otherwise stated, quantization in Sec. 2 refers to *signal quantization* of latent variables, whereas model quantization

Table 2 Frequently used acronyms and metrics in this survey.

Acronym	Description
General concepts and model families	
LIC	Learned image coding
VAE	Variational autoencoder
CNN	Convolutional neural network
GAN	Generative adversarial network
Objectives, metrics, and complexity	
R–D	Rate–distortion
R–D–P	Rate–distortion–perception
BPP	Bits per pixel
MSE	Mean squared error
PSNR	Peak signal-to-noise ratio
MS-SSIM	Multi-scale structural similarity
LPIPS	Learned perceptual image patch similarity
FID	Frechet inception distance
BD-rate	Bjontegaard delta rate
CTC	Common test conditions
KMACs/pixel	Thousand multiply–accumulate operations per pixel
Codecs, standards, and anchors	
HEVC	High Efficiency Video Coding
VVC	Versatile Video Coding
BPG	H.265/HEVC image-profile anchor
VTM Intra	VVC Test Model in intra/image-profile configuration
JPEG AI	JPEG Artificial Intelligence standard
IEEE 1857.11	Neural network-based image coding standard

for hardware deployment is discussed separately in Sec. 3.2. Methods focused primarily on learned post-processing, learned enhancement, purely lossless compression, neural rendering, or implicit scene representation are outside the core scope unless they directly clarify still-image LIC mechanisms or deployment boundaries.

In rate–distortion theory, high-dimensional vector quantization (VQ) can approach optimal compression performance [118, 125], but its practical use is severely limited by the curse of dimensionality. As the source dimension increases, both the size of the reproduction codebook and the complexity of nearest-codeword search grow exponentially [22]. For high-dimensional signals such as images, prevailing image codecs therefore rely on transform coding, which decomposes compression into a set of more tractable modules: transform, quantization, and entropy coding. These modules are jointly designed and optimized under the R–D principle [123, 384], forming the foundation of modern image compression systems.

Since the mid-20th century, researchers and engineers have primarily relied on physical principles and statistical methods to design and optimize the aforementioned modules. Representative milestones include entropy-optimal prefix codes such as Huffman coding [156], and energy-compacting linear transforms such as the Discrete Cosine Transform (DCT) [5]. While early attempts to incorporate neural networks into image coding dated back to the late 1980s [74, 81, 88, 171], their impact remained limited due to a lack of efficient training methods and computational resources. It was not until

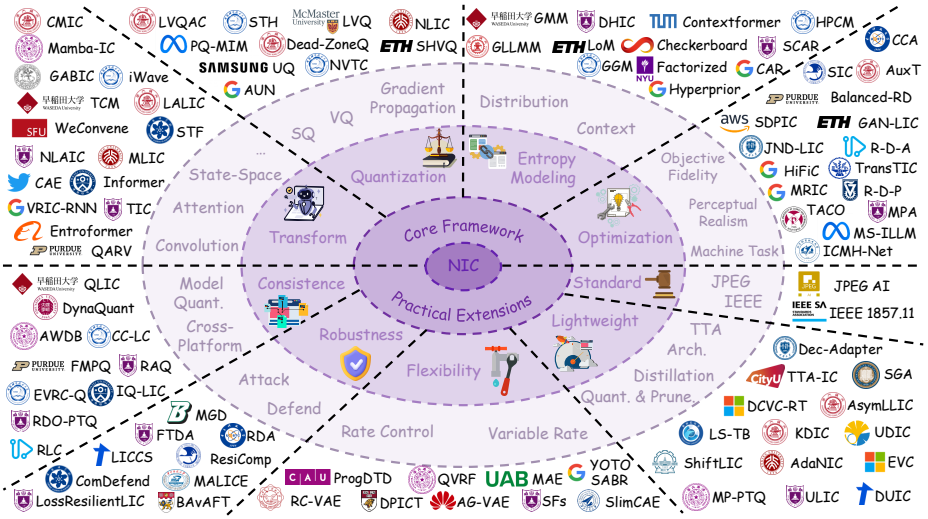


Fig. 2 Taxonomy of representative works in end-to-end LIC. Existing studies are organized from two complementary perspectives: core framework, including transform, quantization, entropy modeling, and optimization; and practical extensions, including flexibility, robustness, consistency, standardization, etc.

the resurgence of deep learning that end-to-end LIC solutions began to demonstrate competitive performance, marking a turning point around 2015–2017 [19, 56, 354].

Most end-to-end LIC frameworks are built upon the variational autoencoder (VAE) paradigm [124, 189], which offers a probabilistic interpretation of transform coding. As illustrated in Fig. 1, these frameworks retain the classical coding pipeline of transform, quantization, and entropy coding, but implement its key components using neural networks and optimize them jointly. In this formulation, an analysis transform f_θ maps an input image x into a latent representation $y = f_\theta(x)$, which is then quantized into discrete symbols $\hat{y}$. These symbols are compressed by an entropy coder according to a learned probability model, while a synthesis transform g_θ reconstructs the image as $\hat{x} = g_\theta(\hat{y})$. At the communication-system level, the learned encoder and entropy coder must produce a standards- or implementation-specific bitstream that satisfies storage, channel, network, and decoder-consistency constraints before the reconstruction is consumed by either human viewers or machine-vision modules.

The main strength of the VAE formulation lies in its ability to unify representation learning and probabilistic modeling within a single end-to-end optimization framework. In practice, learned codecs are trained by minimizing a Lagrangian R–D objective that balances reconstruction distortion and expected code length. This formulation generalizes classical transform coding while providing a flexible foundation for incorporating learned priors, context models, adaptive quantization, and increasingly expressive neural transforms.

In the following sections, we review transform design, quantization mechanisms, entropy modeling, rate estimation, and end-to-end optimization strategies.

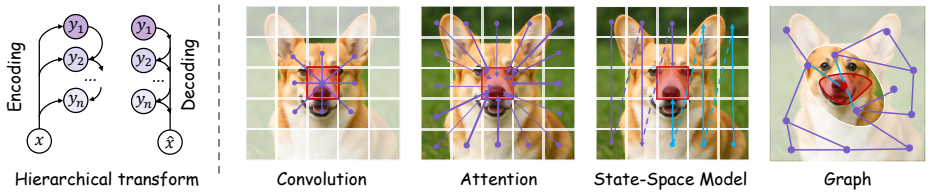


Fig. 3 Representative transform paradigms in LIC, including hierarchical, convolutional, attention-based, state-space, and graph-based transforms. These designs differ in their receptive fields, dependency modeling mechanisms, and computational characteristics: convolution aggregates local neighborhoods, attention establishes explicit long-range interactions, state-space models propagate context along scan paths, and graph transforms capture irregular relations through nodes and edges. The red box denotes the target region; lines indicate dependency connections, which may be local, global, or irregular depending on the transform; arrows represent sequential state propagation.

2.1 Transform

Similar to classical DCT or wavelets, the learned analysis transform f_θ aims to produce representations that are compact and approximately decorrelated, thereby facilitating subsequent quantization and entropy coding. Unlike handcrafted linear transforms, neural transforms are highly nonlinear and data-adaptive, enabling more flexible approximation of complex image manifolds, thereby yielding improved R–D performance.

2.1.1 Convolutional transforms

Since the inception of modern LIC, convolutional neural networks (CNNs) have served as the dominant backbone for learned transforms, owing to their strong inductive bias toward local spatial correlations and their computational efficiency on modern hardware. The origin of this line of research can be traced back to around 2015, when works such as [18, 23] demonstrated that deep convolutional architectures can learn effective nonlinear analysis–synthesis mappings directly from data, often surpassing traditional codecs at comparable bitrates.

Subsequent studies [7, 17, 19, 66, 67, 169, 170, 218, 219, 352] progressively improved CNN-based transforms through deeper architectures, residual connections, content-adaptive modulation, enlarged receptive fields, and carefully designed nonlinearities. Despite their effectiveness, conventional CNN-based transforms remain primarily biased toward local aggregation, making it difficult to explicitly capture long-range spatial dependencies. To mitigate this limitation, non-local-attention-based image compression (NLAIC) [57, 241] incorporated non-local attention into CNN-based transforms, allowing distant spatial positions to interact directly within an otherwise convolutional architecture. This represents an important intermediate step from purely local convolutional transforms toward hybrid designs with explicit global-context modeling. It also points to the later shift toward Transformer-based transforms, where self-attention becomes a central mechanism for representation learning.

2.1.2 Transformer-based transforms

Building upon this trend toward global-context modeling, Transformer architectures [253, 292, 363] have been increasingly introduced into neural image compression. Around 2022, Transformer-based LIC [14, 186, 194, 257, 258, 274, 303, 451, 453] emerged as a major research direction, introducing sequence modeling paradigms into transform design. By leveraging self-attention, these models capture long-range dependencies and dynamically adapt their receptive fields, yielding more expressive latent representations than purely convolutional approaches. This is particularly advantageous for high-resolution images and structured visual content.

However, the quadratic complexity of standard self-attention limits scalability in practical deployment. To address this issue, a variety of efficient architectures have been proposed, including window-based attention [453], hierarchical attention [186], and spatio-channel attention [194, 437]. Hybrid CNN–Transformer architectures [245] further attempt to balance local inductive bias and global modeling capability. More recently, linear-complexity attention mechanisms [100] have been explored to improve scalability for high-resolution and resource-constrained scenarios.

2.1.3 Mamba-based transforms

Beyond Transformers, state-space models (SSMs) have recently been explored as an alternative paradigm for learned transform design. Instead of explicitly computing pairwise token interactions as in self-attention, SSMs aggregate contextual information through recurrent state transitions, offering linear complexity with respect to sequence length. Built upon selective SSMs, Mamba further introduces input-dependent state propagation, enabling content-adaptive long-range modeling with improved efficiency. Recent Mamba-based codecs, including MambaIC [416], MambaVC [304], CASSIC [305], and CMIC [60], incorporate such selective state-space representations into LIC to efficiently capture long-range spatial dependencies. Early evidence suggests that these methods may offer a trade-off between the representation capability of Transformers and the efficiency of CNN-based architectures, although their deployment maturity remains less established than that of CNN–hyperprior families.

2.1.4 Hierarchical transform paradigm

While most LIC models adopt a single-scale VAE architecture, more recent approaches have explored hierarchical transform designs to improve representation capacity. In this paradigm, an image is encoded into a hierarchy of latent variables at multiple spatial resolutions[3]. Such representations naturally support coarse-to-fine modeling, where low-resolution latents capture global structures and higher-resolution latents progressively refine local details.

Early attempts in this direction [37, 150, 281] introduced coarse-to-fine refinement frameworks by aggregating multi-scale latent representations through concatenation. More principled approaches, such as the QARV series [92, 93, 440], explicitly

[3]Unlike hyperprior or autoregressive latent variables used solely for entropy estimation, these hierarchical latents directly contribute to image reconstruction.

model latent variables as additive residual components across multiple layers. This hierarchical decomposition enables progressive reconstruction and reduces reliance on computationally expensive spatial autoregressive entropy models. Extensions to video coding further integrate spatial and temporal hierarchies within a unified framework [259, 260], demonstrating the generality of hierarchical representation learning.

2.1.5 Alternative transform architectures

In parallel, several alternative transform paradigms have been explored. They can be viewed as complementary ways to introduce non-grid dependency modeling, invertible likelihood-preserving mappings, or frequency/subband priors into learned transforms. Graph neural networks (GNNs) [61, 337, 349, 395] model images as irregular graphs, enabling a more flexible representation of structural relationships than grid-based convolutions. Invertible neural networks and normalizing flows [109, 114, 143, 377, 386] aim to construct bijective transforms, enabling theoretically information-preserving latent mappings and exact likelihood estimation.

Another line of work explores frequency-domain representations. Methods such as [110, 172, 368] and FTIC [215] introduce frequency-aware transform designs to better capture frequency-dependent structures, while iWave [266] and WeConvene [107] explicitly leverage wavelet-domain representations for multi-scale subband decomposition.

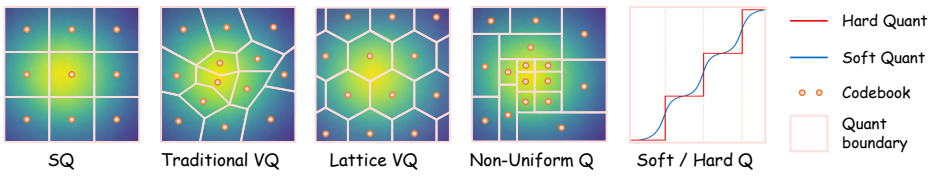


Fig. 4 Illustration of representative quantization mechanisms in LIC. Scalar quantization (SQ) uses independent scalar intervals, while vector quantization (VQ) and lattice VQ form learned or structured partitions of the latent space. Non-uniform quantization adapts decision boundaries to latent statistics, and soft-to-hard quantization bridges differentiable training and hard discretization. Orange dots denote codewords and pale lines denote quantization boundaries.

2.1.6 Discussion and Perspective

The evolution of transform design in LIC reflects a clear transition from handcrafted local transforms to highly expressive, data-driven nonlinear transforms. Existing approaches increasingly converge toward three major directions: (i) global context modeling, (ii) hierarchical representation learning, and (iii) resource-efficient computation.

Despite substantial progress, several fundamental challenges remain open. First, hierarchical representations are often difficult to optimize due to redundancy in representations across latent layers. Recent work such as DHIC [78] attributes this issue to insufficient inter-layer disentanglement and introduces explicit regularization to

enforce frequency-aware coarse-to-fine decomposition. Interestingly, DHIC achieves state-of-the-art performance using a purely CNN-based backbone, suggesting that a well-structured representation hierarchy can reduce the reliance on more complex backbone architectures such as Transformers.

Second, an inherent trade-off persists between global modeling capability and computational efficiency. Although Transformers and SSMs substantially improve representation power, their practical deployment is still constrained by latency, memory consumption, and hardware compatibility. Consequently, CNN-based models like PICO [351] remain highly competitive in real-world systems, particularly for low-latency and resource-constrained applications. A practical design implication is that CNN-based transforms remain attractive for low-latency deployment, attention or SSM modules are useful when long-range dependency modeling dominates, and hierarchical designs provide a natural bridge to scalability and progressive reconstruction.

2.2 Quantization

Quantization[4] is the key operation that bridges continuous neural representations and discrete symbols in end-to-end LIC. It maps real-valued latent variables into a finite set of symbols, making them amenable to lossless entropy coding. As such, quantization directly determines not only the distortion introduced by discretization, but also the statistical structure of the symbols to be modeled and compressed.

Compared with classical codecs, quantization in neural image compression introduces two coupled challenges. The first is an optimization challenge: hard quantization, typically implemented by rounding or nearest-neighbor assignment, is non-differentiable and therefore incompatible with standard gradient-based training. The second is a compact representation challenge: the choice of quantization granularity, partition geometry, and codebook structure strongly affects rate–distortion efficiency, entropy-modeling difficulty, and computational complexity.

These two challenges determine how we organize the following discussion. We first review scalar quantization (SQ), because it is the dominant choice in modern LIC and exposes the training–inference mismatch most directly: discrete rounding is used at test time, but a differentiable relaxation is needed during learning. This motivates the discussion of additive noise, universal quantization, straight-through estimation, and soft-to-hard relaxation as training strategies for SQ-based codecs.

We then turn to methods that change the quantization structure itself. Vector quantization (VQ), product quantization (PQ), and lattice vector quantization (LVQ) increase representational capacity by capturing inter-dimensional dependencies or imposing structured partitions, usually at higher optimization or search cost. Non-uniform and adaptive quantizers instead reshape decision intervals according to latent statistics, acting as implicit bit-allocation mechanisms. This organization links the optimization issue to the first subsubsection and the representation issue to the second.

[4] In this subsection, we focus on *signal quantization*, while model quantization for deployment is discussed separately in Sec. 3.2.

2.2.1 Scalar Quantization and Training Relaxations

Scalar quantization (SQ) remains the dominant paradigm in neural image compression due to its simplicity and efficiency. In SQ, each latent element is quantized independently, typically by rounding it to the nearest integer or predefined level. However, scalar rounding is non-differentiable, which prevents direct gradient-based optimization. Therefore, most SQ-based learned codecs rely on differentiable training relaxations to reconcile discrete inference with gradient-based learning.

Additive Uniform Noise (AUN). A seminal strategy is to replace hard rounding with additive uniform noise during training. Specifically, the quantized latent $\hat{y} = \mathrm{ROUND}(y)$ is approximated as

$$\tilde{y} = y + u, \quad u \sim \mathcal{U}(-0.5, 0.5), \tag{1}$$

as introduced by Ballé *et al.* [19, 23]. This formulation yields a differentiable relaxed objective and enables end-to-end rate–distortion optimization. However, the resulting rate term corresponds to the differential entropy of the perturbed variable rather than the discrete entropy of quantized latents, leading to an objective mismatch between training and actual coding.

Universal Quantization (UQ). To alleviate this mismatch, Choi *et al.* [73] introduced universal quantization [415] with subtractive dithering:

$$z = \mathrm{ROUND}(y + u) - u, \quad u \sim \mathcal{U}(-0.5, 0.5). \tag{2}$$

This formulation preserves exact quantization in the forward pass while making the quantization error statistically independent of the source. It establishes an equivalence between the discrete entropy of quantized variables and the differential entropy of their noise-perturbed counterparts. As a result, the rate optimized during training becomes more consistent with the true coding rate, mitigating the mismatch inherent in AUN.

Straight-Through Estimator (STE). STE [29] performs hard rounding in the forward pass while replacing the true derivative of the quantizer with an identity mapping during backpropagation, e.g., $\partial\hat{y}/\partial y \approx 1$. Although simple and computationally efficient, STE introduces biased gradients and ignores the statistical effect of quantization noise, which may lead to suboptimal convergence.

Soft-to-Hard Annealing. Soft-to-hard strategies introduce differentiable relaxations that gradually approach hard scalar quantization during training. For example, SHVQ [2] employs temperature-controlled soft assignments that progressively converge to nearest-neighbor decisions. Similarly, STH [131] first trains with soft quantization and then fine-tunes with hard quantization, thereby reducing the gap between differentiable optimization and discrete inference. Further analyses of differentiable quantization relaxations are provided in [1, 420].

Gradient Correction. Recent work revisits scalar quantization from an optimization perspective. For instance, [297] identifies a gradient mismatch between entropy modeling and quantization, and introduces correction terms into the loss function to better align training dynamics with actual coding objectives.

Overall, SQ-based learned codecs benefit from a simple discretization rule and mature differentiable training relaxations. When combined with expressive entropy models, SQ provides an effective solution and has therefore become the de facto standard in modern neural image compression.

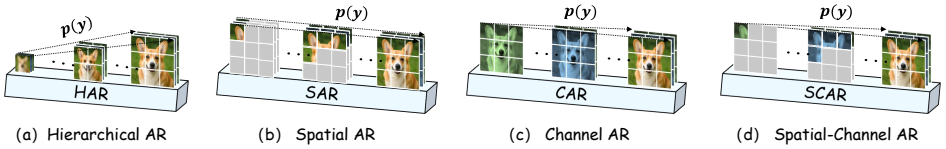
2.2.2 Vector Quantization (VQ) and Its Variants

In contrast to SQ, vector quantization (VQ) jointly quantizes groups of latent variables by mapping each latent vector to an entry in a learned codebook. Formally, a latent vector is replaced by its nearest codeword, allowing VQ to capture inter-dimensional dependencies that scalar quantization ignores. From a rate–distortion perspective, VQ is theoretically attractive because it can realize more flexible partitions of the latent space than independent scalar quantization. However, these benefits come at the cost of increased training complexity, codebook collapse risks, and higher encoding latency. A further difficulty is that hard nearest-codeword assignment makes it challenging for the entropy loss to directly shape the encoder-induced index distribution, which weakens the coupling between codebook learning and entropy coding. As a result, many VQ variants aim to preserve the expressiveness of joint quantization while improving trainability, and computational efficiency.

Product Quantization (PQ). Product quantization provides a compromise between scalar quantization and full vector quantization by decomposing high-dimensional latent vectors into multiple low-dimensional subspaces, each quantized independently with a small codebook [161]. This decomposition substantially reduces the complexity of nearest-neighbor search while preserving part of VQ's dependency-modeling capability. Recent studies have explored PQ in neural compression. For example, NVTC [102] introduces hierarchical PQ to improve representation efficiency, while PQ-MIM [97] integrates PQ into masked image modeling for perceptual compression.

Lattice Vector Quantization (LVQ). Lattice VQ replaces learned codebooks with structured lattice points, enabling efficient tessellation of the latent space [164]. Compared with scalar quantization, LVQ yields more efficient Voronoi partitions and can improve rate–distortion performance while maintaining relatively low computational complexity. However, classical LVQ relies on fixed lattice structures, which limits its adaptability to non-stationary latent distributions. To address this limitation, LVQAC [431] combines LVQ with spatially adaptive companding, allowing the effective quantization resolution to vary across spatial locations. More recently, adaptive lattice VQ [388] further generalizes this idea by dynamically adjusting lattice density to support variable-rate compression.

Non-Uniform and Adaptive Quantization. Beyond uniform quantization, non-uniform and adaptive schemes aim to better match the distribution of latent variables. Dead-zone quantizers [222] allocate larger quantization intervals around zero, which is particularly effective for sparse latent representations. Contextual sequential quantization (CSQ) [116] further introduces spatially and contextually adaptive quantization levels, enabling finer discretization in perceptually or semantically important regions. These methods can be interpreted as implicit bit allocation mechanisms, where



(a) Hierarchical AR (b) Spatial AR (c) Channel AR (d) Spatial-Channel AR

Fig. 5 Representative context modeling strategies for entropy estimation. Hierarchical, spatial autoregressive, channel autoregressive, and spatial–channel autoregressive models exploit different dependency structures to improve probability estimation of quantized latents.

quantization precision is modulated according to local signal statistics to improve rate–distortion efficiency.

2.2.3 Discussion and Perspective

The evolution of quantization in neural image compression reflects a fundamental trade-off among representation capability, optimization tractability, and entropy-modeling efficiency. Scalar quantization has become the dominant paradigm due to its simplicity, scalability, and compatibility with modern entropy models, while differentiable relaxations such as additive noise relaxation and straight-through estimation enable end-to-end optimization despite the non-differentiability of rounding.

In contrast, vector quantization provides more expressive discrete representations and is theoretically appealing from a rate–distortion perspective, but its practical adoption remains limited by codebook complexity, optimization instability, and entropy-modeling difficulty. In particular, hard nearest-codeword assignment prevents the entropy loss from directly shaping the encoder-induced index distribution, making joint optimization of codebook learning and entropy coding challenging. Recent works, such as RDVQ [173], revisit VQ from an explicit rate–distortion optimization perspective by introducing differentiable relaxations for discrete codebook assignments, thereby improving the coupling between token distribution learning and entropy modeling.

Meanwhile, progress in VQ-based generative modeling and ultra-low-bitrate compression has renewed interest in structured discrete latent representations [43, 99, 181, 192]. A promising direction is to combine the scalability and entropy-modeling friendliness of scalar quantization with the representation power of structured VQ, while maintaining stable and rate-aware end-to-end optimization.

2.3 Entropy Modeling

Entropy modeling aims to characterize the probability distribution of latent variables such that the estimated distribution closely matches the true one, thereby minimizing statistical redundancy. The expected coding rate can be expressed as

$$R = \mathbb{E}_{\hat{y}\sim q(\hat{y})}[-\log_2 p_\psi(\hat{y})], \tag{3}$$

where $q(\hat{y})$ denotes the empirical distribution of quantized latents and $p_\psi(\hat{y})$ denotes the learned entropy model. Therefore, the effectiveness of learned image compression critically depends on how accurately $p_\psi(\hat{y})$ approximates $q(\hat{y})$. For a full codec, this

cross-entropy term is combined with auxiliary-rate terms, such as the bitrate of hyper-latents, and with the discrete probability tables or cumulative distribution functions used by the actual entropy coder.

Early neural compression methods relied on fixed, non-adaptive entropy coding based on empirical symbol statistics. For instance, DeepCoder [56] directly applied Huffman coding to quantized deep features through piece-wise histogram estimation. However, such approaches fail to capture contextual dependencies among latent variables, limiting compression efficiency.

Modern LIC has shifted toward learnable entropy modeling, which can be broadly decomposed into two components: (i) a *distribution model* that specifies the parametric form of $p_\psi(\hat{y})$, and (ii) a *context model* that conditions this distribution on auxiliary side information or previously decoded symbols to better approximate $q(\hat{y})$.

2.3.1 Parametric Distribution

The choice of distribution family determines the expressiveness of the entropy model. Early works adopt simple parametric distributions, while recent approaches explore increasingly flexible formulations.

Basic unimodal distributions. The Gaussian model (GM) is the most widely adopted choice, assuming

$$p(y|\mu,\sigma) = \mathcal{N}(y;\mu,\sigma^2), \tag{4}$$

and serves as the default in most modern codecs [276], mainly due to its analytical tractability.

Alternative unimodal distributions have also been explored. The Laplacian model (LaM) exhibits sharper peaks and heavier tails, making it better suited for sparse or edge-dominated latents [447]. The logistic distribution model (LoM) and its discretized variant [272] provide improved flexibility for modeling quantized variables, particularly in lossless or near-lossless settings.

Mixture models. To capture complex latent distributions beyond a single component, mixture models have become increasingly popular. A representative example is the Gaussian mixture model (GMM):

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

which significantly improves modeling capacity for textured and non-stationary regions [57, 68, 450].

Subsequent works further generalize this idea by combining multiple distribution families, leading to hybrid models such as Gaussian–Logistic–Laplacian mixtures model (GLLMM) [105], which demonstrate stronger fitting capability across diverse latent statistics. More recently, generalized Gaussian models (GGM) [422] provide a unified formulation with adaptive shape parameters, enabling better modeling of skewness, kurtosis, and sparsity.

2.3.2 Context Modeling

Context modeling aims to exploit dependencies among latent variables to improve probability estimation. Its evolution reflects a progression from independence assumptions to increasingly expressive conditional models.

Factorized models. The earliest end-to-end framework [19] assumes independence among latent variables:

$$p(y) = \prod_i p(y_i), \tag{6}$$

which enables simple and fully parallelizable entropy coding. This assumption neglects spatial and channel correlations, resulting in suboptimal rate–distortion performance.

Hyperprior models. To capture global statistical dependencies, Ballé *et al.* [21] introduced the hyperprior, which models the distribution of latents conditioned on auxiliary variables z:

$$p(y|z) = \prod_i p(y_i|z). \tag{7}$$

Here, z is generated via a hyper encoder (typically with spatial downsampling) and conveys side information such as local mean and scale. Due to its significantly lower spatial resolution, z is usually modeled with a factorized prior at a relatively low bitrate cost, thereby substantially improving entropy estimation accuracy and reducing the overall bitrate. Hyperpriors can therefore be interpreted as *learned side information* capturing global statistics, and have become a standard component in modern LIC frameworks.

Autoregressive (AR) models. While hyperpriors capture global dependencies, they are less effective at modeling fine-grained local correlations. Autoregressive (AR) models address this limitation by conditioning each latent variable on previously decoded neighbors [276]:

$$p(y) = \prod_i p(y_i|z, y_{<i}). \tag{8}$$

Depending on the spatial-channel dependency structure, AR models can be categorized into:

- **Spatial Autoregressive (SAR)**, modeling local spatial neighborhoods [210, 276];
- **Channel Autoregressive (CAR)**, capturing inter-channel dependencies [275];
- **Spatial–channel Autoregressive (SCAR)**, jointly modeling both dimensions via 3D convolutions or grouped designs [57, 220].

Despite their strong modeling capabilities, AR models suffer from inherently sequential decoding, which limits parallelism. To alleviate this issue, various acceleration strategies have been proposed, including checkerboard masking [103, 106, 137],

channel grouping [136, 139, 275], and hybrid schemes [174, 175, 194, 195, 214, 257] that trade off dependency modeling and decoding efficiency. These methods aim to approximate full autoregression using a limited number of parallelizable steps.

Hierarchical autoregressive (HAR) models. Recent advances extend entropy modeling to hierarchical and multi-scale settings, inspired by hierarchical VAEs [356, 360]. In HAR, the latent space is decomposed into multiple levels:

$$p(y) = \prod_{l=1}^{L} p(y^{(l)}|y^{(<l)}), \tag{9}$$

where higher-level latents capture global structures and lower-level latents encode fine details.

Two distinct paradigms are identified. In *reconstruction-oriented* HAR, hierarchical latents directly contribute to image reconstruction, often in a residual [78, 92, 150, 450]. These models jointly improve representation and entropy modeling. In contrast, *entropy-oriented* hierarchies (e.g., HPCM [228]) introduce multiscale latent variables solely to facilitate probability estimation, without contributing to reconstruction.

2.3.3 Discussion and Perspective

More broadly, entropy modeling in learned image compression is closely related to generative modeling. Both aim to approximate the underlying data distribution, but differ in their objectives: generative models focus on *sampling fidelity*, whereas compression models emphasize *accurate likelihood estimation*. In particular, learned compression requires calibrated probabilities that can be converted into code lengths by an entropy coder, whereas many generative models can tolerate approximate likelihoods when sample quality is the main objective. From this perspective, learned compression can be viewed as a constrained form of probabilistic generative modeling under rate–distortion optimization.

This connection explains why advances in generative modeling have consistently influenced the design of entropy models. Early compression frameworks were built upon VAEs, which provide a natural probabilistic formulation for rate estimation. Subsequent works have incorporated ideas from autoregressive models and generative adversarial networks (GANs) to improve perceptual quality and dependency modeling. More recently, emerging paradigms such as diffusion models [142, 334], normalizing flows [262, 311], and frequency-domain autoregressive modeling [65] offer new opportunities for more accurate and flexible density estimation.

2.4 Rate-Distortion Optimization

Most modern LICs can be interpreted as variational compression models trained by backpropagation [314]. Under common VAE-style formulations [33, 124], the training objective decomposes into two competing terms: the expected code length needed to describe the latent variables and the distortion between the input and reconstruction. This makes the connection between variational inference and classical R–D optimization explicit.

Ballé *et al.* [22] and Duan *et al.* [92, 93] further showed that, for single-layer and hierarchical VAEs, the variational objective can be cast as a rate–distortion (R–D) optimization problem[5] [342].

Concretely, the objective is to minimize distortion under a bitrate constraint, or equivalently, to minimize bitrate under a distortion budget. Introducing a Lagrange multiplier λ, the problem is formulated as

$$\min_{\theta,\psi} \mathcal{L}_{\mathrm{RD}} = \mathbb{E}_{x\sim p_X}[R(\hat{y}) + \lambda D(x, \hat{x})], \tag{10}$$

where $\hat{y} = \mathrm{Q}(f_\theta(x))$ and $\hat{x} = g_\theta(\hat{y})$. Here, f_θ and g_θ denote the analysis and synthesis transforms, respectively; θ are the transform parameters; Q is the quantizer; and $R(\cdot)$, parameterized by ψ, estimates the expected code length via entropy modeling; $D(\cdot,\cdot)$ is the distortion metric.

This formulation highlights a central design axis in learned image compression: the choice of the distortion function $D(\cdot,\cdot)$, which fundamentally determines how information is allocated under bitrate constraints. In the following, we categorize existing approaches into three major paradigms: objective fidelity-oriented optimization, perceptual realism-oriented optimization, and machine task-oriented optimization. These paradigms correspond to different definitions of useful information under a bitrate constraint: pixel fidelity, human perceptual realism, and machine-task utility.

2.4.1 Objective Fidelity-Oriented Optimization

Early and still dominant approaches optimize distortion in the pixel domain, aiming to faithfully reconstruct the original signal using point-wise or structure-aware similarity measures.

Mean Squared Error (MSE) remains the most widely adopted distortion function due to its simplicity, differentiability, and its direct connection to classical Gaussian rate–distortion theory. Under a Gaussian source assumption, minimizing MSE is equivalent to maximizing the log-likelihood of a Gaussian observation model:

$$\min \tfrac{1}{2\sigma^2}\|x - \tilde{x}\|_2^2 \quad \Leftrightarrow \quad \max \log \mathcal{N}(x|\tilde{x}, \sigma^2). \tag{11}$$

From a variational inference perspective, such pixel-wise losses correspond to fixed likelihood models: MSE implies an i.i.d. Gaussian assumption, while ℓ_1 loss corresponds to a Laplacian model [120, 189]. More generally, ℓ_p losses can be interpreted as assuming generalized Gaussian distributions. Although individual images may violate these assumptions, dataset-level statistical averaging makes such losses effective in practice.

To better align with human visual quality sensation, Structural Similarity (SSIM) [378] and its multiscale extension MS-SSIM [380] were introduced. Unlike pixel-wise errors, these metrics explicitly model local luminance, contrast, and structural

[5] In classical information theory, distortion usually refers to signal fidelity, such as mean squared error (MSE) or PSNR. In learned compression, the distortion term may also encode perceptual realism or task-driven criteria. We use R–D for the general Lagrangian form and explicitly use rate–distortion–perception (R–D–P) or rate–distortion–accuracy (R–D–A) when realism or machine-task objectives are central.

correlations, placing greater emphasis on structural consistency rather than absolute intensity differences.

Beyond the explicit choice of distortion metric, recent studies further show that the final R–D performance is strongly influenced by the training dynamics of learned codecs. Since transform learning, entropy modeling, and quantization are tightly coupled, stable convergence and balanced gradients become critical to coding efficiency. For example, CCA [136] improves the utilization of early autoregressive context through causal adjustment. AuxT [216] introduces lightweight auxiliary transforms to precondition uneven latent energy distributions and accelerate convergence. CMD-LIC [438] reduces instability in entropy modeling through a Sampling-then-Moving-Average strategy, while balanced R–D optimization [439] treats training as a multi-objective problem and explicitly balances gradient magnitudes.

In addition, optimizer design has also attracted increasing attention. While most LIC methods rely on first-order optimizers such as Adam [188], recent work, such as SOAP [364, 434], explores second-order curvature modeling [132, 407] to better capture the local geometry of the R–D objective and improve convergence stability. These developments suggest that coding performance is determined not only by the distortion metric itself, but also by the optimization strategy used to navigate the highly non-convex and coupled R–D landscape.

2.4.2 Perceptual Realism-Oriented Optimization

Despite their success, fidelity-driven objectives often lead to over-smoothed reconstructions, particularly at low bitrates. A growing body of evidence since 2017 has shown that minimizing pixel-wise distortion does not necessarily correlate with human perceptual quality. This discrepancy arises because pixel-domain losses penalize all deviations uniformly, whereas human perception is primarily sensitive to structures, textures, and semantics [128, 271]. As a result, modern research has shifted from *objective fidelity* toward *perceptual realism*, where the goal is not only to reproduce the input signal exactly, but also to generate reconstructions that are visually plausible to human observers.

Theoretical foundations: rate–distortion–perception trade-off. The empirical success of perceptual optimization was formalized by Blau and Michaeli [31, 32], who established a fundamental trade-off among rate, distortion, and perception. They showed that, for a fixed bitrate, improving perceptual quality—defined as reducing the divergence between the distribution of reconstructed images and that of natural images—necessarily increases distortion measured in a signal fidelity sense, and vice versa. Here, perceptual realism is quantified as the *distributional proximity* between $p_{\hat{x}}$ and p_x. This result provides a unifying interpretation of perceptual compression methods: improving realism corresponds to allocating limited bitrate toward modeling the natural image distribution rather than preserving exact pixel values.

Subsequent works extended this framework to more general settings, including perceptual metrics such as learned perceptual image patch similarity (LPIPS) for empirical analysis [190], conditional coding with side information [133, 283], and randomness [134, 284], as well as Gaussian sources [317]. These extensions further

reveal how additional priors, side information, or randomness can effectively relax the trade-off boundary.

In addition, several recent works explore perceptual compression from alternative theoretical perspectives. Xu *et al.* [389] proved that conditional generative codecs satisfy an idempotence property, and further showed that an unconditional generative model constrained by idempotence is theoretically equivalent to a conditional generative codec, enabling the conversion of MSE-optimized codecs into perceptual ones. SIC [235] introduces Synonymous Variational Inference (SVI), which reformulates perceptual compression as variational inference over semantic equivalence classes (synsets), providing a semantic-level interpretation of perceptual reconstruction.

Guided by these theoretical insights, practical perceptual optimization requires *quantifiable surrogates* that can approximate human perceptual judgments during training. Existing approaches mainly instantiate this idea from two complementary perspectives: *distribution-level alignment*, which encourages reconstructions to match the natural image distribution, and *representation-level similarity*, which compares reconstructed and original images in perceptually meaningful feature spaces.

Adversarial learning. Early perceptual compression methods relied on proxy models to approximate perceptual importance. Representative approaches, including SPIC [300], CDPIC [91], OBIC [385], and SDPIC [294], employ auxiliary semantic or saliency networks to guide spatial bit allocation toward perceptually important regions. However, these methods largely depend on handcrafted or task-specific perceptual proxies and lack a unified formulation of perceptual quality.

A major breakthrough came with generative adversarial networks (GANs), which implicitly characterize perceptual quality through distribution matching. Rather than minimizing point-wise distortion, GAN-based methods introduce an adversarial objective that encourages reconstructed images to match the distribution of natural images, thereby promoting outputs that lie on the manifold of realistic visual signals [121, 254, 308].

Building upon this idea, Liu *et al.* [240], Agustsson *et al.* [3] and Huang *et al.* [152] first incorporated adversarial learning into learned image compression, demonstrating that visually plausible textures can be synthesized at extremely low bitrates (e.g., below 0.1 bits per pixel (bpp)). More recently, [4] and [159] proposed conditional GAN-based frameworks capable of producing multiple reconstruction modes from a single compressed representation, enabling flexible trade-offs between perceptual realism and distortion fidelity at the decoder side.

From an optimization perspective, GAN training remains challenging due to the instability of min–max optimization and the sensitivity to discriminator design. Consequently, substantial effort has been devoted to improving training stability and perceptual quality through enhanced discriminator architectures [69, 295], improved adversarial objectives [296], and refined training strategies [369, 375]. Despite these challenges, adversarial learning remains one of the most effective approaches for enforcing distribution-level realism in perceptual compression.

Deep feature-based perceptual learning. A more principled and stable direction arises from the observation that deep features extracted from pretrained vision networks correlate well with human perceptual judgments [53, 85, 293]. Early work

by Liu *et al.* [240] first introduced perceptual losses based on Visual Geometry Group (VGG) features into learned image compression and combined them with adversarial training, demonstrating significant perceptual quality improvements over purely distortion-driven optimization.

This idea was formalized by Zhang *et al.* [441], who proposed LPIPS, a perceptual similarity metric defined as distances in deep feature spaces (e.g., VGG [331] or AlexNet [198]). Compared with pixel-wise distortion, LPIPS performs *representation-level comparison*, enabling better modeling of high-level structures and semantic consistency. From a modeling perspective, LPIPS can be interpreted as an ℓ_2 distance in a learned perceptual feature space, analogous to MSE in the pixel domain but defined over semantically meaningful representations.

Feature-based perceptual losses have since become standard components in perceptual compression. HiFiC [273] systematically integrates LPIPS-style losses with adversarial objectives to achieve controllable trade-offs between fidelity and realism, while subsequent works such as PO-ELIC [140] further refine this paradigm through improved optimization and training strategies.

Beyond LPIPS, alternative metrics such as Deep Image Structure and Texture Similarity (DISTS) [84] decompose perceptual similarity into structural and texture components, offering improved robustness. Some work [127, 329, 375] gradually incorporates DISTS into optimization.

Alternative methods. Recent works incorporate richer priors, including vision foundation models [285, 381], text-guided compression [209, 306], and perceptual thresholds such as just noticeable difference (JND) [288, 353]. These approaches highlight a growing trend toward leveraging high-level semantics and human perceptual characteristics to guide compression, especially in the low-bitrate regime.

Diffusion-based generative models have also recently emerged as a promising direction for perceptual compression, owing to their strong capability in modeling complex data distributions [62, 127, 180, 199, 224, 232, 256, 291, 310, 324, 329, 390, 394, 399, 429]. However, existing approaches are still at an early stage, with open questions regarding rate range, coding efficiency, and integration with practical entropy coding frameworks. We therefore treat them as an emerging frontier rather than a mature codec family in the present survey.

2.4.3 Machine Task-Oriented Optimization

Beyond human perception, images are increasingly consumed by machine vision systems in applications such as autonomous driving, robotics, and embodied AI. In these scenarios, the ultimate objective is not visual quality, but *task performance*, giving rise to a new paradigm of rate–distortion–accuracy (R–D–A) optimization [47, 263].

Reconstruction-then-understanding. A straightforward approach follows the conventional pipeline: reconstruct the image and then perform downstream tasks. In this setting, the key challenge is to preserve task-relevant semantic information during compression.

Early works incorporate task loss (e.g., classification or detection loss) into the R–D objective [46, 263], enabling joint optimization for both reconstruction and task accuracy. Interestingly, several studies [104, 293, 406] observe that perceptual-optimized

Table 3 Overview of representative learned image compression methods.

Method	Pub.	Core Components			Params (M) ↓	KMACs/ pixel ↓	BD-rate (%) ↓
		Arch.	Context	Dist.			
Factorized [19]	ICLR'17	CNN	Factorized	UM	7.03	204.00	67.80
Hyperprior [21]	ICLR'18	CNN	Hyperprior	GM	11.81	208.97	30.14
JointAR [276]	NeurIPS'18	CNN	SAR	GM	25.50	224.80	10.61
ChannelAR [275]	ICIP'20	CNN	CAR	GM	28.72	243.00	7.22
NLAIC [57]	TIP'20	Attention	SCAR (3D)	GM	65.50	823.54	5.92
GMM [68]	CVPR'20	CNN	SAR	GMM	29.63	512.80	4.55
ELIC [139]	CVPR'22	CNN	SCAR	GM	36.93	573.88	-3.22
TIC [258]	DCC'22	Attention	SCAR	GM	28.40	506.72	-4.58
TCM [245]	CVPR'23	Attention	CAR	GM	76.57	1823.58	-10.70
QARV [92]	TPAMI'24	CNN	HAR	GM	93.40	718.96	-5.81
FLIC [215]	ICLR'24	Attention	CAR	GM	70.97	1096.04	-12.97
MLIC++ [175]	TOMM'25	CNN	SCAR	GM	116.72	1282.81	-11.83
MambaIC [416]	CVPR'25	Mamba	SCAR	GM	75.78	1284.86	-15.72
HPCM [228]	ICCV'25	CNN	HAR	GM	89.71	1261.29	-19.19
DHIC [78]	ICLR'26	CNN	HAR	GM	106.93	977.73	**-19.73**

Note: **Pub.**: publication venue and year; **Arch.**: main architecture; **Dist.**: distribution model. Because BD-rate and complexity values depend on anchor version, color pipeline, test-set split, interpolation protocol, and whether results are reproduced or reported by the original papers, the table should be read as a normalized high-level comparison rather than a bit-exact cross-paper leaderboard.

compression often yields better task performance, suggesting that human perception and machine vision share common sensitivity to structural and semantic features. However, task-oriented compression reveals a critical insight: *not all information is equally important for downstream tasks*. This motivates scalable or layered coding schemes, where a base layer preserves task-relevant semantics, and enhancement layers progressively refine visual quality [149, 401]. Such designs enable bitrate-efficient transmission tailored to machine-centric applications.

A limitation of this paradigm is that pretrained vision models are typically trained on clean, uncompressed images, leading to domain mismatch when applied to compressed reconstructions. To address this, recent works [47, 207, 249, 371] explore joint optimization of compression and task networks. For example, ICMH-Net [249] explicitly models how to aggregate latent representations to balance human and machine objectives, while recent multi-path aggregation [432] extends this idea to multi-task scenarios.

Understanding in latent space. An alternative paradigm bypasses reconstruction entirely and performs inference directly in the compressed latent domain. This is motivated by the observation that high-level vision tasks operate on abstract representations rather than raw pixels.

This direction can be traced back to early studies on collaborative intelligence [70, 358], where neural networks are partitioned between edge devices and the cloud to reduce communication overhead. Around the same period, CodedVision [322] explored a unified framework that jointly optimizes image compression and understanding, and further extended the paradigm to multi-task learning. These early studies demonstrate that compressed latent representations can directly support downstream tasks, suggesting that a single learned representation may simultaneously serve both compression and machine vision objectives. Subsequent works [71, 332] further refine this paradigm through improved latent representations and task-aware optimization strategies.

From an information-theoretic perspective, this paradigm reflects a hierarchy of semantic abstraction: as tasks progress from reconstruction to segmentation, detection, and classification, the required information content gradually decreases. Higher-level tasks rely less on fine-grained signal details and more on abstract semantic cues. Therefore, the key to achieving high performance lies in selecting task-relevant, non-redundant information from a unified latent representation.

Recent works explore various mechanisms to achieve this, including scalable latent representations [72, 246], gating mechanisms for adaptive feature selection [244], and self-supervised representation learning [101]. More recently, approaches based on pre-trained codecs leverage lightweight adapters or prompt-based modulation [63, 94] to enable task-specific inference without modifying the underlying codec.

2.4.4 Discussion and Perspective

The development of perceptual and task-oriented compression reflects a broader shift from *signal reconstruction* to *information utilization*. In Shannon and Weaver's classical view, communication can be considered at three levels: syntactic accuracy, semantic meaning, and pragmatic effectiveness [319]. Traditional source coding has primarily operated at the syntactic level, aiming to minimize bitrate while preserving reconstruction fidelity. In contrast, perceptual compression moves toward the semantic level by prioritizing visually meaningful structures over exact pixel values, while machine vision-oriented compression further approaches the pragmatic level by optimizing for downstream task performance.

This trajectory is closely aligned with the emerging paradigm of semantic communication [25, 154, 320], where the goal is not to transmit all signal details, but to preserve information relevant to perception, interpretation, or decision-making. From a unified optimization perspective, signal fidelity, perceptual realism, and task accuracy can be viewed as different forms of distortion measures. A central open challenge is therefore to develop compact representations that can flexibly support reconstruction, perception, and inference within a unified coding framework.

2.5 Benchmark and Evaluation

Over the past decade, learned image compression has evolved from proof-of-concept models to highly optimized end-to-end systems. This rapid progress has made benchmarking increasingly challenging. Unlike traditional codecs, which are typically evaluated under standardized common test conditions (CTC), learned codecs are strongly coupled with training data, optimization objectives, architectural design, and implementation details.

To provide a clear and controlled comparison, this survey adopts a fidelity-oriented evaluation protocol, focusing on PSNR-based rate–distortion (R–D) performance and computational complexity. Although perceptual realism and machine task-oriented compression have received increasing attention, their evaluation often depends on task-specific objectives, datasets, and perceptual metrics, making direct cross-method comparison difficult. By contrast, MSE-optimized codecs provide a relatively

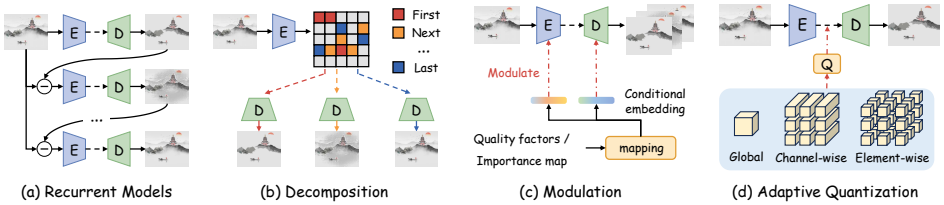


Fig. 6 Representative paradigms for variable-rate coding in learned image compression. Existing approaches introduce rate adaptability through recurrent refinement, latent decomposition or progressive transmission, feature modulation, and adaptive quantization.

standardized and reproducible basis for benchmarking. Many architectural and modular advances developed under objective fidelity-oriented settings can also benefit perceptual-realism and task-oriented codecs.

2.5.1 Evaluation

Table 3 summarizes representative learned image compression methods using a unified set of metrics that characterize both compression efficiency and computational cost.

Distortion metric. We use PSNR as the primary fidelity metric and report BD-rate [30] against the VTM Intra [35] anchor under the H.266/VVC image-profile configuration to summarize R–D performance.

Complexity metric. We measure computational cost using KMACs/pixel, which provides a hardware-agnostic proxy for model complexity. We also report parameter counts to reflect model size and deployment overhead. Together, these metrics provide a consistent view of efficiency across heterogeneous codec designs. Nevertheless, KMACs/pixel does not directly determine wall-clock latency, memory bandwidth, entropy-coder runtime, or energy consumption, so it should be interpreted as an architecture-level complexity proxy.

2.5.2 Discussion and Perspective

This benchmarking protocol highlights the central trade-off between compression efficiency and computational complexity. As shown in Table 3, recent improvements in R–D performance are often accompanied by substantially increased computational cost, underscoring the importance of system efficiency-aware codec design.

As the field moves from performance-driven exploration toward practical deployment, lightweight, robust, and reproducible benchmark frameworks become increasingly important. This need motivates the development of open and efficient reference implementations, which we introduce in the next section.

To make this transition explicit, Table 4 summarizes the evidence base and deployment readiness of major LIC research directions. The matrix is a qualitative author assessment based on evidence density, protocol maturity, and deployment validation. The readiness levels are qualitative: *High* indicates mature benchmarks and clear practical entry points; *Medium* indicates active progress but incomplete protocol or system validation; and *Low* indicates that evidence remains fragmented or strongly application-dependent.

3 Practical Extensions

Toward real-world deployment, both academia and industry have devoted substantial efforts to improving flexibility, controllability, and system efficiency. In this section, we review practical extensions that enable learned codecs under operational constraints encountered in communication, storage, and edge-intelligence pipelines.

3.1 Variable-Rate Coding and Rate Control

Variable-rate coding (VRC) and rate control are essential for practical image codecs, as they enable a single model to adapt to varying bandwidth constraints and user preferences without retraining or maintaining multiple models.

3.1.1 Variable-rate coding

Variable-rate coding aims to support multiple rate–distortion (R–D) operating points within a unified framework. Unlike traditional codecs, where bitrate is primarily controlled through explicit quantization parameter (QP) adjustment, learned image compression requires dedicated mechanisms to regulate bitrate while maintaining stable reconstruction quality. Existing approaches mainly differ in how and where rate adaptability is introduced into the codec pipeline.

Recurrent Models. Early variable-rate learned codecs were primarily based on recurrent architectures. Toderici *et al.* [354, 355] proposed recurrent neural networks (RNNs) for LIC, where each iteration progressively encodes additional residual information to refine reconstruction quality. In this framework, bitrate is directly controlled by the number of recurrent iterations, naturally enabling variable-rate coding. Subsequent works [157, 165, 177, 340] further improved network architecture and training stability.

However, these approaches operate in the pixel-domain residual space and require sequential inference, resulting in increased latency and computational cost at higher bitrates. To alleviate this issue, Yang *et al.* [397] proposed slimmable compressive autoencoders (SlimCAEs), where subnetworks with different capacities correspond to different rate levels. Although this design reduces iterative decoding overhead, model complexity still scales with the target bitrate, motivating the development of unified architectures with approximately constant complexity across different rate points.

Feature Decomposition. Another line of work introduces rate scalability through structured decomposition of latent representations within a single forward pass. Rippel *et al.* [312] proposed bitplane decomposition of latent features, where different bitplanes progressively encode finer image details. Similarly, Nakanishi *et al.* [281] and Cai *et al.* [37] employed multi-scale transforms to generate hierarchical latent representations.

These methods naturally support progressive transmission and multi-rate decoding without repeated encoding. However, different decomposition levels are often strongly coupled, making joint optimization across all rate points difficult. In addition, because rate adaptation is performed through discrete bitplanes or hierarchical scales, the achievable bitrates are inherently coarse, leading to sparse coverage of the R–D curve.

Feature Modulation. A more flexible strategy introduces continuous modulation mechanisms into latent representations or transform modules. Chen *et al.* [54] observed that latent representations corresponding to different bitrates exhibit strong similarity, and that bitrate variation can be effectively controlled through latent scaling. Building upon this insight, subsequent works incorporated modulation mechanisms into transform layers, normalization parameters, or latent activations [22, 39, 40, 73, 80, 111, 129, 237, 347, 396]. Typically, the Lagrangian multiplier λ, which controls the R–D trade-off, is mapped into a learned embedding that modulates the codec behavior (e.g., convolutional kernels, normalization parameters, or latent activations).

Compared with recurrent or decomposition-based approaches, feature modulation enables approximately continuous rate control within a single model while maintaining nearly constant complexity across different operating points. Furthermore, several works introduced spatially adaptive quality maps [40, 179, 233, 268, 335], enabling fine-grained bit allocation across image regions. This mechanism is particularly promising for region-of-interest (ROI) coding [38], although determining optimal quality maps or ROI regions remains an open problem.

Adaptive Quantization. Inspired by traditional codecs, adaptive quantization remains one of the most fundamental mechanisms for rate control in learned compression (see Fig. 6d). The core idea is to scale latent representations before quantization, thereby adjusting the effective quantization step size and controlling bitrate.

Depending on the granularity of scaling, existing methods can be broadly divided into: (i) *global scaling*, where a single scaling factor is applied to all latent variables [96, 223, 301, 357, 446], and (ii) *latent-wise scaling*, where different latent channels or elements are assigned different quantization factors [6, 7, 89, 111, 352]. Compared with global scaling, latent-wise strategies provide finer bitrate control and better adaptation to heterogeneous latent statistics.

3.1.2 Progressive Coding

Progressive coding can be regarded as a constrained form of variable-rate coding. While general variable-rate codecs aim to provide multiple R–D operating points within a single model, progressive codecs further require these operating points to be nested and incrementally decodable. In other words, a lower-rate reconstruction should be obtained from a prefix or ordered subset of the same bitstream, and additional symbols should progressively refine the reconstruction quality. This requirement imposes an explicit ordering constraint on the latent or transmitted symbols.

Early recurrent codecs naturally supported progressive reconstruction through iterative residual refinement [15], but suffered from high computational overhead due to repeated inference. More recent approaches instead operate on a shared latent representation and achieve progressiveness through structured decomposition or masking strategies. In these methods, latent representations are partitioned into ordered subsets according to importance, allowing partial decoding using only a subset of transmitted symbols.

Representative strategies include trit-plane coding [163, 213], nested quantization grids [212, 261], channel variance-based masking [135], learnable importance

maps [211], and self-supervised feature priors such as DINO [45] for importance estimation [16]. Another related paradigm is *resolution-scalable compression* [418], where a low-resolution image is reconstructed first and spatial details are progressively refined as more bits become available. More recently, hierarchical autoregressive modeling has also been explored for progressive transmission [78, 419, 443], where different autoregressive stages correspond to progressively refined reconstructions.

ProgDTD [144] further observed that latent bottlenecks naturally exhibit an intrinsic importance ordering. By modifying the training objective, the codec learns to organize latent elements according to their contribution to reconstruction quality, enabling progressive transmission without additional architectural overhead. Building on this perspective, Presta *et al.* [302] introduced element-wise importance ranking and masking strategies, achieving state-of-the-art progressive coding performance while remaining compatible with standard compression pipelines. LossResilientLIC [318] further incorporates network-aware progressive coding to improve robustness under packet loss, extending progressive transmission from rate scalability toward resilient communication.

3.1.3 Rate Control

While variable-rate coding provides multiple operating points, rate control (RC) focuses on selecting appropriate coding parameters so that the actual bitrate closely matches a target rate with minimal distortion. Formally, rate control can be formulated as a constrained optimization problem:

$$\min_{\{\alpha_i\}} D, \quad \text{s.t.} \quad R \leq R_{\texttt{target}}, \tag{12}$$

where $\{\alpha_i\}$ denotes tunable coding parameters, such as quantization scales or modulation factors, and R, D, and $R_{\texttt{target}}$ denote the actual bitrate, distortion, and target bitrate, respectively. Existing rate control methods for LICs can be broadly grouped into three categories.

Repeated encoding search. A direct strategy is to repeatedly encode the image with different parameter settings until the target bitrate is reached. For example, JPEG AI [12] adopts a coarse-to-fine strategy, where a candidate model is first selected according to its nominal bitrate range, followed by fine adjustment of latent modulation parameters. Jia *et al.* [166] further propose a two-stage rate control scheme: a model closest to the target bitrate is first selected, and then a bitrate–parameter curve is estimated by evaluating multiple parameter settings around the target. An iterative search is finally performed to identify the configuration that minimizes distortion under the bitrate constraint. Although search-based repeated encoding can achieve accurate rate control, it requires multiple encoding passes at inference time, resulting in considerable computational overhead and latency.

Bitrate–parameter mapping. To reduce repeated encoding, another line of work explicitly models the relationship between target bitrate and coding parameters. Shi *et al.* [330] investigate the R–λ relationship in learned compression. Zhang *et al.* [428] propose a Rate-Feature-Level (RFL) model that uses both the target bitrate and image content features to predict suitable quantization levels. Xue *et al.* [393]

introduce an auxiliary rate estimator that leverages shallow, high-resolution encoder features to predict the mapping from target bitrate to λ. Similarly, Pan *et al.* [286] formulate rate control as a classification problem over discrete model and parameter indices, using neural predictors to select the appropriate configuration for a given image and target rate. These methods substantially reduce encoding overhead and are more suitable for real-time applications, but their accuracy depends on how well the predictor generalizes to unseen content and target rates.

Block-level rate allocation. Beyond global rate control, spatially fine-grained allocation aims to exploit content heterogeneity within an image. Wang *et al.* [376] divide an image into blocks and independently model the bitrate–parameter relationship of each block through repeated encoding. A greedy optimization algorithm is then used to allocate bits across blocks under a global bitrate constraint. To reduce the computational burden, Dong *et al.* [87] exploit statistical correlations among neighboring blocks and predict bitrate–parameter curves for most blocks using only a small subset of sampled blocks. This strategy greatly reduces the number of required re-encodings while maintaining high rate-control accuracy. Such methods provide more precise spatial bit allocation, especially for images with strong content non-uniformity.

3.1.4 Discussion and Perspective

Variable-rate capability has become a de facto requirement for practical LIC. Empirical evidence suggests that simple modulation-based VRC mechanisms can provide rate flexibility without noticeably degrading rate–distortion performance compared with fixed-rate models [92, 169], while requiring limited complexity overhead. Conceptually, feature modulation in compression is also related to conditional embedding mechanisms widely used in generative models and large language models, where external control variables modulate intermediate representations. This connection may inspire more expressive conditioning frameworks for flexibility of LIC.

Compared with still-image compression, learned video rate control is more constrained by temporal dependency, buffer dynamics, and stricter bandwidth requirements [126, 226, 417, 435]. The bit budget assigned to the current frame affects both its own R–D behavior and the prediction quality of later frames; this sequential dependency motivates joint λ–R/λ–D modeling and, in some settings, reinforcement-learning-based allocation [79, 108, 126, 176, 309, 315, 412].

Although rate control for still images is less constrained by temporal dynamics, emerging scenarios such as ultra-high-resolution imaging, immersive 360° content, and bandwidth-limited edge deployment are increasing the demand for accurate and efficient bitrate control. Future research should focus on robust bitrate–parameter prediction, spatially adaptive and perceptually driven bit allocation, and tighter integration between rate control, semantic importance, and downstream task objectives.

3.2 Model Quantization & Cross-Platform Consistency

Despite significant progress in rate–distortion performance, LIC models are typically implemented using floating-point arithmetic, which introduces non-determinism across different hardware platforms and software environments. Even minor discrepancies in

floating-point precision or rounding behavior can lead to entropy decoding mismatch or reconstruction drift, ultimately breaking bitstream compatibility. Such issues are particularly critical in compression systems, where strict cross-platform consistency and deterministic decoding are fundamental requirements.

To address these challenges, model quantization has emerged as a key technique for enabling efficient deployment and standardized interoperability in LIC. Early work by Ballé et al. [20] pioneered integer-only learned compression by redesigning the entire codec in the integer domain, thereby eliminating floating-point inconsistencies at the source. This work achieved fully deterministic cross-platform decoding and laid the foundation for subsequent research on the model quantization of LICs.

3.2.1 Post-Training Quantization (PTQ) versus Quantization-Aware Training (QAT)

From a methodological perspective, quantization strategies can be broadly categorized into quantization-aware training (QAT) and post-training quantization (PTQ).

Quantization-Aware Training (QAT) simulates quantization effects during training, allowing the model to adapt to low-precision representations. Early works apply uniform quantization to convolutional weights [343, 345], and later extend to both weights and activations [344]. However, directly applying those quantization schemes developed for computer vision models to LIC often leads to noticeable performance degradation. This is mainly because LIC latents are not merely intermediate features, but are eventually quantized into discrete integer symbols for entropy coding. To preserve rate–distortion flexibility, these latents often span a much larger numerical range than standard neural activations, making low-bit quantization of LIC models more challenging.

To address this issue, range-adaptive quantization (RAQ) [145] introduces layer-wise dynamic range calibration and bit-width adaptation, jointly optimizing weights, biases, and activations. QLIC [346] further mitigates quantization error by redistributing channels with large quantization distortion into multiple sub-channels, effectively smoothing the dynamic range. IQ-LIC [162] incorporates layer-wise quantization loss as an additional supervision signal to guide training. Beyond standalone quantization, QAT has also been combined with model pruning [146] and variable-rate frameworks [408], enabling joint optimization of compression efficiency, model size, and bitrate flexibility.

While QAT generally preserves compression performance and yields deployable integer models, it requires full retraining with access to training data, incurs substantial computational cost, and reduces flexibility once the quantized model is fixed.

Post-Training Quantization (PTQ) converts pretrained floating-point models into low-precision representations without retraining, offering a more practical and flexible deployment pipeline. Although originally developed for classification and detection tasks, PTQ has recently been adapted to LIC models.

He *et al.* [138] combine classical PTQ techniques [225, 280] with deterministic entropy coding to achieve cross-platform consistent decoding. Rate–distortion-optimized PTQ (RDO-PTQ) [325] further analyze the relationship between quantization error and rate–distortion performance, showing that minimizing quantization error alone does not necessarily yield optimal compression performance, and propose incorporating task-aware loss functions into PTQ optimization.

A key insight is provided by Koyuncu et al. [196], who identify that cross-platform inconsistency primarily arises from discrepancies in probability estimation of the entropy model between encoder and decoder. Based on this observation, they demonstrate that quantizing the entropy model alone is sufficient to ensure consistent decoding. This finding is further validated in their subsequent work within the JPEG AI framework [12, 197].

3.2.2 Mixed-Precision Quantization

From the perspective of precision granularity, quantization strategies can be divided into uniform quantization and mixed-precision quantization (MPQ). Uniform quantization assigns a fixed bit-width (e.g., INT8) to all layers, enabling efficient hardware acceleration and straightforward implementation. However, it does not account for the heterogeneous sensitivity of different layers or channels to quantization noise. Mixed-precision quantization addresses this limitation by assigning different bit-widths across modules, layers, or channels based on their impact on rate–distortion performance. The central challenge lies in designing effective criteria for bit-width allocation.

Existing approaches explore various strategies. For instance, FMPQ [147] and MP-PTQ [411] use the fractional change in rate–distortion loss as a sensitivity metric. Other methods, such as [398] and DynaQuant [27], introduce routing mechanisms that dynamically select bit-widths from a predefined set of candidates.

Another emerging direction considers transforming the feature distribution prior to quantization. Since latent representations in LIC often exhibit long-tailed and highly non-uniform distributions, performing quantization in a transformed domain can significantly reduce information loss. For example, AWPB [410] applies adaptive warping to better align the distribution with quantization bins.

3.2.3 Cross-Platform Consistency

Beyond quantization, additional mechanisms have been proposed to ensure cross-platform consistency. Pang et al. [289] introduce a safeguarding mechanism that transmits auxiliary guard bits to tolerate numerical deviations during decoding. However, this approach may significantly increase bitstream size, especially at low bitrates.

As discussed earlier, inconsistency largely stems from entropy model mismatch. Motivated by this, Yang et al. [400] propose removing the entropy model altogether and adopting chain coding, achieving competitive performance compared with BPG (the H.265/HEVC image-profile anchor), while inherently avoiding probability mismatch issues.

3.2.4 Discussion and Perspective

Model quantization is essential for deploying learned codecs on resource-constrained devices, where latency, memory footprint, and power consumption are tightly limited. However, practical deployment remains constrained by hardware support. Most existing accelerators are optimized for uniform precision formats such as 16-bit floating point (FP16) and INT8, whereas fine-grained mixed-bit operations, such as channel-wise 4-bit or 6-bit integer (INT4/INT6) operations, are not yet efficiently supported. This gap limits the practical benefits of mixed-precision quantization despite its potential rate–distortion and complexity advantages.

A key direction is therefore to design hardware-friendly quantization schemes that balance compression performance and deployability. For example, layer-wise activation quantization combined with channel-wise weight quantization under standardized bit-width formats, such as INT8, may offer an effective compromise between flexibility and hardware efficiency [49]. In parallel, emerging low-precision floating-point formats, including 8-bit floating point (FP8) and 4-bit floating point (FP4), provide promising alternatives to integer quantization [202, 361, 370].

Another important challenge lies in cross-platform consistency, especially for learned video compression. In still image compression, quantizing the entropy model is often sufficient to ensure deterministic decoding for the tested architectures and entropy-decoding paths. In video compression, however, decoded features are reused as references for subsequent frames [77, 217, 242, 252, 255], so small numerical discrepancies may accumulate and propagate temporally. Fully quantized and deterministic decoder implementations are therefore crucial for reliable deployment. Recent works have begun to address this issue [76, 362], but robust hardware-consistent quantized video codecs remain an open problem.

3.3 Model Robustness

Beyond rate–distortion performance, robustness is a critical criterion for evaluating the real-world viability of LIC systems. Unlike traditional codecs with manually designed and largely deterministic modules, learned codecs rely on highly nonlinear transforms and data-driven entropy models. As a result, small perturbations in the input or latent space may lead to severe reconstruction artifacts, inaccurate entropy estimation, or substantial bitrate inflation [55]. In this survey, robustness covers several threat and stress models: benign distribution shift, adversarial evasion, training-time backdoors, repeated compression, packet loss or bitstream corruption, and error propagation into downstream machine-vision tasks.

Robustness is important for two reasons [36]. First, learned codecs may exhibit unstable behavior even under benign inputs, producing atypical artifacts that differ from those of classical codecs, as also observed in recent evaluations of JPEG AI [12]. Second, compression is increasingly used as a front-end module in downstream vision pipelines. Perturbations introduced or amplified during compression may propagate to recognition, detection, or other machine vision models, thereby compromising the reliability of the entire system.

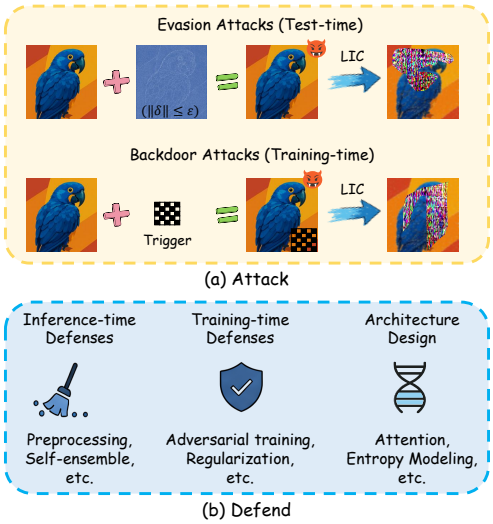


Fig. 7 Illustration of attacks and defenses for LIC models. Attacks include test-time evasion perturbations and training-time backdoor triggers, while defenses can be introduced through architecture design, robust training, and inference-time protection.

Robustness evaluation in LIC differs from that in standard vision tasks. Perturbations may affect both distortion and bitrate, and the entropy model itself becomes an attack surface. Moreover, compression has a dual role: it can remove adversarial noise and serve as a preprocessing defense [168, 282], but it may also amplify perturbations or introduce artifacts that degrade reconstruction and downstream task performance.

3.3.1 Attacks

Given an input image x, an adversary seeks a perturbation δ such that the adversarial input $x^* = x + \delta$ degrades codec behavior under a bounded perturbation budget. The attack objective may target reconstruction quality, coding cost, or downstream task performance. Accordingly, attacks on LIC can be broadly grouped into test-time evasion attacks and training-time backdoor attacks.

Evasion attacks. Evasion attacks construct adversarial inputs at inference time. In the white-box setting, gradients through the codec can be used to optimize perturbations against R–D objectives. For example, Chen *et al.* [55] propose a fast threshold-constrained distortion attack (FTDA) that increases reconstruction distortion while maintaining imperceptibility. Subsequent works improve attack effectiveness by designing compression-specific objectives [201, 269], exploiting transform-domain representations such as wavelet-domain attacks [178], or directly targeting bitrate inflation [383].

In the black-box setting, transferability-based attacks [250, 290] are commonly adopted, where perturbations generated on surrogate codecs are transferred to unseen models. Recent studies further extend adversarial objectives from human-oriented reconstruction to machine vision tasks [341], showing that attacks at the compression stage can propagate through the pipeline and degrade downstream predictions.

Backdoor attacks. Backdoor attacks poison the training process by implanting hidden triggers, causing the codec to behave normally on clean inputs but abnormally when a trigger is present. Compared with evasion attacks, backdoors are more stealthy and difficult to detect, posing substantial risks when learned codecs are trained or fine-tuned on untrusted data.

For example, Yu *et al.* [413, 414] introduce frequency-domain trigger injection based on discrete cosine transform (DCT) representations, where structured triggers are embedded into a subset of training samples. Xue *et al.* [392] further address the vulnerability of backdoor triggers to compression by enforcing feature consistency, allowing trigger patterns to survive the compression process.

3.3.2 Defenses

Defenses for LIC can be categorized according to where robustness is introduced: architecture-level design, training-time robustification, and inference-time protection.

Architecture-level design. One direction is to improve robustness through codec design, especially in entropy modeling. Since learned entropy models are sensitive to distribution mismatch, simplifying or stabilizing their structure can reduce vulnerability to perturbations. Liu *et al.* [247, 248] replace complex autoregressive components with simpler independent distribution models, and suggest that attention mechanisms may improve robustness. Such designs reveal a trade-off between R–D efficiency and stability: stronger entropy models improve compression performance but may also increase sensitivity to adversarial or out-of-distribution inputs.

Training-time robustification. Training-based defenses expose the codec to perturbations during optimization. Representative strategies include adversarial training [55, 270, 449] and consistency regularization. For instance, Kim *et al.* [185] introduce latent consistency losses to mitigate error accumulation under repeated compression, while Cao *et al.* [42] combine knowledge distillation with adversarial training. These methods can improve robustness, but typically increase training cost and may degrade clean-image R–D performance.

Inference-time protection. Inference-time defenses aim to improve robustness without retraining the codec. Common strategies include input preprocessing, geometric self-ensemble, randomized smoothing, and anomaly detection based on latent or bitstream statistics [55, 336]. Some methods also modify numerical precision or latent representations at inference time [269]. These approaches are flexible and easy to deploy, but their protection is often partial and may be bypassed by adaptive attacks.

3.3.3 Practical Robustness Considerations

Beyond adversarial attacks, learned codecs must also be robust to non-adversarial perturbations encountered in practical deployment.

Repeated compression. Repeated encoding and decoding can accumulate errors in the latent space, leading to progressive reconstruction degradation [185]. This effect can be more pronounced at high bitrates, where finer representation makes the codec more sensitive to small perturbations [55].

Transmission errors. In practical transmission scenarios, packet loss or bitstream corruption may cause partial latent omission and severe reconstruction degradation. Traditional systems typically improve robustness through channel-side forward error correction (FEC) [278], whereas learned codecs can enable joint source–channel optimization by incorporating robustness directly into latent representations and reconstruction objectives [318, 373, 421]. Such designs can achieve improved resilience to unreliable channels under the same bandwidth budget. Incorporating transmission perturbations during training further enhances robustness to dynamic network conditions.

Interaction with downstream tasks. Compression interacts nontrivially with downstream machine vision systems. On one hand, it can suppress adversarial noise and improve robustness as a preprocessing operation [168]. On the other hand, compression artifacts or quantization noise can be exploited to attack downstream models [251]. Recent evidence also suggests that high-fidelity compression can improve robustness in certain task pipelines [307].

3.3.4 Discussion and Perspective

Robustness in learned image compression should be understood as a system-level property involving adversarial security, numerical stability, transmission reliability, and downstream task robustness. Recent studies reveal strong transferability of adversarial perturbations across different codecs and architectures [36], suggesting that vulnerabilities may arise from shared properties of learned representations and entropy modeling rather than from isolated architectural choices.

Emerging generative and perceptual compression paradigms further complicate robustness analysis. Compared with deterministic distortion-oriented codecs, these methods may involve stochastic decoding, distribution sampling, and perceptual objectives, whose robustness behavior is not yet well understood [204, 405, 430]. Existing R–D-based attack and defense strategies may therefore not directly generalize to such settings.

Looking forward, robust learned compression requires unified evaluation protocols that jointly consider bitrate, reconstruction quality, adversarial vulnerability, and downstream task performance. Developing codec architectures and training objectives that balance compression efficiency, stability, and task reliability remains an important open direction.

3.4 Lightweight Deployment

Under PSNR/MS-SSIM evaluation on common still-image benchmarks, several LIC models now match or exceed conventional anchors such as JPEG [365] and VTM Intra, the H.266/VVC image-profile anchor [298]. Nevertheless, classical standards still dominate real-world deployment due to their mature status in practice. As a fundamental system component, image compression must operate under strict constraints on latency, memory, and power consumption. This motivates the development of *lightweight LIC* models that balance compression performance with practical efficiency.

Recent research has increasingly explored several directions, including knowledge distillation, model compression (e.g., pruning and quantization), architecture-level design, and test-time adaptation.

3.4.1 Knowledge Distillation

Knowledge distillation (KD) transfers knowledge from a high-capacity teacher model to a lightweight student model [122]. In learned image compression, KD differs from its use in classification: the objective is not only to match output logits, but also to transfer structured representations, reconstruction behavior, and entropy-modeling capability.

Early approaches mainly rely on output-level supervision, such as reconstruction consistency or entropy-related statistics [106]. More recent methods perform finer-grained distillation in the latent or intermediate feature space [10]. KDIC [59] introduces multi-level distillation across network blocks, improving training stability and student performance. EVC [367] adopts a mask-decay mechanism that gradually transforms a large model into a smaller one, implicitly transferring knowledge without explicit distillation losses. CSLIC [374] further combines neural architecture search (NAS) with KD to construct encoder–decoder configurations at multiple complexity levels, enabling scalable complexity–compression trade-offs.

3.4.2 Quantization and Pruning

Model compression techniques, including quantization and pruning, are widely used to reduce model size and computational cost. Quantization converts floating-point parameters and activations into low-bit representations, enabling efficient integer inference [119], particularly for edge deployment. As model quantization has been discussed in Sec. 3.2, we omit detailed discussion here.

Pruning removes redundant weights to reduce computation [448]. In LIC, pruning is often applied to transform networks to accelerate inference. For example, Liao et al. [236] propose structured pruning for efficient decoder design, achieving substantial speedup with limited performance degradation. Hossain et al. [146] further combine pruning and quantization for joint model compression, achieving better efficiency–performance trade-offs.

3.4.3 Architecture-Level Design

Another important direction is to design inherently lightweight codec architectures by simplifying network structures, reducing sequential dependencies, and improving parallelism [86]. This strategy is often more deployment-friendly, since efficiency is considered from the beginning of codec design.

A central bottleneck is autoregressive entropy modeling, which improves probability estimation but limits decoding speed. To address this issue, Ali *et al.* [8] propose non-autoregressive compression models that improve throughput by relaxing inter-latent dependencies. AsymLIC [372] reduces decoding complexity through an asymmetric encoder–decoder design, shifting more computation to the encoder while keeping the decoder lightweight.

In addition, operator-level designs further improve efficiency. Lightweight attention modules [58, 141], efficient convolutions, and shift-based operations have been explored to reduce redundant computation. For example, ShiftLIC [26] replaces large-kernel convolutions with spatial–channel shift operations and small kernels, while Li *et al.* [221] propose dynamic concatenated convolution (DCC) to reduce filter redundancy. ShallowNTC [403] adopts a shallow, even nearly linear, decoding transform reminiscent of JPEG-style reconstruction, achieving performance competitive with BPG (H.265/HEVC image-profile anchor) at less than 50 KMACs/pixel decoding complexity. FastNIC [423] incorporates switchable priors and achieves BPG-level or better performance on Kodak with encoding and decoding complexities below 12 and 10 KMACs/pixel, respectively. Dynamic transform routing [350] further adaptively allocates computation according to input content, offering a promising path toward content-aware complexity control.

3.4.4 Test-Time Adaptation

Test-time adaptation (TTA) [50, 234] aims to improve inference performance without modifying the model architecture or increasing model size.

One key motivation is the amortization gap in VAE-based compression models, where the encoder may produce suboptimal latent representations, especially under domain shifts [404]. Latent refinement methods address this issue by directly optimizing latent variables during encoding. For example, Campos et al. [41] update latents via gradient descent under the rate–distortion objective. Yang et al. [404] further jointly refine latent variables and side information while mitigating discretization errors.

Another line of work focuses on decoder-side adaptation, where a subset of model parameters is updated and transmitted. Van Rozendaal et al. [313] update decoder and entropy model parameters using rate–distortion optimization, while Lam et al. [205] restrict updates to convolutional biases. Zou et al. [452] introduce lightweight multiplicative parameters for efficient adaptation.

To reduce the overhead of parameter transmission, recent approaches adopt parameter-efficient transfer learning (PETL) techniques [148], introducing lightweight adaptation modules (e.g., adapters) for efficient decoder refinement [264, 323, 359]. Chen et al. [52] further study cross-domain TTA and propose a Bayesian regularization framework to improve distribution modeling in a plug-and-play manner.

3.4.5 Discussion and Perspective

While recent advances have demonstrated the strong potential of learned codecs, bridging the gap between research prototypes and real-world systems, traditionally dominated by decades-old standards, remains a critical challenge. In this context, achieving both superior compression performance and low computational complexity is essential for practical adoption.

Moreover, existing LIC methods are primarily designed for natural images, with limited understanding of their generalization to various data domains, such as screen content [323, 366], artificial-intelligence-generated content (AIGC) [90, 115], and scientific imagery [82, 130, 328]. These data often exhibit distinct characteristics, including

wider dynamic ranges, higher bit-depths, and complex spatial or temporal structures. Systematic evaluation and modeling for such data remain largely underexplored.

In addition, domain generalization and continual learning [95, 382] are important yet insufficiently studied problems. Developing compression models that can robustly handle distribution shifts and evolving data characteristics is a promising direction for future research.

3.5 Standardization Progress

With the rapid maturation of learning-based image compression, international standardization bodies have initiated parallel efforts to bridge the gap between research prototypes and deployable codecs. Early explorations can be traced back to the JPEG XL activity around 2018, where next-generation coding tools began to incorporate learning-based components. Building upon this trend, two major standardization tracks have emerged: IEEE 1857.11-2024 and JPEG AI, standardized as International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) 6048 and International Telecommunication Union Telecommunication Standardization Sector (ITU-T) T.840. Both adopt a classical VAE-style architecture: an end-to-end neural network compresses an image into a compact latent tensor for entropy coding and transmission, and the receiver decodes this tensor to reconstruct the image. Recent survey works [113, 239] also provide overviews of these standardization efforts.

3.5.1 JPEG AI

The JPEG AI standard [11, 12], developed by the ISO/IEC Joint Photographic Experts Group (JPEG), was initiated in 2019 and published as ISO/IEC 6048 / ITU-T T.840, with Part 1 published in 2025 as ISO/IEC 6048-1:2025 / ITU-T T.840.1. It represents the first international standard that fully embraces end-to-end learning-based image compression.

Architecture and Coding Framework. A key feature of JPEG AI is its *multi-branch architecture*, which supports multiple decoder configurations with different computational complexities. Representative decoder complexities range from approximately 8 to 214 KMACs/pixel, enabling flexible deployment across heterogeneous platforms. Notably, a mid-level configuration (around 23 KMACs/pixel) has been demonstrated to be deployable on mobile devices, highlighting the practicality of the design.

To improve coding efficiency and reduce memory consumption, JPEG AI adopts a structured color processing strategy. The input image is first converted to a YUV color space (BT.709), where the luminance (Y) component is encoded independently using a more expressive neural network. The chrominance components (U and V) are then encoded conditionally based on the decoded luminance signal. This *conditional coding* mechanism effectively exploits inter-channel redundancy while reducing peak memory usage, making it particularly suitable for resource-constrained scenarios.

Extended Functionalities. The learned latent representation in JPEG AI enables several practical functionalities beyond conventional reconstruction.

Progressive decoding is supported by ordering latent channels according to their relative importance, allowing incremental reconstruction with gradually improving quality as more channels are decoded. In addition, *partial (spatial) decoding* is enabled by tiling the latent tensor, such that different spatial regions can be decoded independently. This facilitates efficient region-of-interest (ROI) processing and scalable decoding.

Beyond reconstruction, JPEG AI explicitly supports compressed-domain processing. Three representative tasks have been defined, including image classification, super-resolution, and image denoising. This design reflects a forward-looking trend toward integrating image compression with downstream vision and image processing tasks, and it is expected that additional compressed-domain functionalities will be incorporated in future extensions.

Training, Evaluation, and Implementation. JPEG AI is optimized using a weighted combination of distortion metrics, typically including mean squared error (MSE) and structural similarity (SSIM). More importantly, the standard development emphasizes perceptual quality rather than pixel-wise fidelity.

To ensure reliable and consistent evaluation, the JPEG AI Common Test and Training Conditions (CTTC) define a comprehensive set of perceptual quality metrics, including multi-scale structural similarity (MS-SSIM) [380], information-weighted structural similarity (IW-SSIM) [379], Video Multi-Method Assessment Fusion (VMAF)[6] [24], visual information fidelity (VIF) [321], peak signal-to-noise ratio with human-visual-system and masking terms (PSNR-HVS-M) [299], normalized Laplacian pyramid distance (NLPD) [206], and feature similarity index measure (FSIM) [425]. Conventional PSNR is deliberately excluded due to its limited correlation with perceived visual quality. Under these criteria, JPEG AI demonstrates substantial improvements in perceptual coding efficiency compared with traditional codecs.

From a system perspective, JPEG AI adopts a heterogeneous deployment strategy. Entropy decoding is typically executed on central processing units (CPUs) to guarantee bit-exact behavior, while other neural network components are accelerated using neural processing units (NPUs) or graphics processing units (GPUs). This separation balances computational efficiency with standard compliance and facilitates practical deployment across diverse hardware platforms.

Experience and Emerging Studies. As the first international standard for end-to-end learning-based image compression, JPEG AI has attracted significant attention from both academia and industry. A growing body of recent work has begun to systematically analyze its underlying principles, practical performance, and potential extensions.

Several studies focus on understanding the core design and functionality of the standard. For example, [98] provides a comprehensive overview of the technical features and architectural characteristics of JPEG AI. Complementarily, [167] investigates its variable-rate coding mechanism, including a three-dimensional quality control framework, a fast bitrate matching algorithm, and corresponding training strategies.

[6] https://github.com/Netflix/vmaf

Beyond architectural analysis, recent efforts have also examined the robustness and generalization of JPEG AI. [193] conducts a large-scale evaluation under various perturbations and adversarial conditions, showing that although JPEG AI achieves strong compression performance, it remains vulnerable to distribution shifts and adversarial attacks, similar to other LIC models.

In terms of practical applications, JPEG AI has been explored in several emerging scenarios. These include improved rate control strategies [287], domain adaptation for remote sensing imagery [424], and extensions toward compressed-domain computer vision tasks [9].

3.5.2 IEEE 1857.11

The IEEE 1857.11-2024 Standard [7] [338] emerged from the "Future Video Coding Study Group" (FVC-SG) and was formally developed by the IEEE 1857.11 Subworking Group (SWG), established in 2021. It was approved by the IEEE Standards Association (IEEE SA) Board on September, 2024, and officially published on December, 2024.

Architecture and Coding Framework. IEEE 1857.11 defines three profiles—Base, Main, and High—targeting different complexity–performance trade-offs and enabling adaptation to diverse deployment scenarios.

The reference implementations include three representative models: BEE [442], NIC [57], and iWave [265, 266]. The Base profile (BEE) focuses on low complexity and fast decoding, incorporating efficient design choices such as wavefront parallel processing (WPP), adaptive quantization (AQ), and feature scaling. The Main profile (NIC) introduces more advanced components, including non-local attention, hierarchical priors, and 3D masked entropy models, achieving a balance between coding efficiency and computational complexity. The High profile (iWave) targets maximum performance by employing a learned wavelet-like transform within a VAE-style framework, enabling highly efficient representation of image content.

Training, Evaluation, and Implementation. During its verification phase, IEEE 1857.11 evaluated models on datasets spanning multiple resolutions, including 6K, 4K, 2K, and 768×512 (the latter corresponding to the Kodak dataset). Its Common Training Conditions (CTC) mandate the use of a large-scale ultra-HD dataset [243], with training images randomly cropped into 256×256 patches.

For performance evaluation, the standard primarily targets improvements in MS-SSIM and PSNR. Compared with BPG (the H.265/HEVC image-profile anchor), IEEE 1857.11 achieves bitrate savings of over 30% at equivalent quality levels under these metrics.

Experience and Emerging Studies. As a recently established standard following JPEG AI, IEEE 1857.11 has attracted increasing attention in the research community. It has been widely adopted as a baseline in subsequent studies [199, 231, 409].

[7] https://sagroups.ieee.org/fvc/

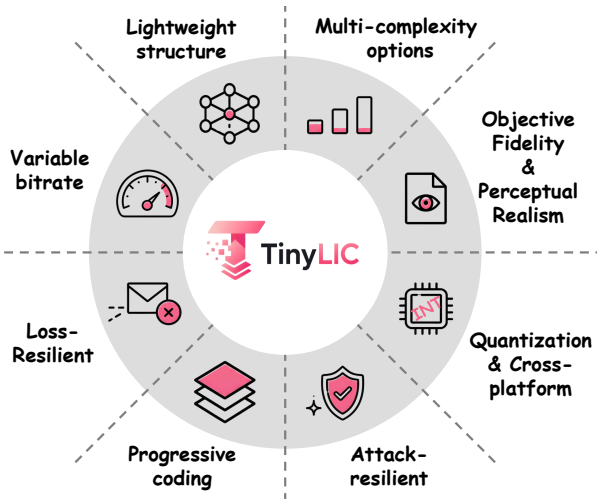


Fig. 8 TinyLIC as a practical benchmark for learned image coding.

3.5.3 Discussion and Perspective

As emerging standards, learning-based image coding frameworks have already demonstrated clear performance advantages over conventional formats such as JPEG. AI-based image coding may become an important paradigm and is likely to coexist with conventional standards in emerging AI-enabled visual communication systems. Beyond improved rate–distortion performance, its integration with intelligent systems—such as autonomous agents [391] and embodied intelligence [187]—enables functionalities that are difficult to achieve with traditional hand-crafted codecs. In addition, a key advantage of learning-based approaches lies in their adaptability. While the overall coding framework remains fixed, performance can be continuously improved through advances in training strategies, model architectures, and data, without requiring fundamental changes to the coding pipeline. This property provides a scalable path for long-term evolution.

4 TinyLIC as a Reproducible Case Study for Deployment-Oriented LIC

This section presents **TinyLIC** as a reproducible case study for deployment-oriented learned image coding, rather than as the primary contribution of the survey. Its role is to operationalize the design questions reviewed above: how R–D performance changes when complexity, rate control, numerical consistency, robustness, progressive delivery, and network resilience are treated as coupled deployment constraints. TinyLIC therefore serves as an open, lightweight reference implementation and benchmark that makes these trade-offs measurable under a common architecture and evaluation protocol. Unlike recent codecs optimized primarily for raw R–D performance, TinyLIC provides a transparent and extensible baseline for analyzing rate, distortion, complexity, controllability, robustness, and deployment behavior under reproducible

conditions. The purpose of this section is therefore explanatory: TinyLIC is used to illustrate how architectural choices and deployment tools interact, while the broader survey remains the main contribution of the article.

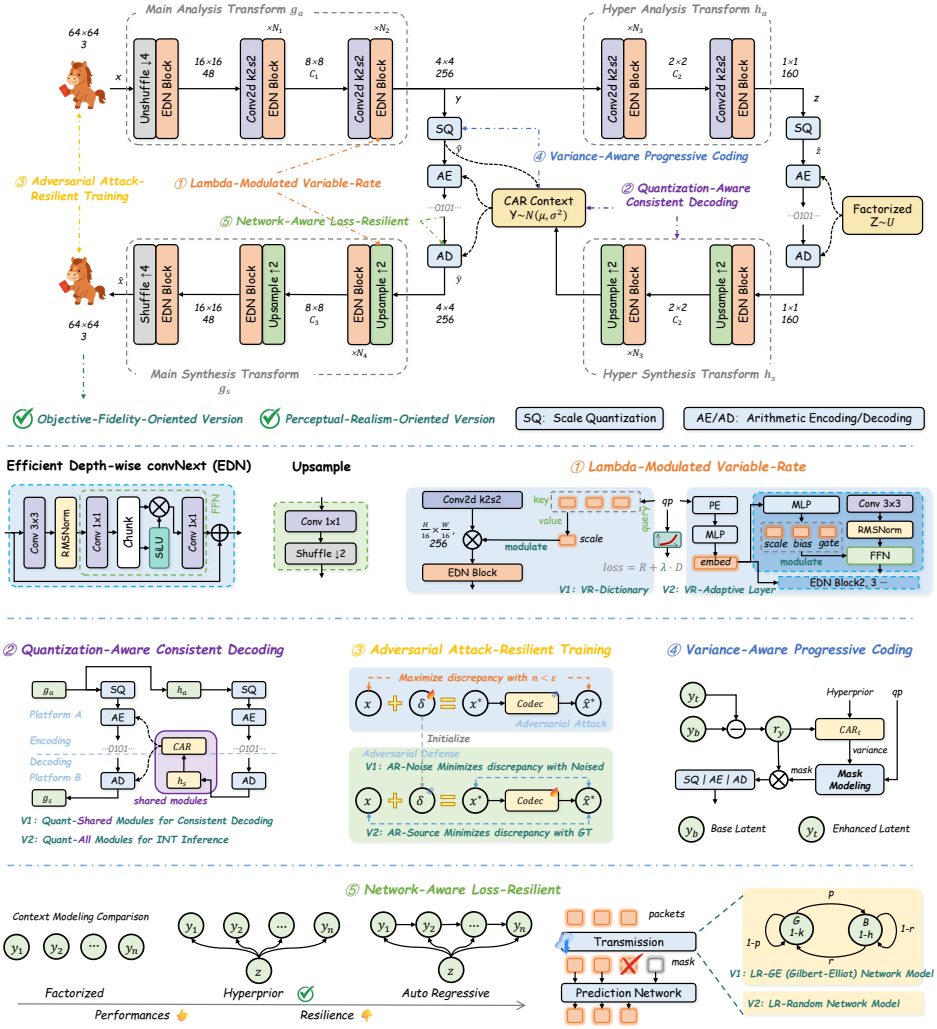


Fig. 9 TinyLIC employs a lightweight VAE-based architecture using a channel-autoregressive entropy model and efficient depthwise ConvNeXt-style blocks. It further integrates five configurable tools for variable-rate coding, consistent decoding, adversarially robust training, progressive transmission, and loss resilience.

4.1 Design Philosophy

TinyLIC is guided by three principles: *simplicity*, *diagnosability*, and *extensibility*. First, a compact architecture keeps computational cost and module contributions

transparent. Second, key system-level tools are exposed as optional modules, enabling controlled analysis of rate, distortion, complexity, and robustness trade-offs. Third, a modular design permits new transforms, entropy models, rate-control strategies, or deployment utilities to be integrated without modifying the core infrastructure.

4.1.1 Efficient Architecture

As illustrated in Fig. 9, TinyLIC follows a VAE-based pipeline:

$$y = g_a(x), \quad \hat{y} = Q(y), \quad \hat{x} = g_s(\hat{y}), \tag{13}$$

where g_a and g_s denote the main analysis and synthesis transforms, and $Q(\cdot)$ denotes scalar quantization. A hyperprior branch further estimates side information:

$$z = h_a(y), \quad \hat{z} = Q(z), \quad \psi = h_s(\hat{z}), \tag{14}$$

where ψ provides entropy parameters for coding $\hat{y}$. The overall training objective follows the standard Lagrangian R–D formulation:

$$\mathcal{L} = R(\hat{y}, \hat{z}) + \lambda D(x, \hat{x}), \tag{15}$$

where R denotes the estimated bitrate and D is either a fidelity-oriented or perceptual distortion measure.

The main transforms are built primarily from convolutional blocks rather than expensive attention modules. This design reflects the observation that, under strict complexity constraints, well-designed convolutional networks often provide a more favorable performance–complexity trade-off.

Lightweight design. The basic unit of TinyLIC is an efficient depthwise ConvNeXt-style block, termed the EDN block. Its design is selected through ablations on activation functions, normalization, residual placement, and re-parameterization, with the goal of maximizing R–D gain under low-complexity budgets.

For entropy modeling, TinyLIC uses Gaussian likelihood estimation with a lightweight channel autoregressive (CAR) context model:

$$p(\hat{y}|\hat{z}) = \prod_{t=1}^{T} p(\hat{y}_t \mid \hat{y}_{<t}, h_s(\hat{z})), \quad \hat{y}_t \sim \mathcal{N}(\mu_t, \sigma_t^2), \tag{16}$$

where $\hat{y}_t$ denotes the t-th channel group and T is the number of autoregressive steps. By default, TinyLIC uses a 4-step CAR model, which balances coding efficiency and computational overhead.

In this implementation, CAR is more suitable for lightweight deployment than spatial autoregressive models. Spatial autoregression usually depends on masked convolutions. The mask limits the accessible spatial context, but the convolution itself is still computed over the entire feature map. Therefore, SAR reduces context visibility rather than the actual spatial computation. In contrast, CAR only involves

the required channel groups when predicting the current group, while unused channels can be excluded from computation. This makes its complexity easier to control by adjusting the number of channel groups, which is more friendly to lightweight deployment.

Complexity-aware scaling. TinyLIC offers three complexity configurations, approximately 10, 20, and 50 KMACs per pixel, achieved by scaling channel width and block depth while preserving the codec topology. These tiers target smart glasses and other edge devices, mobile platforms, and higher-end platforms such as laptops and servers, respectively. This design enables systematic analysis of the scaling relationship between model complexity and compression performance.

Fidelity- and perceptual-realism-oriented variants. The fidelity-oriented variant employs conventional R–D optimization with MSE loss, serving as a baseline for PSNR benchmarking. The perceptual-realism-oriented variant incorporates feature-space losses (AlexNet- and VGG-based, following LPIPS [427]) and adversarial training to enhance visual quality at low bitrates. Together, these variants enable controlled study of the rate–distortion–perception trade-off.

4.1.2 Practical Trade-offs

Beyond architectural efficiency, TinyLIC is designed to study system-level trade-offs that are often overlooked in existing studies. As shown in Fig. 9, TinyLIC integrates five optional tools: λ-modulated variable-rate coding, quantization-aware consistent decoding, adversarially attack-resilient training, variance-aware progressive coding, and network-aware loss-resilient coding. These tools are configurable rather than mandatory, reflecting the reality that enhancing properties like robustness or deployability often incurs R–D or complexity costs.

Controllability: λ-modulated variable-rate coding. Rate control is essential for practical codecs. TinyLIC supports flexible rate control through a λ-modulated variable-rate framework, where the Lagrange multiplier λ serves as a conditioning signal to control the R–D operating point. Feature modulation is adopted because it is simple, effective, and largely plug-and-play: it does not modify the codec topology and can be integrated into different transform and entropy-model architectures. Given an intermediate feature f, modulation is formulated as

$$f' = f + \gamma(\lambda) \odot (\alpha(\lambda) \odot f + \beta(\lambda)), \tag{17}$$

where $\alpha(\lambda)$, $\beta(\lambda)$, and $\gamma(\lambda)$ denote the learned scale, bias, and gate parameters used for modulation, respectively. TinyLIC implements two different strategies.

- **VR-Adaptive Layer.** The adaptive-layer strategy uses a lightweight multilayer perceptron (MLP) to map λ into modulation parameters dynamically, enabling continuous rate control through MLP-based modulation.
- **VR-Dictionary.** The dictionary-based strategy defines a discrete set of modulation parameters indexed by QP, where each QP is trained to correspond to a specific λ. During inference, the target bitrate is first mapped to a QP, which then retrieves the associated modulation parameters.

Deployability: quantization-aware consistent decoding. A practical codec must produce numerically consistent decoding results across heterogeneous platforms. However, floating-point implementations may introduce small discrepancies, especially in entropy-parameter networks executed on different devices or libraries. Since entropy decoding depends on identical probability estimation between encoder and decoder, such discrepancies may cause decoding mismatch or failure.

TinyLIC therefore introduces quantization-aware training and INT8 inference. The quantization operator is defined as

$$Q(x) = \text{clip}\left(\left\lfloor \frac{x}{s} \right\rceil, q_{\min}, q_{\max}\right) s, \tag{18}$$

where s denotes the quantization scale. Weights are quantized channel-wise, while activations are generally quantized layer-wise. For depth-wise convolutions, channel-wise activation quantization can also be applied since channel interactions are absent. TinyLIC supports two quantization modes.

- **Quant-Shared.** Only the modules shared between encoder and decoder are quantized, including the hyper-synthesis transform and entropy-parameter network. This is sufficient to ensure cross-platform entropy decoding consistency while preserving most transform modules in floating point.
- **Quant-All.** For hardware accelerators such as NPUs, TinyLIC additionally supports full INT8 inference through quantization-aware training initialized from pretrained 32-bit floating-point (FP32) models. Although full quantization may slightly reduce R–D performance, it improves hardware compatibility, inference speed, memory efficiency, and energy efficiency.

Reliability: adversarial attack-resilient training. Neural codecs may be sensitive to out-of-distribution inputs, repeated compression, and adversarial perturbations. This issue is particularly important when compression is used as a preprocessing stage for downstream vision tasks, where perturbations targeting the codec may propagate through the entire perception pipeline. TinyLIC incorporates adversarially robust training strategies inspired by existing robustness studies [55] without introducing additional runtime overhead.

- **AR-Source.** AR-Source minimizes the discrepancy between the reconstruction from the adversarial input and the clean source image, encouraging the codec to recover clean content from perturbed inputs, using adversarial examples generated with a Carlini–Wagner (CW)-style optimization objective [44].
- **AR-Noise.** AR-Noise minimizes the discrepancy between the reconstruction of x^* and that of a noise-perturbed reference, encouraging consistent reconstruction under small input perturbations [55, 200].

Progressiveness: variance-aware progressive coding. Progressive transmission is important for bandwidth-limited or coarse-to-fine reconstruction scenarios.

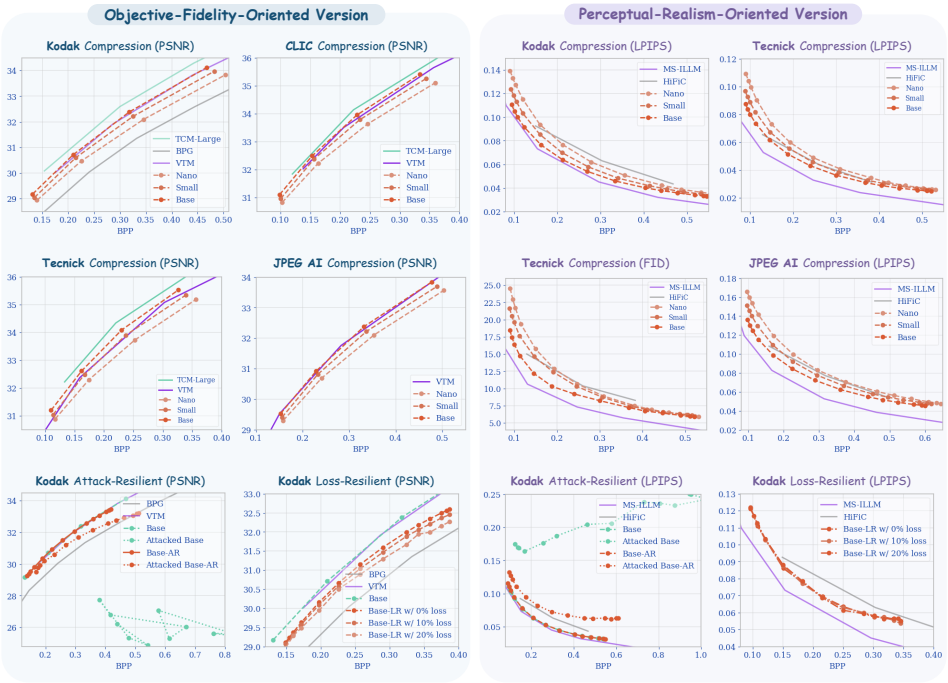


Fig. 10 Rate–distortion and perceptual performance of TinyLIC under different complexity budgets and practical tool configurations.

TinyLIC adopts a variance-aware progressive coding strategy inspired by recent efficient progressive coding methods [302]. This design requires only minor modifications to the standard VAE pipeline and introduces little additional complexity.

The key observation is that latent elements with larger predicted variances are often more informative. During encoding, latent elements are ranked according to the variance predicted by the entropy model and transmitted progressively. Under bandwidth constraints, untransmitted elements are replaced by their predicted means:

$$\tilde{y}_i = m_i \hat{y}_i + (1 - m_i)\mu_i, \tag{19}$$

where $m_i \in \{0, 1\}$ denotes whether the i-th latent element is transmitted. This enables coarse-to-fine reconstruction from partial bitstreams while preserving the standard VAE-based codec structure.

Resilience: network-aware loss-resilient coding. In practical communication systems, compressed bitstreams may suffer from packet loss, burst errors, and latency constraints. A codec optimized only for ideal transmission may fail when part of the bitstream is missing. TinyLIC therefore incorporates network-aware loss-resilient training by simulating packet loss during optimization.

For this module, TinyLIC removes the autoregressive entropy model and adopts a hyperprior-only architecture. This is based on the performance–resilience trade-off:

autoregressive dependencies improve entropy modeling under error-free transmission, but may amplify error propagation when packets are lost. Recent studies such as [318, 373] show that different autoregressive dependencies can significantly affect robustness under transmission loss. Therefore, the loss-resilient version keeps the transform architecture unchanged while simplifying the entropy model. Two network models are considered during training.

- **LR-GE (Gilbert–Elliot model).** The Gilbert–Elliot model is a two-state Markov model alternating between *good* and *bad* states, enabling simulation of burst packet losses.
- **LR-Random model.** The random loss model assumes independent packet loss with a fixed probability, providing a simple baseline for stochastic transmission errors.

4.2 Results and Analysis

4.2.1 Implementation Details

Training protocol. TinyLIC is trained on the MLIC-100K dataset [175]. Unless otherwise specified, all models are trained for 400 epochs with a batch size of 32. The initial learning rate is set to 2×10^{-4} and decayed to 2×10^{-5} after 300 epochs. Fidelity-oriented models are optimized with the standard R–D objective using MSE distortion, while realism-oriented models additionally incorporate perceptual losses as described in Sec. 4. For reproducibility, the benchmark should report the optimizer, training hardware, entropy-coder implementation, random-seed policy, and code/checkpoint availability together with the released implementation.

Test datasets. We evaluate TinyLIC on Kodak[8], Tecnick [13], CLIC Professional validation[9], and the JPEG AI test set [12], covering diverse resolutions and content distributions. Kodak contains 24 images of size 768×512; Tecnick contains high-quality 1200×1200 images; CLIC Professional validation contains 41 images with approximately 2K resolution; and the JPEG AI test set contains 50 natural images ranging from 1K to 4K. For high-resolution JPEG AI images, inference is performed using non-overlapping or tiled patches no larger than 1024×1024 to avoid memory overflow.

Compared methods. We include representative traditional and learned codecs as anchors. For objective fidelity-oriented evaluation, VTM Intra 22.0 under the H.266/VVC image-profile configuration [34] and BPG (the H.265/HEVC image-profile anchor) [28] are used as conventional baselines. We compare with recent lightweight learned codecs, including EVC [367], AsymLIC [372], ShiftLIC [26], CSLIC [374], ShallowNTC [403], and FastNIC [423], which emphasize different complexity–compression trade-offs through efficient architectures, scalable designs, or lightweight entropy models.

For perceptual realism-oriented evaluation, we compare with representative perceptual codecs such as HiFiC [273] and MS-ILLM [279]. HiFiC is a classical GAN-based

[8] https://r0k.us/graphics/kodak/
[9] https://clic2025.compression.cc/

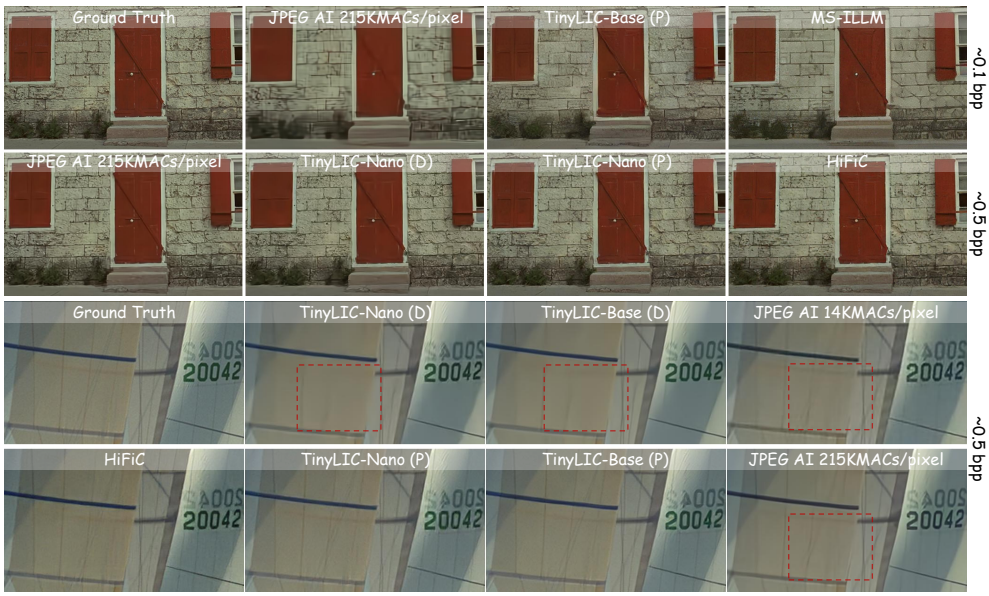


Fig. 11 Qualitative comparison of TinyLIC with representative codecs. Fidelity-oriented variants (D) preserve structural fidelity, while perceptual-realism-oriented variants (P) recover sharper textures and more visually realistic details.

neural image codec optimized for perceptual quality, while MS-ILLM represents a recent strong baseline.

4.2.2 Rate–Distortion–Perception Performance

Table 5 summarizes the performance of TinyLIC under different complexity budgets and practical tool configurations. The fidelity-oriented results are reported as PSNR BD-rate with VTM Intra as the H.266/VVC image-profile anchor, while realism-oriented results are reported using LPIPS and FID BD-rate with MS-ILLM as the anchor.

Complexity scaling. TinyLIC exhibits a clear performance–complexity scaling trend. Increasing the complexity budget from 10 to 20 KMACs/pixel significantly improves fidelity-oriented performance, reducing the Kodak PSNR BD-rate from 14.78% to 5.43%. The 50 KMACs/pixel Base model further matches or slightly surpasses the VTM Intra anchor on Kodak, CLIC, and Tecnick, reaching -0.79%, -1.76%, and -4.21% BD-rate, respectively. TinyLIC adopts tile-based coding for JPEG AI test images due to memory and deployment considerations, whereas VTM Intra performs full-resolution coding. This protocol mismatch slightly disadvantages TinyLIC in R–D evaluation.

These results suggest that a compact convolutional architecture with lightweight entropy modeling can achieve competitive R–D performance without heavy attention mechanisms. Performance gains also diminish at higher complexity levels, indicating saturation in model capacity. This observation aligns with recent scaling-law

studies [229], which show that diminishing returns persist even in learned codecs approaching 1B parameters.

Perceptual-realism-oriented performance. For perceptual optimization, TinyLIC improves over HiFiC under the reported LPIPS/FID BD-rate protocol. The perceptual gap to MS-ILLM consistently decreases as complexity increases. For example, the FID BD-rate on Tecnick improves from 91.57% for Nano to 35.47% for Base TinyLIC. These results demonstrate that the same lightweight backbone can also serve as an effective platform for studying perceptual compression.

4.2.3 Effect of Practical Tools

The lower portion of Table 5 evaluates the practical system tools introduced in Sec. 4. The results demonstrate a consistent trade-off between compression efficiency and the added system functionalities.

Variable-rate behavior. The variable-rate variants maintain or slightly improve compression performance over the fixed-rate baseline. This improvement is partly related to the BD-rate computation protocol: BD-rate relies on interpolation over discrete R–D points, and sparse fixed-rate operating points may introduce larger interpolation errors. In our evaluation, the fixed-rate setting contains only four operating points, whereas the variable-rate models use 15, leading to denser bitrate coverage and a more reliable R–D curve estimation.

Among the variants, VR-Adaptive Layer achieves the best fidelity-oriented performance, reaching -1.50%, -2.75%, and -5.02% BD-rate gain over VTM Intra on Kodak, CLIC, and Tecnick, respectively. Nevertheless, we use VR-Dictionary in the following sections for its simplicity, as it only scales one activation layer in each of the analysis and synthesis transforms while retaining effective variable rate.

Quantization-aware decoding. Quant-Shared introduces only a limited performance drop relative to the VR-Dictionary baseline, while enabling consistent cross-platform entropy decoding. By contrast, Quant-All introduces a larger but still moderate R–D penalty, which is expected because all transform and entropy-model components are quantized for hardware-friendly INT8 inference.

Progressive and robust coding. Variance-aware progressive coding introduces a moderate penalty in both distortion and perceptual metrics, reflecting the cost of enforcing importance-aware latent ordering. Attack-resilient training results in a smaller degradation, with AR-Noise consistently outperforming AR-Source on clean-image evaluation. Importantly, these tools preserve the same codec backbone, enabling controlled studies of robustness and progressiveness without redesigning the architecture.

Loss-resilient coding. The loss-resilient variants exhibit the largest clean-image performance degradation. This is expected because the autoregressive entropy model is removed to reduce error propagation under packet loss. The similar behavior of LR-GE and LR-Random further suggests that the primary cost originates from the resilience-oriented entropy-model simplification rather than the specific channel model itself.

4.2.4 Visual Comparison

We further provide qualitative comparisons to illustrate the behavior of TinyLIC under different complexity levels and optimization objectives. The visual results include distortion-oriented reconstructions, perception-oriented reconstructions, and comparisons with JPEG AI operating points of different complexities.

The perceptual realism-oriented models produce sharper and more perceptually plausible textures in the shown examples at low bitrates, although they may introduce more synthesized high-frequency content. TinyLIC achieves a favorable balance between visual quality and computational efficiency. These qualitative results are consistent with the quantitative findings and further confirm that TinyLIC can serve as a compact yet effective platform for analyzing practical learned image compression systems.

4.2.5 Model Complexity

Table 6 reports the computational complexity and parameter distribution of TinyLIC. The three variants provide a clear lightweight scaling path from Nano to Base, covering approximately 10, 20, and 50 KMACs/pixel decoding budgets. Even the largest Base model contains only 10.567M parameters, while the Nano model requires 9.99 KMACs/pixel for decoding with 3.161M parameters, making it suitable for resource-constrained deployment.

4.3 Discussion and Perspective

TinyLIC is not intended to maximize raw R–D performance, but to provide a lightweight, reproducible, and configurable framework for practical learned image compression. By combining a compact backbone with modular system-level tools, it enables controlled analysis of the interplay among compression efficiency, computational complexity, deployability, and robustness.

More broadly, TinyLIC reflects an important transition in learned image coding research. Early studies primarily focused on improving R–D performance under idealized settings, whereas practical deployment requires codecs to satisfy a much wider set of constraints, including hardware efficiency, numerical consistency, transmission reliability, and flexible rate adaptation. These objectives are often coupled and sometimes conflicting, making codec design increasingly a system-level optimization problem rather than a purely algorithmic one. Our experiments suggest that many practical functionalities can be integrated into a unified lightweight framework with manageable overhead, although each additional constraint typically introduces some performance trade-off.

In this sense, TinyLIC serves not only as a lightweight reference codec, but also as a practical research platform for studying the broader design space of neural image compression, where coding performance must be balanced together with complexity, robustness, and real-world usability.

5 Concluding Remarks & Future Perspective

Remarks. Over the past decade, end-to-end LIC has matured from proof-of-concept neural compression into an increasingly deployable codec paradigm. This review traces that transition through nonlinear transforms, differentiable signal quantization, entropy modeling, and rate–distortion optimization, emphasizing how data-driven joint optimization replaces hand-crafted module design while preserving the core compression goal of compact and faithful representation.

The field is now judged not only by PSNR or perceptual R–D gains, but also by computational complexity, memory footprint, numerical consistency, hardware compatibility, robustness, rate control, and transmission reliability. TinyLIC is used as a lightweight and reproducible reference case to make these coupled trade-offs visible under controlled conditions, rather than as a claim of best raw compression performance.

The standardization progress of JPEG AI and IEEE 1857.11 shows that learned image coding is entering industrial validation, with early foundations for bitstream syntax, model profiles, complexity levels, conformance testing, and interoperability. Remaining challenges in model evolution, hardware support, and long-term maintenance define the next stage of practical adoption.

Perspective. The long-term potential of LIC lies in enabling *coding for intelligence*: a coding paradigm in which compressed representations are no longer optimized solely for human-view reconstruction, but can also flexibly support downstream perception, analysis, and decision-making. Although most existing studies still focus on compression and reconstruction for human visual quality, recent efforts in both academia and industry have increasingly explored machine-task-oriented coding [426] and unified frameworks that jointly consider human visual reconstruction and machine perception [402]. However, many existing solutions remain at an early stage. Some rely on task-specific encoders, while others require joint optimization with downstream task networks, which limits scalability and creates barriers to rapid deployment. A more promising direction is to construct a unified compressed bitstream that can, through lightweight control or adaptive decoding, support multiple objectives such as human-view reconstruction, machine perception, and semantic analysis [94, 432, 433]. This perspective is also consistent with biological vision, where the visual system does not preserve all sensory details uniformly, but rapidly extracts compact and task-relevant representations to support higher-level cognition, decision-making, and memory. Such a mechanism provides an intuitive basis for the emerging view that compression and intelligence are deeply connected [155, 444].

At the representation level, emerging paradigms beyond conventional latent-grid VAEs are also expanding the design space of learned coding. Implicit neural representations (INRs) model signals as continuous coordinate-conditioned functions, enabling flexible resolution scaling and compact continuous representations [333]. Meanwhile, Gaussian-based representations [182] provide an alternative explicit scene representation through learnable Gaussian primitives, achieving highly efficient rendering and view synthesis. Recent studies have begun exploring compression directly in INR and Gaussian domains [184, 203, 326, 327, 339], suggesting that future codecs may move beyond traditional pixel- or latent-grid representations toward more general

neural scene representations. These paradigms are particularly attractive for immersive media, free-viewpoint rendering, and AI-native visual communication, where the goal is not merely reconstructing pixels, but representing underlying scene structure, geometry, and semantics in a compact and editable form.

The rapid development of large models further expands the design space of learned coding. Large language, vision, and multimodal models have demonstrated strong capabilities in semantic understanding, context prediction, and visual generation, while parallel progress in model compression and edge deployment is making their practical use increasingly feasible [445]. This convergence suggests a new route for image and video coding: instead of transmitting all visual details explicitly, future codecs may transmit compact structural, semantic, or latent information, and rely on powerful generative priors to recover perceptually plausible content. Recent studies have already shown that large-model-based context modeling can substantially improve lossless compression efficiency across text, images, video, and audio [230]. Similarly, extremely low-bitrate visual reconstruction can be achieved by guiding diffusion models with semantic descriptions generated by large language models and coarse latent features extracted from variational autoencoders [43, 181]. These results indicate that foundation models may shift the relevant evaluation frontier from signal fidelity alone toward rate–distortion–perception and semantic reliability, especially in ultra-low-bitrate scenarios such as emergency communication, satellite links, deep-space communication, and deep-sea exploration. At the same time, this paradigm also raises new challenges: reconstructed content may become more generative than strictly faithful, making hallucination, semantic consistency, controllability, and computational cost central issues for future evaluation.

Although learned image and video coding share many core techniques, their deployment timelines are likely to differ. For still-image coding, the combination of strong compression performance, manageable complexity, mobile NPU support, and standardization progress suggests nearer-term deployment potential than learned video coding. For video coding, the situation is more challenging due to stricter requirements on latency, power consumption, memory bandwidth, temporal consistency, and random access [51, 112, 183, 227, 362, 387]. Its adoption will likely depend on continued advances in semiconductor technology, NPU efficiency, model compression, and standardization ecosystems.

Overall, learned coding is moving from neural compression toward AI-native media representation. Future codecs will need to jointly optimize *compression efficiency*, *system deployability*, and *intelligence compatibility*. The convergence of these dimensions may define a new coding paradigm in which compression is no longer only a tool for bitrate reduction, but also an interface between visual sensing, communication, perception, and intelligent decision-making.

Statements and Declarations

Funding. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62431011.

Competing interests. The authors declare no competing interests.

Data and code availability. Materials associated with TinyLIC, including the open project website and benchmark information, are accessible at https://njuvision.github.io/TinyLIC/. Public datasets and third-party implementations discussed in this survey are available from their original sources as cited in the manuscript.

Author contributions. Ming Lu contributed to manuscript drafting. Junqi Shi contributed to experimental studies and the open benchmark. Zhan Ma contributed to manuscript polishing and resource support. All authors reviewed and approved the manuscript.

References

[1] Agustsson E, Theis L (2020) Universally quantized neural compression. Advances in neural information processing systems 33:12367–12376

[2] Agustsson E, Mentzer F, Tschannen M, et al (2017) Soft-to-hard vector quantization for end-to-end learning compressible representations. In: Advances in Neural Information Processing Systems, pp 1141–1151

[3] Agustsson E, Tschannen M, Mentzer F, et al (2019) Generative adversarial networks for extreme learned image compression. In: IEEE International Conference on Computer Vision (CVPR), pp 221–231

[4] Agustsson E, Minnen D, Toderici G, et al (2023) Multi-realism image compression with a conditional generator. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 22324–22333

[5] Ahmed N, Natarajan T, Rao KR (1974) Discrete cosine transform. IEEE transactions on Computers 100(1):90–93

[6] Akbari M, Liang J, Han J, et al (2020) Learned variable-rate image compression with residual divisive normalization. In: 2020 IEEE International Conference on Multimedia and Expo (ICME), IEEE, pp 1–6

[7] Akbari M, Liang J, Han J, et al (2021) Learned multi-resolution variable-rate image compression with octave-based residual blocks. IEEE Transactions on Multimedia 23:3013–3021

[8] Ali MS, Kim Y, Qamar M, et al (2023) Towards efficient image compression without autoregressive models. Advances in Neural Information Processing Systems 36:7392–7404

[9] Alkhateeb A, Gnutti A, Guerrini F, et al (2024) Jpeg ai compressed domain face detection. In: 2024 IEEE 26th International Workshop on Multimedia Signal Processing (MMSP), IEEE, pp 1–6

[10] Allemand F, Fiandrotti A, Chaudhuri S, et al (2025) Efficient learned image compression through knowledge distillation. arXiv preprint arXiv:250910366

[11] Alshina E, Ascenso J, Ebrahimi T (2024) Jpeg ai: The first international standard for image coding based on an end-to-end learning-based approach. IEEE MultiMedia 31(4):60–69

[12] Ascenso J, Alshina E, Ebrahimi T (2023) The jpeg ai standard: Providing efficient human and machine visual data consumption. IEEE MultiMedia 30(1):100–111. https://doi.org/10.1109/MMUL.2023.3245919

[13] Asuni N, Giachetti A, et al (2014) Testimages: a large-scale archive for testing visual devices and basic image processing algorithms. In: STAG, pp 63–70

[14] Bai Y, Yang X, Liu X, et al (2022) Towards end-to-end image compression and analysis with transformers. In: Proceedings of the AAAI conference on artificial intelligence, pp 104–112

[15] Baig MH, Koltun V, Torresani L (2017) Learning to inpaint for image compression. Advances in Neural Information Processing Systems 30

[16] Baldassarre F, El-Nouby A, Jégou H (2023) Variable rate allocation for vector-quantized autoencoders. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 1–5

[17] Ballé J (2018) Efficient nonlinear transforms for lossy image compression. In: IEEE Picture Coding Symposium (PCS), pp 248–252

[18] Ballé J, Laparra V, Simoncelli EP (2015) Density modeling of images using a generalized normalization transformation. arXiv preprint arXiv:151106281

[19] Ballé J, Laparra V, Simoncelli EP (2017) End-to-end optimized image compression. In: 5th International Conference on Learning Representations, ICLR 2017

[20] Ballé J, Johnston N, Minnen D (2018) Integer networks for data compression with latent-variable models. In: International Conference on Learning Representations

[21] Ballé J, Minnen D, Singh S, et al (2018) Variational image compression with a scale hyperprior. In: International Conference on Learning Representations

[22] Ballé J, Chou PA, Minnen D, et al (2020) Nonlinear transform coding. IEEE Journal of Selected Topics in Signal Processing 15(2):339–353

[23] Ballé J, Laparra V, Simoncelli EP (2016) End-to-end optimization of nonlinear transform codes for perceptual quality. In: 2016 Picture Coding Symposium (PCS), pp 1–5, https://doi.org/10.1109/PCS.2016.7906310

[24] Bampis CG, Li Z, Bovik AC (2018) Spatiotemporal feature integration and model fusion for full reference video quality assessment. IEEE Transactions on Circuits and Systems for Video Technology 29(8):2256–2270

[25] Bao J, Basu P, Dean M, et al (2011) Towards a theory of semantic communication. In: 2011 IEEE Network Science Workshop, IEEE, pp 110–117

[26] Bao Y, Tan W, Jia C, et al (2025) Shiftlic: lightweight learned image compression with spatial-channel shift operations. IEEE Transactions on Circuits and Systems for Video Technology

[27] Bao Y, Cheng Y, Liu Y, et al (2026) Dynaquant: Dynamic mixed-precision quantization for learned image compression. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 14475–14483

[28] Bellard F (2016) Bpg image format (2014). URL https://bellard org/bpg 4

[29] Bengio Y, Léonard N, Courville A (2013) Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:13083432

[30] Bjontegaard G (2001) Calculation of average psnr differences between rd-curves. VCEG-M33

[31] Blau Y, Michaeli T (2018) The perception-distortion tradeoff. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6228–6237

[32] Blau Y, Michaeli T (2019) Rethinking lossy compression: The rate-distortion-perception tradeoff. In: International Conference on Machine Learning, PMLR, pp 675–685

[33] Blei DM, Kucukelbir A, McAuliffe JD (2017) Variational inference: A review for statisticians. Journal of the American statistical Association 112(518):859–877

[34] Bross B, Chen J, Ohm JR, et al (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

[35] Bross B, Wang YK, Ye Y, et al (2021) Overview of the versatile video coding (vvc) standard and its applications. IEEE Transactions on Circuits and Systems for Video Technology 31(10):3736–3764

[36] Bychkov G, Abud K, Kovalev E, et al (2025) Nic-robustbench: A comprehensive open-source toolkit for neural image compression and robustness analysis. arXiv preprint arXiv:250619051

[37] Cai C, Chen L, Zhang X, et al (2018) Efficient variable rate image compression with multi-scale decomposition network. IEEE Transactions on Circuits and Systems for Video Technology 29(12):3687–3700

[38] Cai C, Chen L, Zhang X, et al (2019) End-to-end optimized roi image compression. IEEE Transactions on Image Processing 29:3442–3457

[39] Cai S, Zhang Z, Chen L, et al (2022) High-fidelity variable-rate image compression via invertible activation transformation. In: Proceedings of the 30th ACM International Conference on Multimedia, pp 2021–2031

[40] Cai S, Chen L, Zhang Z, et al (2024) I2c: Invertible continuous codec for high-fidelity variable-rate image compression. IEEE Transactions on Pattern Analysis and Machine Intelligence 46(6):4262–4279

[41] Campos J, Meierhans S, Djelouah A, et al (2019) Content adaptive optimization for neural image compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp 0–0

[42] Cao Z, Bao Y, Meng F, et al (2024) Enhancing adversarial training with prior knowledge distillation for robust image compression. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 3430–3434

[43] Careil M, Muckley MJ, Verbeek J, et al (2024) Towards image compression with perfect realism at ultra-low bitrates. In: The Twelfth International Conference on Learning Representations

[44] Carlini N, Wagner D (2017) Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp), Ieee, pp 39–57

[45] Caron M, Touvron H, Misra I, et al (2021) Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9650–9660

[46] Chamain LD, Cheung ScS, Ding Z (2019) Quannet: Joint image compression and classification over channels with limited bandwidth. In: 2019 IEEE International Conference on Multimedia and Expo (ICME), IEEE, pp 338–343

[47] Chamain LD, Racapé F, Bégaint J, et al (2021) End-to-end optimized image compression for machines, a study. In: 2021 Data Compression Conference (DCC), IEEE, pp 163–172

[48] Chen B, Yin S, Chen P, et al (2024) Generative visual compression: A review. In: 2024 IEEE International Conference on Image Processing (ICIP), IEEE, pp 3709–3715

[49] Chen C, Zhang H, Guo K, et al (2024) Exploring efficient hardware accelerator for learning-based image compression. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 44(6):2204–2217

[50] Chen D, Wang D, Darrell T, et al (2022) Contrastive test-time adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 295–305

[51] Chen H, Xu B, Wang S, et al (2024) Toward adaptive volumetric video streaming: A joint network-viewport adaptation framework. IEEE Communications Magazine 63(3):197–203

[52] Chen K, Zhang P, Qin T, et al (2025) Test-time adaptation for image compression with distribution regularization. In: 13th International Conference on Learning Representations (ICLR 2025), International Conference on Learning Representations, ICLR, pp 23686–23703

[53] Chen LH, Bampis CG, Li Z, et al (2020) Proxiqa: A proxy approach to perceptual optimization of learned image compression. IEEE Transactions on Image Processing 30:360–373

[54] Chen T, Ma Z (2020) Variable bitrate image compression with quality scaling factors. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 2163–2167

[55] Chen T, 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

[56] Chen T, Liu H, Shen Q, et al (2017) Deepcoder: A deep neural network based video compression. In: 2017 IEEE Visual Communications and Image Processing (VCIP), IEEE, pp 1–4

[57] Chen T, Liu H, Ma Z, et al (2021) End-to-end learnt image compression via non-local attention optimization and improved context modeling. IEEE Transactions on Image Processing 30:3179–3191

[58] Chen X, Xie F, Li Q, et al (2026) A lightweight transformer model with high-throughput for image compression in 6g-enabled intelligent transportation systems. IEEE Transactions on Intelligent Transportation Systems

[59] Chen Y, Lyu Z, He B, et al (2025) Knowledge distillation for learned image compression. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 4996–5006

[60] Chen Y, Lyu Z, He B, et al (2025) Content-aware mamba for learned image compression. arXiv preprint arXiv:250802192

[61] Chen Y, He B, Lyu Z, et al (2026) Adaptive learned image compression with graph neural networks. arXiv preprint arXiv:260325316

[62] Chen Y, Zhou C, Li J, et al (2026) Next-frame decoding for ultra-low-bitrate image compression with video diffusion priors. arXiv preprint arXiv:260315129

[63] Chen YH, Weng YC, Kao CH, et al (2023) Transtic: Transferring transformer-based image compression from human perception to machine perception. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 23297–23307

[64] Chen Z, Sun H, Zhang L, et al (2024) Survey on visual signal coding and processing with generative models: Technologies, standards, and optimization. IEEE Journal on Emerging and Selected Topics in Circuits and Systems 14(2):149–171

[65] Chen Z, Fan J, Yu Z, et al (2025) Frequency-aware autoregressive modeling for efficient high-resolution image synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 17140–17149

[66] Cheng Z, Sun H, Takeuchi M, et al (2018) Deep convolutional autoencoder-based lossy image compression. In: 2018 Picture Coding Symposium (PCS), pp 253–257, https://doi.org/10.1109/PCS.2018.8456308

[67] Cheng Z, Sun H, Takeuchi M, et al (2019) Deep residual learning for image compression. In: Cvpr workshops, p 0

[68] Cheng Z, Sun H, Takeuchi M, et al (2020) Learned image compression with discretized gaussian mixture likelihoods and attention modules. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7939–7948

[69] Cheng Z, Fu T, Hu J, et al (2021) Perceptual image compression using relativistic average least squares gans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 1895–1900

[70] Choi H, Bajić IV (2018) Near-lossless deep feature compression for collaborative intelligence. In: 2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP), IEEE, pp 1–6

[71] Choi H, Bajić IV (2021) Latent-space scalability for multi-task collaborative intelligence. In: 2021 IEEE International Conference on Image Processing (ICIP), IEEE, pp 3562–3566

[72] Choi H, Bajić IV (2022) Scalable image coding for humans and machines. IEEE Transactions on Image Processing 31:2739–2754

[73] Choi Y, El-Khamy M, Lee J (2019) Variable rate deep image compression with a conditional autoencoder. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 3146–3154

[74] Chua LO, Lin T (1988) A neural network approach to transform image coding. International Journal of Circuit Theory and Applications 16(3):317–324

[75] Clarke R (1985) Transform coding of images. Academic Press Professional, Inc.

[76] Conceição R, Porto M, Peng WH, et al (2025) Cross-platform neural video coding: A case study. In: 2025 IEEE International Symposium on Circuits and Systems (ISCAS), IEEE, pp 1–5

[77] Cong W, Kong Y, Lu M, et al (2025) Integrating adaptive sampling for optimal learned video compression. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 1–5

[78] Cong W, Shi J, Lu M, et al (2026) Taming hierarchical image coding optimization: A spectral regularization perspective. In: The Fourteenth International Conference on Learning Representations

[79] Cong W, Shi J, Wang L, et al (2026) Reinforced rate control for neural video compression via inter-frame rate–distortion awareness. Proceedings of the AAAI Conference on Artificial Intelligence 40(17):14621–14629. https://doi.org/10.1609/aaai.v40i17.38480

[80] Cui Z, Wang J, Gao S, et al (2021) Asymmetric gained deep image compression with continuous rate adaptation. In: Proceedings of the IEEE/CVF Conference on Computer

Vision and Pattern Recognition, pp 10532–10541

[81] Daugman JG (1988) Complete discrete 2-d gabor transforms by neural networks for image analysis and compression. IEEE Transactions on acoustics, speech, and signal processing 36(7):1169–1179

[82] Di S, Liu J, Zhao K, et al (2025) A survey on error-bounded lossy compression for scientific datasets. ACM computing surveys 57(11):1–38

[83] Ding D, Ma Z, Chen D, et al (2021) Advances in video compression system using deep neural network: A review and case studies. Proceedings of the IEEE

[84] Ding K, Ma K, Wang S, et al (2020) Image quality assessment: Unifying structure and texture similarity. IEEE transactions on pattern analysis and machine intelligence 44(5):2567–2581

[85] Ding K, Ma K, Wang S, et al (2021) Comparison of full-reference image quality models for optimization of image processing systems. International Journal of Computer Vision 129(4):1258–1281

[86] Dong M, Sha H, Lu M, et al (2023) Ulic: Ultra lightweight image coder on wearable devices. In: International Forum on Digital TV and Wireless Multimedia Communications, Springer, pp 99–113

[87] Dong M, Lu M, Ma Z (2024) Accelerating block-level rate control for learned image compression. In: 2024 Data Compression Conference (DCC), pp 552–552, https://doi.org/10.1109/DCC58796.2024.00069

[88] Dony RD, Haykin S (1995) Neural network approaches to image compression. Proceedings of the IEEE 83(2):288–303

[89] Dosovitskiy A, Djolonga J (2020) You only train once: Loss-conditional training of deep networks. In: International conference on learning representations

[90] Duan X, Ma S, Liu H, et al (2024) Pku-aigi-500k: A neural compression benchmark and model for ai-generated images. IEEE Journal on Emerging and Selected Topics in Circuits and Systems 14(2):172–184

[91] Duan Y, Zhang Y, Tao X, et al (2019) Content-aware deep perceptual image compression. In: 2019 11th International Conference on Wireless Communications and Signal Processing (WCSP), IEEE, pp 1–6

[92] Duan Z, Lu M, Ma J, et al (2023) Qarv: Quantization-aware resnet vae for lossy image compression. IEEE Transactions on Pattern Analysis and Machine Intelligence

[93] Duan Z, Lu M, Ma Z, et al (2023) Lossy image compression with quantized hierarchical vaes. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 198–207

[94] Duan Z, Ma Z, Zhu F (2023) Unified architecture adaptation for compressed domain semantic inference. IEEE Transactions on Circuits and Systems for Video Technology 33(8):4108–4121

[95] Duan Z, Lu M, Yang J, et al (2024) Towards backward-compatible continual learning of image compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 25564–25573

[96] Dumas T, Roumy A, Guillemot C (2018) Autoencoder based image compression: can the learning be quantization independent? In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 1188–1192

[97] El-Nouby A, Muckley MJ, Ullrich K, et al (2022) Image compression with product quantized masked image modeling. arXiv preprint arXiv:221207372

[98] Esenlik S, Wu Y, Zhang Z, et al (2025) An overview of the jpeg ai learning-based image coding standard. IEEE Transactions on Circuits and Systems for Video Technology

[99] Esser P, Rombach R, Ommer B (2021) Taming transformers for high-resolution image synthesis. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 12868–12878, https://doi.org/10.1109/CVPR46437.2021.01268

[100] Feng D, Cheng Z, Wang S, et al (2025) Linear attention modeling for learned image compression. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 7623–7632

[101] Feng R, Jin X, Guo Z, et al (2022) Image coding for machines with omnipotent feature learning. In: European Conference on Computer Vision, Springer, pp 510–528

[102] Feng R, Guo Z, Li W, et al (2023) Nvtc: Nonlinear vector transform coding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 6101–6110

[103] Feng S, Zhu L, Zhang Y, et al (2025) C-ctx: Cubic-checkerboard context entropy model for learned image compression. IEEE Transactions on Multimedia

[104] Friedland G, Jia R, Wang J, et al (2020) On the impact of perceptual compression on deep learning. In: 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), IEEE, pp 219–224

[105] Fu H, Liang F, Lin J, et al (2023) Learned image compression with gaussian-laplacian-logistic mixture model and concatenated residual modules. IEEE Transactions on Image Processing 32:2063–2076

[106] Fu H, Liang F, Liang J, et al (2024) Fast and high-performance learned image compression with improved checkerboard context model, deformable residual module, and knowledge distillation. IEEE Transactions on Image Processing 33:4702–4715

[107] Fu H, Liang J, Fang Z, et al (2024) Weconvene: Learned image compression with wavelet-domain convolution and entropy model. In: European Conference on Computer Vision, Springer, pp 37–53

[108] Gadot U, Shocher A, Mannor S, et al (2025) Rl-rc-dot: A block-level rl agent for task-aware video compression. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 12533–12542

[109] Gao F, Deng X, Jing J, et al (2023) Extremely low bit-rate image compression via invertible image generation. IEEE Transactions on Circuits and Systems for Video Technology 34(8):6993–7004

[110] Gao G, You P, Pan R, et al (2021) Neural image compression via attentional multi-scale back projection and frequency decomposition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 14677–14686

[111] Gao S, Shi Y, Guo T, et al (2021) Perceptual learned image compression with continuous rate adaptation. 4th Challenge on Learned Image Compression 2(3)

[112] Gao W (2025) Implementations for ai-based image and video coding. In: AI-based Image and Video Coding: Methods, Standards, and Applications. Springer, p 271–303

[113] Gao W (2025) Standards for ai-based image and video coding. In: AI-based Image and Video Coding: Methods, Standards, and Applications. Springer, p 225–269

[114] Gao Y, Li S, Fu M, et al (2025) Approximately invertible neural network for learned image compression. IEEE Transactions on Image Processing

[115] Gao Y, Pan X, Li X, et al (2025) Why compress what you can generate? when gpt-4o generation ushers in image compression fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 371–381

[116] Ge Z, Ma S, Gao W, et al (2024) Nlic: Non-uniform quantization-based learned image compression. IEEE Transactions on Circuits and Systems for Video Technology 34(10):9647–9663. https://doi.org/10.1109/TCSVT.2024.3401872

[117] Gersho A (1978) Principles of quantization. IEEE Transactions on circuits and systems 25(7):427–436

[118] Gersho A, Gray RM (2012) Vector quantization and signal compression, vol 159. Springer Science & Business Media

[119] Gholami A, Kim S, Dong Z, et al (2022) A survey of quantization methods for efficient neural network inference. In: Low-power computer vision. Chapman and Hall/CRC, p 291–326

[120] Goodfellow I, Bengio Y, Courville A, et al (2016) Deep learning, vol 1. MIT press Cambridge

[121] Goodfellow I, Pouget-Abadie J, Mirza M, et al (2020) Generative adversarial networks. Communications of the ACM 63(11):139–144

[122] Gou J, Yu B, Maybank SJ, et al (2021) Knowledge distillation: A survey. International journal of computer vision 129(6):1789–1819

[123] Goyal VK (2001) Theoretical foundations of transform coding. IEEE Signal processing magazine 18(5):9–21

[124] Graves A (2011) Practical variational inference for neural networks. Advances in neural information processing systems 24

[125] Gray R (1984) Vector quantization. IEEE Assp Magazine 1(2):4–29

[126] Gu B, Chen H, Lu M, et al (2025) Adaptive rate control for deep video compression with rate-distortion prediction. In: 2025 Data Compression Conference (DCC), IEEE, pp 33–42

[127] Guo J, Ji Y, Chen Z, et al (2025) Oscar: One-step diffusion codec across multiple bit-rates. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems

[128] Guo P, Su W, Zhang X, et al (2024) Modeling the non-uniform retinal perception for viewport-dependent streaming of immersive video. Multimedia Systems 30(4):237

[129] Guo T, Wang J, Cui Z, et al (2020) Variable rate image compression with content adaptive optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp 122–123

[130] Guo Y, Chong Y, Pan S (2023) Hyperspectral image compression via cross-channel contrastive learning. IEEE Transactions on Geoscience and Remote Sensing 61:1–18

[131] Guo Z, Zhang Z, Feng R, et al (2021) Soft then hard: Rethinking the quantization in neural image compression. In: International Conference on Machine Learning, PMLR, pp 3920–3929

[132] Gupta V, Koren T, Singer Y (2018) Shampoo: Preconditioned stochastic tensor optimization. In: International Conference on Machine Learning, PMLR, pp 1842–1850

[133] Hamdi Y, Gündüz D (2023) The rate-distortion-perception trade-off with side information. In: 2023 IEEE International Symposium on Information Theory (ISIT), IEEE, pp 1056–1061

[134] Hamdi Y, Wagner AB, Gündüz D (2024) The rate-distortion-perception trade-off: The role of private randomness. In: 2024 IEEE International Symposium on Information Theory (ISIT), IEEE, pp 1083–1088

[135] Han C, Duan Y, Tao X, et al (2020) Toward variable-rate generative compression by reducing the channel redundancy. IEEE Transactions on Circuits and Systems for Video Technology 30(7):1789–1802

[136] Han M, Jiang S, Li S, et al (2024) Causal context adjustment loss for learned image compression. Advances in Neural Information Processing Systems 37:133231–133253

[137] He D, Zheng Y, Sun B, et al (2021) Checkerboard context model for efficient learned image compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 14771–14780

[138] He D, Yang Z, Chen Y, et al (2022) Post-training quantization for cross-platform learned image compression. arXiv preprint arXiv:220207513

[139] He D, Yang Z, Peng W, et al (2022) 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, pp 5718–5727

[140] He D, Yang Z, Yu H, et al (2022) Po-elic: Perception-oriented efficient learned image coding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 1764–1769

[141] He Z, Huang M, Luo L, et al (2024) Towards real-time practical image compression with lightweight attention. Expert Systems with Applications 252:124142

[142] Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. Advances in neural information processing systems 33:6840–6851

[143] Ho YH, Chan CC, Peng WH, et al (2021) Anfic: Image compression using augmented normalizing flows. IEEE Open Journal of Circuits and Systems 2:613–626

[144] Hojjat A, Haberer J, Landsiedel O (2023) Progdtd: Progressive learned image compression with double-tail-drop training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 1130–1139

[145] Hong W, Chen T, Lu M, et al (2020) Efficient neural image decoding via fixed-point inference. IEEE Transactions on Circuits and Systems for Video Technology 31(9):3618–3630

[146] Hossain MAF, Zhu F (2024) Structured pruning and quantization for learned image compression. In: 2024 IEEE International Conference on Image Processing (ICIP), IEEE, pp 3730–3736

[147] Hossain MAF, Duan Z, Zhu F (2024) Flexible mixed precision quantization for learne image compression. In: 2024 IEEE International Conference on Multimedia and Expo (ICME), IEEE, pp 1–8

[148] Hu EJ, Shen Y, Wallis P, et al (2022) Lora: Low-rank adaptation of large language models. Iclr 1(2):3

[149] Hu Y, Yang S, Yang W, et al (2020) Towards coding for human and machine vision: A scalable image coding approach. In: 2020 IEEE International Conference on Multimedia and Expo (ICME), IEEE, pp 1–6

[150] Hu Y, Yang W, Liu J (2020) Coarse-to-fine hyper-prior modeling for learned image compression. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 11013–11020

[151] Hu Y, Yang W, Ma Z, et al (2021) Learning end-to-end lossy image compression: A benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence 44(8):4194–4211

[152] Huang C, Liu H, Chen T, et al (2019) Extreme image coding via multiscale autoencoders with generative adversarial optimization. In: IEEE Visual Communications and Image Processing (VCIP), pp 1–4

[153] Huang CH, Wu JL (2024) Unveiling the future of human and machine coding: A survey of end-to-end learned image compression. Entropy 26(5):357

[154] Huang D, Gao F, Tao X, et al (2022) Toward semantic communications: Deep learning-based image semantic coding. IEEE Journal on Selected Areas in Communications 41(1):55–71

[155] Huang Y, Zhang J, Shan Z, et al (2024) Compression represents intelligence linearly. In: First Conference on Language Modeling, URL https://openreview.net/forum?id=SHMj84U5SH

[156] Huffman DA (1952) A method for the construction of minimum-redundancy codes. Proceedings of the IRE 40(9):1098–1101

[157] Islam K, Dang LM, Lee S, et al (2021) Image compression with recurrent neural network and generalized divisive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 1875–1879

[158] Ivins Jr WM (1969) Prints and visual communication, vol 131. mit Press

[159] Iwai S, Miyazaki T, Omachi S (2024) Controlling rate, distortion, and realism: Towards a single comprehensive neural image compression model. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 2900–2909

[160] Jamil S, Piran MJ, Rahman M, et al (2023) Learning-driven lossy image compression: A comprehensive survey. Engineering Applications of Artificial Intelligence 123:106361

[161] Jegou H, Douze M, Schmid C (2010) Product quantization for nearest neighbor search. IEEE transactions on pattern analysis and machine intelligence 33(1):117–128

[162] Jeon GW, Yu S, Lee JS (2023) Integer quantized learned image compression. In: 2023 IEEE International Conference on Image Processing (ICIP), IEEE, pp 2755–2759

[163] Jeon S, Choi KP, Park Y, et al (2023) Context-based trit-plane coding for progressive image compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 14348–14357

[164] Jeong DG, Gibson JD (1989) Lattice vector quantization for image coding. In: International Conference on Acoustics, p 439

[165] Jia C, Liu Z, Wang Y, et al (2019) Layered image compression using scalable auto-encoder. In: IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pp 431–436

[166] Jia P, Brand F, Yu D, et al (2025) Overview of variable rate coding in jpeg ai. IEEE Transactions on Circuits and Systems for Video Technology 35(9):9460–9474. https://doi.org/10.1109/TCSVT.2025.3552971

[167] Jia P, Brand F, Yu D, et al (2025) Overview of variable rate coding in jpeg ai. IEEE Transactions on Circuits and Systems for Video Technology

[168] Jia X, Wei X, Cao X, et al (2019) Comdefend: An efficient image compression model to defend adversarial examples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6084–6092

[169] Jia Z, Li B, Li J, et al (2025) Towards practical real-time neural video compression. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 12543–12552

[170] Jiang F, Tao W, Liu S, et al (2018) An end-to-end compression framework based on convolutional neural networks. IEEE Transactions on Circuits and Systems for Video Technology 28(10):3007–3018. https://doi.org/10.1109/TCSVT.2017.2734838

[171] Jiang J (1999) Image compression with neural networks–a survey. Signal processing: image Communication 14(9):737–760

[172] Jiang S, Yuan H, Li S, et al (2025) Fd-lscic: Frequency decomposition-based learned screen content image compression. IEEE Transactions on Broadcasting

[173] Jiang S, Long W, Han M, et al (2026) Differentiable vector quantization for rate-distortion optimization of generative image compression. arXiv preprint arXiv:260410546

[174] Jiang W, Yang J, Zhai Y, et al (2023) Mlic: Multi-reference entropy model for learned image compression. In: Proceedings of the 31st ACM International Conference on Multimedia, pp 7618–7627

[175] Jiang W, Yang J, Zhai Y, et al (2025) Mlic++: Linear complexity multi-reference entropy modeling for learned image compression. ACM Transactions on Multimedia Computing, Communications and Applications 21(5):1–25

[176] Jiang Z, Zhang X, Xu Y, et al (2020) Reinforcement learning based rate adaptation for 360-degree video streaming. IEEE Transactions on Broadcasting 67(2):409–423

[177] Johnston N, Vincent D, Minnen D, et al (2018) Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 4385–4393

[178] Kalmykov NI, Dibo R, Shen K, et al (2026) T-mla: A targeted multiscale log–exponential attack framework for neural image compression. Information Sciences p 123143

[179] Kao CH, Weng YC, Chen YH, et al (2023) Transformer-based variable-rate image compression with region-of-interest control. In: 2023 IEEE International Conference on Image Processing (ICIP), IEEE, pp 2960–2964

[180] Ke A, Zhang X, Chen T, et al (2025) Ultra lowrate image compression with semantic residual coding and compression-aware diffusion. In: Forty-second International Conference on Machine Learning

[181] Ke A, Zhang X, Chen T, et al (2025) Ultra lowrate image compression with semantic residual coding and compression-aware diffusion. In: Forty-second International Conference on Machine Learning

[182] Kerbl B, Kopanas G, Leimkühler T, et al (2023) 3d gaussian splatting for real-time radiance field rendering. ACM Trans Graph 42(4):139–1

[183] Khan MA, Baccour E, Chkirbene Z, et al (2022) A survey on mobile edge computing for video streaming: Opportunities and challenges. IEEe Access 10:120514–120550

[184] Kim H, Bauer M, Theis L, et al (2024) C3: High-performance and low-complexity neural compression from a single image or video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9347–9358

[185] Kim JH, Jang S, Choi JH, et al (2020) Instability of successive deep image compression. In: Proceedings of the 28th ACM International Conference on Multimedia, pp 247–255

[186] Kim JH, Heo B, Lee JS (2022) Joint global and local hierarchical priors for learned image compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5992–6001

[187] Kim MJ, Pertsch K, Karamcheti S, et al (2025) Openvla: An open-source vision-language-action model. In: Conference on Robot Learning, PMLR, pp 2679–2713

[188] Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980

[189] Kingma DP, Welling M (2013) Auto-encoding variational bayes. arXiv preprint arXiv:13126114

[190] Kirmemis O, Tekalp AM (2021) A practical approach for rate-distortion-perception analysis in learned image compression. In: 2021 Picture Coding Symposium (PCS), IEEE, pp 1–5

[191] Kobayashi H, Bahl LR (1974) Image data compression by predictive coding i: Prediction algorithms. IBM Journal of Research and Development 18(2):164–171

[192] Kong Y, Lu M, Ma Z (2024) Generative refinement for low bitrate image coding using vector quantized residual. IEEE Journal on Emerging and Selected Topics in Circuits and Systems 14(2):185–197

[193] Kovalev E, Bychkov G, Abud K, et al (2024) Exploring adversarial robustness of jpeg ai: methodology, comparison and new methods. arXiv preprint arXiv:241111795

[194] Koyuncu AB, Gao H, Boev A, et al (2022) Contextformer: A transformer with spatio-channel attention for context modeling in learned image compression. In: European conference on computer vision, Springer, pp 447–463

[195] Koyuncu AB, Jia P, Boev A, et al (2024) Efficient contextformer: Spatio-channel window attention for fast context modeling in learned image compression. IEEE Transactions on Circuits and Systems for Video Technology 34(8):7498–7511

[196] Koyuncu E, Solovyev T, Alshina E, et al (2022) Device interoperability for learned image compression with weights and activations quantization. In: 2022 Picture Coding Symposium (PCS), IEEE, pp 151–155

[197] Koyuncu E, Solovyev T, Sauer J, et al (2024) Quantized decoder in learned image compression for deterministic reconstruction. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 3985–3989

[198] Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25

[199] Kuang H, Yang W, Guo Z, et al (2026) Syntax-driven multi-realism image compression with consistency guided diffusion model. IEEE Transactions on Circuits and Systems for Video Technology

[200] Kurakin A, Goodfellow IJ, Bengio S (2018) Adversarial examples in the physical world. In: Artificial intelligence safety and security. Chapman and Hall/CRC, p 99–112

[201] Kurihara J, Sun H (2025) Efficient adversarial attack and training on learned image compression. In: 2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, pp 2453–2458

[202] Kuzmin A, Van Baalen M, Ren Y, et al (2022) Fp8 quantization: The power of the exponent. Advances in Neural Information Processing Systems 35:14651–14662

[203] Ladune T, Philippe P, Jaffuer P, et al (2026) Cool-chic 5.0: Faster encoding and inter-feature entropy modeling for overfitted image compression. arXiv preprint arXiv:260502726

[204] Laidlaw C, Singla S, Feizi S (2021) Perceptual adversarial robustness: Defense against unseen threat models. In: International Conference on Learning Representations (ICLR)

[205] Lam YH, Zare A, Cricri F, et al (2020) Efficient adaptation of neural network filter for video compression. In: Proceedings of the 28th ACM International Conference on Multimedia, pp 358–366

[206] Laparra V, Ballé J, Berardino A, et al (2016) Perceptual image quality assessment using a normalized laplacian pyramid. Electronic Imaging 28:1–6

[207] Le N, Zhang H, Cricri F, et al (2021) 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, pp 1590–1594

[208] Lee DT (2005) Jpeg 2000: Retrospective and new developments. Proceedings of the IEEE 93(1):32–41. https://doi.org/10.1109/JPROC.2004.839613

[209] Lee H, Kim M, Kim JH, et al (2024) Neural image compression with text-guided encoding for both pixel-level and perceptual fidelity. In: International Conference on Machine Learning, PMLR, pp 26715–26730

[210] Lee J, Cho S, Beack SK (2019) Context-adaptive entropy model for end-to-end optimized image compression. In: International Conference on Learning Representations

[211] Lee J, Jeong S, Kim M (2022) Selective compression learning of latent representations for variable-rate image compression. Advances in Neural Information Processing Systems 35:13146–13157

[212] Lee J, Jeong SY, Kim M (2026) Deephq: Learned hierarchical quantizer for progressive deep image coding. ACM Transactions on Multimedia Computing, Communications and Applications 22(1):1–24

[213] Lee JH, Jeon S, Choi KP, et al (2022) Dpict: Deep progressive image compression using trit-planes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 16113–16122

[214] Li C, Luo J, Dai W, et al (2020) Spatial-channel context-based entropy modeling for end-to-end optimized image compression. In: 2020 IEEE International Conference on Visual Communications and Image Processing (VCIP), IEEE, pp 222–225

[215] Li H, Li S, Dai W, et al (2024) Frequency-aware transformer for learned image compression. In: The Twelfth International Conference on Learning Representations

[216] Li H, Li S, Dai W, et al (2025) On disentangled training for nonlinear transform in learned image compression. In: The Thirteenth International Conference on Learning Representations

[217] Li J, Li B, Lu Y (2023) Neural video compression with diverse contexts. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver,

Canada, June 18-22, 2023

[218] Li M, Zuo W, Gu S, et al (2018) Learning convolutional networks for content-weighted image compression. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3214–3223

[219] Li M, Zuo W, Gu S, et al (2018) Learning convolutional networks for content-weighted image compression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

[220] Li M, Ma K, You J, et al (2020) Efficient and effective context-based convolutional entropy modeling for image compression. IEEE Transactions on Image Processing 29:5900–5911

[221] Li M, Wang Z, Shen L, et al (2022) Lightweight image compression based on deep learning. In: CAAI International Conference on Artificial Intelligence, Springer, pp 106–116

[222] Li S, Li H, Dai W, et al (2022) Learned progressive image compression with dead-zone quantizers. IEEE Transactions on Circuits and Systems for Video Technology 33(6):2962–2978

[223] Li S, Dai W, Kan N, et al (2025) Learnable non-uniform quantization with sampling-based optimization for variable-rate learned image compression. IEEE Transactions on Circuits and Systems for Video Technology

[224] Li X, Zhang J, Shi J, et al (2026) Yoda: Yet another one-step diffusion-based video compressor. arXiv preprint arXiv:260101141

[225] Li Y, Gong R, Tan X, et al (2021) Brecq: Pushing the limit of post-training quantization by block reconstruction. In: International Conference on Learning Representations

[226] Li Y, Chen X, Li J, et al (2022) Rate control for learned video compression. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 2829–2833

[227] Li Y, Zhang Z, Chen H, et al (2023) Mamba: Bringing multi-dimensional abr to webrtc. In: Proceedings of the 31st ACM International Conference on Multimedia, pp 9262–9270

[228] Li Y, Zhang H, Li L, et al (2025) Learned image compression with hierarchical progressive context modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 18834–18843

[229] Li Y, Zhang H, Li L, et al (2025) Scaling learned image compression models up to 1 billion. arXiv preprint arXiv:250809075

[230] Li Z, Huang C, Wang X, et al (2025) Lossless data compression by large models. Nature Machine Intelligence pp 1–6

[231] Li Z, Liao J, Tang C, et al (2025) Ustc-td: A test dataset and benchmark for image and video coding in 2020s. IEEE Transactions on Multimedia

[232] Li Z, Zhou Y, Wei H, et al (2025) Rdeic: Accelerating diffusion-based extreme image compression with relay residual diffusion. IEEE Transactions on Circuits and Systems for Video Technology

[233] Liang J, Liu M, Yao C, et al (2023) Sigvic: Spatial importance guided variable-rate image compression. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 1–5

[234] Liang J, He R, Tan T (2025) A comprehensive survey on test-time adaptation under distribution shifts. International Journal of Computer Vision 133(1):31–64

[235] Liang Z, Niu K, Wang C, et al (2025) Synonymous variational inference for perceptual image compression. In: International Conference on Machine Learning, PMLR, pp 37339–37369

[236] Liao L, Li S, Luo J, et al (2022) Efficient decoder for learned image compression via structured pruning. In: 2022 Data Compression Conference (DCC), IEEE, pp 464–464

[237] Lin J, Akbari M, Fu H, et al (2020) Variable-rate multi-frequency image compression using modulated generalized octave convolution. In: 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP), IEEE, pp 1–6

[238] Liu D, Li Y, Lin J, et al (2020) Deep learning-based video coding: A review and a case study. ACM Computing Surveys (CSUR) 53(1):1–35

[239] Liu D, Liu S, Ascenso J, et al (2024) Guest editorial special section on recent standardization efforts for learning-based visual data coding. IEEE Transactions on Circuits and Systems for Video Technology 34(5):3063–3066

[240] Liu H, Chen T, Shen Q, et al (2018) Deep image compression via end-to-end learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 2575–2578

[241] Liu H, Chen T, Shen Q, et al (2019) Practical stacked non-local attention modules for image compression. In: CVPR Workshops, p 0

[242] Liu H, Lu M, Chen Z, et al (2022) End-to-end neural video coding using a compound spatiotemporal representation. IEEE Transactions on Circuits and Systems for Video Technology 32(8):5650–5662

[243] Liu J, Liu D, Yang W, et al (2020) A comprehensive benchmark for single image compression artifact reduction. IEEE Transactions on image processing 29:7845–7860

[244] Liu J, Sun H, Katto J (2022) Improving multiple machine vision tasks in the compressed domain. In: 2022 26th International Conference on Pattern Recognition (ICPR), IEEE, pp 331–337

[245] Liu J, Sun H, Katto J (2023) Learned image compression with mixed transformer-cnn architectures. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14388–14397

[246] Liu K, Liu D, Li L, et al (2021) Semantics-to-signal scalable image compression with learned revertible representations. International Journal of Computer Vision 129(9):2605–2621

[247] Liu K, Wu D, Wang Y, et al (2022) Malice: Manipulation attacks on learned image compression. arXiv preprint arXiv:220513253

[248] Liu K, Wu D, Wu Y, et al (2023) Manipulation attacks on learned image compression. IEEE Transactions on Artificial Intelligence 5(6):3083–3097

[249] Liu L, Hu Z, Chen Z, et al (2023) Icmh-net: Neural image compression towards both machine vision and human vision. In: Proceedings of the 31st ACM International Conference on Multimedia, pp 8047–8056

[250] Liu Y, Chen X, Liu C, et al (2017) Delving into transferable adversarial examples and black-box attacks. In: International Conference on Learning Representations

[251] Liu Y, Gu H, Zhang A, et al (2024) Backdoor attack based on lossy image compression using discrete cosine transform. IEEE Access 12:196488–196497

[252] Liu Y, Huo S, Gu J, et al (2026) Neural b-frame video compression with bi-directional reference harmonization. Advances in Neural Information Processing Systems 38:154369–154394

[253] Liu Z, Lin Y, Cao Y, et al (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10012–10022

[254] Löhdefink J, Hüger F, Schlicht P, et al (2020) Scalar and vector quantization for learned image compression: A study on the effects of mse and gan loss in various spaces. In: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), IEEE, pp 1–8

[255] Lu G, Cai C, Zhang X, et al (2020) Content adaptive and error propagation aware deep video compression. In: European conference on computer vision, Springer, pp 456–472

[256] Lu L, Li Y, Wang Y, et al (2025) Hdcompression: hybrid-diffusion image compression for ultra-low bitrates. arXiv preprint arXiv:250207160

[257] Lu M, Chen F, Pu S, et al (2022) High-efficiency lossy image coding through adaptive neighborhood information aggregation. arXiv preprint arXiv:220411448

[258] Lu M, Guo P, Shi H, et al (2022) Transformer-based image compression. In: 2022 Data Compression Conference (DCC), IEEE, pp 469–469

[259] Lu M, Duan Z, Cong W, et al (2024) High-efficiency neural video compression via hierarchical predictive learning. arXiv preprint arXiv:241002598

[260] Lu M, Duan Z, Zhu F, et al (2024) Deep hierarchical video compression. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 8859–8867

[261] Lu Y, Zhu Y, Yang Y, et al (2021) Progressive neural image compression with nested quantization and latent ordering. In: 2021 IEEE International conference on image processing (ICIP), IEEE, pp 539–543

[262] Lu Y, Sun Q, Wang X, et al (2025) Bidirectional normalizing flow: From data to noise and back. arXiv preprint arXiv:251210953

[263] Luo X, Talebi H, Yang F, et al (2021) The rate-distortion-accuracy tradeoff: Jpeg case study. In: 2021 Data compression conference (DCC), IEEE, pp 354–354

[264] Lv Y, Xiang J, Zhang J, et al (2023) Dynamic low-rank instance adaptation for universal neural image compression. In: Proceedings of the 31st ACM International Conference on Multimedia, pp 632–642

[265] Ma H, Liu D, Xiong R, et al (2019) iwave: Cnn-based wavelet-like transform for image compression. IEEE Transactions on Multimedia 22(7):1667–1679

[266] Ma H, Liu D, Yan N, et al (2020) End-to-end optimized versatile image compression with wavelet-like transform. IEEE Transactions on Pattern Analysis and Machine Intelligence 44(3):1247–1263

[267] Ma S, Zhang X, Jia C, et al (2019) Image and video compression with neural networks: A review. IEEE Transactions on Circuits and Systems for Video Technology 30(6):1683–1698

[268] Ma Y, Zhai Y, Yang C, et al (2021) Variable rate roi image compression optimized for visual quality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and

Pattern Recognition, pp 1936–1940

[269] Madden J, Dorje L, Li X (2025) Bitstream collisions in neural image compression via adversarial perturbations. arXiv preprint arXiv:250319817

[270] Madry A, Makelov A, Schmidt L, et al (2018) Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations

[271] Marr D (2010) Vision: A computational investigation into the human representation and processing of visual information. MIT press

[272] Mentzer F, Agustsson E, Tschannen M, et al (2019) Practical full resolution learned lossless image compression. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 10629–10638

[273] Mentzer F, Toderici GD, Tschannen M, et al (2020) High-fidelity generative image compression. Advances in neural information processing systems 33:11913–11924

[274] Mentzer F, Agustson E, Tschannen M (2023) M2t: Masking transformers twice for faster decoding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 5340–5349

[275] Minnen D, Singh S (2020) Channel-wise autoregressive entropy models for learned image compression. In: 2020 IEEE International Conference on Image Processing (ICIP), IEEE, pp 3339–3343

[276] Minnen D, Ballé J, Toderici GD (2018) Joint autoregressive and hierarchical priors for learned image compression. Advances in neural information processing systems 31

[277] Mishra D, Singh SK, Singh RK (2022) Deep architectures for image compression: a critical review. Signal Processing 191:108346

[278] Mizuochi T (2006) Recent progress in forward error correction and its interplay with transmission impairments. IEEE Journal of Selected Topics in Quantum Electronics 12(4):544–554

[279] Muckley MJ, El-Nouby A, Ullrich K, et al (2023) Improving statistical fidelity for neural image compression with implicit local likelihood models. In: International Conference on Machine Learning, PMLR, pp 25426–25443

[280] Nagel M, Fournarakis M, Amjad RA, et al (2021) A white paper on neural network quantization. arXiv preprint arXiv:210608295

[281] Nakanishi KM, Maeda Si, Miyato T, et al (2018) Neural multi-scale image compression. In: Asian Conference on Computer Vision, Springer, pp 718–732

[282] Nie X, Li H, Jiang W, et al (2025) Badcomp: Backdoor attack against object detection using image compression operation. In: GLOBECOM 2025-2025 IEEE Global Communications Conference, IEEE, pp 1543–1548

[283] Niu X, Gündüz D, Bai B, et al (2023) Conditional rate-distortion-perception trade-off. In: 2023 IEEE International Symposium on Information Theory (ISIT), IEEE, pp 1068–1073

[284] Niu X, Bai B, Guo N, et al (2025) Rate–distortion–perception trade-off in information theory, generative models, and intelligent communications. Entropy 27(4):373

[285] Oquab M, Darcet T, Moutakanni T, et al (2024) Dinov2: Learning robust visual features without supervision. Transactions on Machine Learning Research Journal

[286] Pan X, Ding G, Chen Z, et al (2025) An efficient neural rate control for jpeg-ai. IEEE Transactions on Circuits and Systems for Video Technology pp 1–1. https://doi.org/

10.1109/TCSVT.2025.3614007

[287] Pan X, Ding G, Chen Z, et al (2025) An efficient neural rate control for jpeg-ai. IEEE Transactions on Circuits and Systems for Video Technology

[288] Pan Z, Zhang G, Peng B, et al (2024) Jnd-lic: Learned image compression via just noticeable difference for human visual perception. IEEE Transactions on Broadcasting 71(1):217–228

[289] Pang J, Lodhi MA, Ahn J, et al (2024) Towards reproducible learning-based compression. In: 2024 IEEE 26th International Workshop on Multimedia Signal Processing (MMSP), IEEE, pp 1–6

[290] Papernot N, McDaniel P, Goodfellow I (2016) Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:160507277

[291] Park C, Lee JC, Ko JH (2026) Single-step diffusion for image compression at ultra-low bitrates. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 6391–6400

[292] Parmar N, Vaswani A, Uszkoreit J, et al (2018) Image transformer. In: International conference on machine learning, PMLR, pp 4055–4064

[293] Patel Y, Appalaraju S, Manmatha R (2019) Deep perceptual compression. arXiv preprint arXiv:190708310

[294] Patel Y, Appalaraju S, Manmatha R (2021) Saliency driven perceptual image compression. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 227–236

[295] Pei Y, Liu Y, Ling N (2023) Mobilevit-gan: a generative model for low bitrate image coding. In: 2023 IEEE International Conference on Visual Communications and Image Processing (VCIP), IEEE, pp 1–5

[296] Pei Y, Liu Y, Ling N, et al (2023) An end-to-end deep generative network for low bitrate image coding. In: 2023 IEEE International Symposium on Circuits and Systems (ISCAS), IEEE, pp 1–5

[297] Peng C, Ye Y, Gao W (2026) Correcting quantization-induced gradient mismatch in neural image compression. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 8331–8339

[298] Pfaff J, Filippov A, Liu S, et al (2021) Intra prediction and mode coding in vvc. IEEE Transactions on Circuits and Systems for Video Technology 31(10):3834–3847. https://doi.org/10.1109/TCSVT.2021.3072430

[299] Ponomarenko N, Silvestri F, Egiazarian K, et al (2007) On between-coefficient contrast masking of dct basis functions. In: Proceedings of the third international workshop on video processing and quality metrics, Scottsdale USA

[300] Prakash A, Moran N, Garber S, et al (2017) Semantic perceptual image compression using deep convolution networks. In: 2017 Data Compression Conference (DCC), IEEE, pp 250–259

[301] Presta A, Tartaglione E, Fiandrotti A, et al (2025) Stanh: Parametric quantization for variable rate learned image compression. IEEE Transactions on Image Processing

[302] Presta A, Tartaglione E, Fiandrotti A, et al (2025) Efficient progressive image compression with variance-aware masking. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, pp 7692–7700

[303] Qian Y, Sun X, Lin M, et al (2022) Entroformer: A Transformer-based Entropy Model for Learned Image Compression. In: International Conference on Learning Representations

[304] Qin S, Wang J, Zhou Y, et al (2024) Mambavc: Learned visual compression with selective state spaces. arXiv preprint arXiv:240515413

[305] Qin S, Wang J, Zhou Y, et al (2025) Cassic: Towards content-adaptive state-space models for learned image compression. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 15727–15736

[306] Qin SY, Chen B, Huang YJ, et al (2026) Perceptual image compression with textual side information. Pattern Recognition 169:111848

[307] R¨aber S, Aczel T, Plesner A, et al (2025) Keep it real: Challenges in attacking compression-based adversarial purification. In: NeurIPS 2025 Workshop: Reliable ML from Unreliable Data

[308] Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:151106434

[309] Rafailov R, Sharma A, Mitchell E, et al (2023) Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems 36:53728–53741

[310] Relic L, Azevedo R, Gross M, et al (2024) Lossy image compression with foundation diffusion models. In: European Conference on Computer Vision, Springer, pp 303–319

[311] Rezende D, Mohamed S (2015) Variational inference with normalizing flows. In: International conference on machine learning, PMLR, pp 1530–1538

[312] Rippel O, Bourdev L (2017) Real-time adaptive image compression. In: International Conference on Machine Learning, pp 2922–2930

[313] van Rozendaal T, Huijben I, Cohen TS (2021) Overfitting for fun and profit: Instance-adaptive data compression. In: 9th International Conference on Learning Representations, ICLR 2021, ICLR

[314] Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. nature 323(6088):533–536

[315] Schulman J, Wolski F, Dhariwal P, et al (2017) Proximal policy optimization algorithms. arXiv preprint arXiv:170706347

[316] Senoo T, Girod B (1990) Vector quantization for entropy coding of image subbands. Vision and Modeling Group, Media Laboratory, Massachusetts Institute of ...

[317] Serra G, Stavrou PA, Kountouris M (2024) On the computation of the gaussian rate–distortion–perception function. IEEE Journal on Selected Areas in Information Theory 5:314–330

[318] Sha H, Dong M, Luo Q, et al (2025) Towards loss-resilient image coding for unstable satellite networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 12506–12514

[319] Shannon CE, Weaver W (1948) A mathematic theory of communication. Bell Syst Tech J 27(3):379–423

[320] Shao Y, Cao Q, Gündüz D (2024) A theory of semantic communication. IEEE Transactions on Mobile Computing 23(12):12211–12228

[321] Sheikh HR, Bovik AC (2006) Image information and visual quality. IEEE Transactions on image processing 15(2):430–444

[322] Shen Q, Cai J, Liu L, et al (2018) Codedvision: Towards joint image understanding and compression via end-to-end learning. In: Pacific Rim Conference on Multimedia, Springer, pp 3–14

[323] Shen S, Yue H, Yang J (2023) Dec-adapter: Exploring efficient decoder-side adapter for bridging screen content and natural image compression. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 12887–12896

[324] Sheng X, Zhu L, Zhang T, et al (2026) Cadc: Content adaptive diffusion-based generative image compression. arXiv preprint arXiv:260221591

[325] Shi J, Lu M, 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

[326] Shi J, Jiang M, Lu M, et al (2024) Hiner: neural representation for hyperspectral image. In: Proceedings of the 32nd ACM International Conference on Multimedia, pp 9837–9846

[327] Shi J, Chen Z, Li H, et al (2025) On quantizing neural representation for variable-rate video coding. In: The Thirteenth International Conference on Learning Representations

[328] Shi J, Zhang Q, Lu M, et al (2025) Compression as restoration: A unified implicit approach to self-supervised hyperspectral image representation. IEEE Journal of Selected Topics in Signal Processing

[329] Shi J, Lu M, Li X, et al (2026) Dit-ic: Aligned diffusion transformer for efficient image compression. arXiv preprint arXiv:260313162

[330] Shi Y, Zhang K, Wang J, et al (2022) Variable-rate image compression based on side information compensation and r-λ model rate control. IEEE Transactions on Circuits and Systems for Video Technology 33(7):3488–3501

[331] Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations

[332] Singh S, Abu-El-Haija S, Johnston N, et al (2020) End-to-end learning of compressible features. In: 2020 IEEE International Conference on Image Processing (ICIP), IEEE, pp 3349–3353

[333] Sitzmann V, Martel J, Bergman A, et al (2020) Implicit neural representations with periodic activation functions. Advances in neural information processing systems 33:7462–7473

[334] Sohl-Dickstein J, Weiss E, Maheswaranathan N, et al (2015) Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning, pmlr, pp 2256–2265

[335] Song M, Choi J, Han B (2021) Variable-rate deep image compression through spatially-adaptive feature transform. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 2380–2389

[336] Song M, Choi J, Han B (2024) A training-free defense framework for robust learned image compression. arXiv preprint arXiv:240111902

[337] Spadaro G, Presta A, Tartaglione E, et al (2024) Gabic: Graph-based attention block for image compression. In: 2024 IEEE International Conference on Image Processing (ICIP), IEEE, pp 1802–1808

[338] Standard I (2024) Ieee standard for neural network-based image coding. IEEE Std 185711-2024 pp 1–159. https://doi.org/10.1109/IEEESTD.2024.10810316

[339] Strümpler Y, Postels J, Yang R, et al (2022) Implicit neural representations for image compression. In: European conference on computer vision, Springer, pp 74–91

[340] Su R, Cheng Z, Sun H, et al (2020) Scalable learned image compression with a recurrent neural networks-based hyperprior. In: 2020 IEEE International Conference on Image Processing (ICIP), IEEE, pp 3369–3373

[341] SUI Y, LI Z, DING D, et al (2024) Transferable learned image compression-resistant adversarial perturbations. In: 2024 Data Compression Conference, Institute of Electrical and Electronics Engineers Inc., p 582

[342] Sullivan GJ, Wiegand T (1998) Rate-distortion optimization for video compression. IEEE Signal Processing Magazine 15(6):74–90

[343] Sun H, Cheng Z, Takeuchi M, et al (2020) End-to-end learned image compression with fixed point weight quantization. In: 2020 IEEE International Conference on Image Processing (ICIP), IEEE, pp 3359–3363

[344] Sun H, Yu L, Katto J (2021) End-to-end learned image compression with quantized weights and activations. arXiv preprint arXiv:211109348

[345] Sun H, Yu L, Katto J (2021) Learned image compression with fixed-point arithmetic. In: 2021 Picture Coding Symposium (PCS), IEEE, pp 1–5

[346] Sun H, Yu L, Katto J (2022) Q-lic: Quantizing learned image compression with channel splitting. IEEE Transactions on Circuits and Systems for Video Technology 35(4):3798–3811

[347] Sun Z, Tan Z, Sun X, et al (2021) Interpolation variable rate image compression. In: Proceedings of the 29th ACM international conference on multimedia, pp 5574–5582

[348] Sze V, Chen YH, Emer J, et al (2017) Hardware for machine learning: Challenges and opportunities. In: 2017 IEEE custom integrated circuits conference (CICC), IEEE, pp 1–8

[349] Tang Z, Wang H, Yi X, et al (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

[350] Tao L, Gao W, Li G, et al (2023) Adanic: towards practical neural image compression via dynamic transform routing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 16879–16888

[351] Tatwawadi K, Rahimzadeh P, Sun Z, et al (2026) What matters in practical learned image compression. 2605.05148

[352] Theis L, Shi W, Cunningham A, et al (2017) Lossy image compression with compressive autoencoders. arXiv preprint arXiv:170300395

[353] Tian T, Wang H, Zuo L, et al (2020) Just noticeable difference level prediction for perceptual image compression. IEEE Transactions on Broadcasting 66(3):690–700

[354] Toderici G, O'Malley SM, Hwang SJ, et al (2016) Variable rate image compression with recurrent neural networks. International Conference on Learning Representations

[355] Toderici G, Vincent D, Johnston N, et al (2017) Full resolution image compression with recurrent neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 5306–5314

[356] Tomczak J, Welling M (2018) Vae with a vampprior. In: International conference on artificial intelligence and statistics, PMLR, pp 1214–1223

[357] Tong K, Wu Y, Li Y, et al (2023) Qvrf: A quantization-error-aware variable rate framework for learned image compression. In: 2023 IEEE International Conference on Image Processing (ICIP), IEEE, pp 1310–1314

[358] Torfason R, Mentzer F, Agustsson E, et al (2018) Towards image understanding from deep compression without decoding. In: International Conference on Learning Representations

[359] Tsubota K, Akutsu H, Aizawa K (2023) Universal deep image compression via content-adaptive optimization with adapters. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 2529–2538

[360] Vahdat A, Kautz J (2020) Nvae: A deep hierarchical variational autoencoder. Advances in neural information processing systems 33:19667–19679

[361] Van Baalen M, Kuzmin A, Nair SS, et al (2023) Fp8 versus int8 for efficient deep learning inference. arXiv preprint arXiv:230317951

[362] Van Rozendaal T, Singhal T, Le H, et al (2024) Mobilenvc: Real-time 1080p neural video compression on a mobile device. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 4323–4333

[363] Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008

[364] Vyas N, Morwani D, Zhao R, et al (2025) Soap: Improving and stabilizing shampoo using adam for language modeling. In: The Thirteenth International Conference on Learning Representations

[365] Wallace GK (1991) The JPEG still picture compression standard. Communications of the ACM 34(4):30–44

[366] Wang F, Shen L, Teng Q, et al (2024) Dscic: Deep screen content image compression. IEEE Transactions on Circuits and Systems for Video Technology 34(11):11965–11979

[367] Wang GH, Li J, Li B, et al (2023) Evc: Towards real-time neural image compression with mask decay. arXiv preprint arXiv:230205071

[368] Wang J, Ling Q (2024) Fdnet: Frequency decomposition network for learned image compression. IEEE Transactions on Circuits and Systems for Video Technology 34(11):11241–11255

[369] Wang J, Duan Y, Tao X, et al (2021) Semantic perceptual image compression with a laplacian pyramid of convolutional networks. IEEE Transactions on Image Processing 30:4225–4237

[370] Wang J, Liu H, Feng D, et al (2024) Fp4-quantization: Lossless 4bit quantization for large language models. In: 2024 IEEE International Conference on Joint Cloud Computing (JCC), IEEE, pp 61–67

[371] Wang S, Wang Z, Wang S, et al (2021) End-to-end compression towards machine vision: Network architecture design and optimization. IEEE Open Journal of Circuits and Systems 2:675–685

[372] Wang S, Cheng Z, Feng D, et al (2024) Asymllic: Asymmetric lightweight learned image compression. In: 2024 IEEE International Conference on Visual Communications and Image Processing (VCIP), IEEE, pp 1–5

[373] Wang S, Dai J, Qin X, et al (2025) Resicomp: Loss-resilient image compression via dual-functional masked visual token modeling. IEEE Transactions on Circuits and Systems for Video Technology

[374] Wang S, Cheng Z, Feng D, et al (2026) Distilling complexity-scalable learned image compression models via neural architecture search. IEEE Transactions on Circuits and Systems for Video Technology

[375] Wang X, Chen T, Ma Z (2021) Subjective quality optimized efficient image compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 1911–1915

[376] Wang X, Lu M, Ma Z (2022) Block-level rate control for learnt image coding. In: 2022 Picture Coding Symposium (PCS), pp 157–161, https://doi.org/10.1109/PCS56426.2022.10018043

[377] Wang Y, Xiao M, Liu C, et al (2020) Modeling lost information in lossy image compression. arXiv preprint arXiv:200611999

[378] Wang Z (2004) Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 13(4):600–612

[379] Wang Z, Li Q (2010) Information content weighting for perceptual image quality assessment. IEEE Transactions on image processing 20(5):1185–1198

[380] Wang Z, Simoncelli EP, Bovik AC (2003) Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, pp 1398–1402

[381] Wei H, Zhou Y, Jia Y, et al (2025) A lightweight model for perceptual image compression via implicit priors. Neural Networks p 108279

[382] Wu C, Wu Q, Ma R, et al (2024) Continual cross-domain image compression via entropy prior guided knowledge distillation and scalable decoding. IEEE Transactions on Circuits and Systems for Video Technology 34(9):8080–8092

[383] Wu C, Wu Q, Wei H, et al (2025) On the adversarial robustness of learning-based image compression against rate-distortion attacks. IEEE Transactions on Multimedia

[384] Wyner A, Ziv J (1976) The rate-distortion function for source coding with side information at the decoder. IEEE Transactions on information Theory 22(1):1–10

[385] Xia Q, Liu H, Ma Z (2020) Object-based image coding: A learning-driven revisit. In: IEEE International Conference on Multimedia and Expo, pp 1–6

[386] Xie Y, Cheng KL, Chen Q (2021) Enhanced invertible encoding for learned image compression. In: Proceedings of the ACM International Conference on Multimedia

[387] Xu B, Chen H, Ma Z (2023) Karma: Adaptive video streaming via causal sequence modeling. In: Proceedings of the 31st ACM International Conference on Multimedia, pp 1527–1535

[388] Xu H, Wu X, Zhang X (2025) Multirate neural image compression with adaptive lattice vector quantization. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 7633–7642

[389] Xu T, Zhu Z, He D, et al (2024) Idempotence and perceptual image compression. In: The Twelfth International Conference on Learning Representations

[390] Xu T, Li J, Li B, et al (2025) Picd: Versatile perceptual image compression with diffusion rendering. In: Proceedings of the Computer Vision and Pattern Recognition

Conference, pp 28436–28445

[391] Xu Y, Wang Z, Wang J, et al (2025) Aguvis: Unified pure vision agents for autonomous gui interaction. In: International Conference on Machine Learning, PMLR, pp 69772–69805

[392] Xue M, Wang X, Sun S, et al (2023) Compression-resistant backdoor attack against deep neural networks. Applied Intelligence 53(17):20402–20417

[393] Xue N, Zhang Y (2023) Lambda-domain rate control for neural image compression. In: Proceedings of the 5th ACM International Conference on Multimedia in Asia, pp 1–7

[394] Xue N, Jia Z, Li J, et al (2025) One-step diffusion-based image compression with semantic distillation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems

[395] Yang C, Ma Y, Yang J, et al (2021) Graph-convolution network for image compression. In: 2021 IEEE International Conference on Image Processing (ICIP), IEEE, pp 2094–2098

[396] Yang F, Herranz L, Van De Weijer J, et al (2020) Variable rate deep image compression with modulated autoencoder. IEEE Signal Processing Letters 27:331–335

[397] Yang F, Herranz L, Cheng Y, et al (2021) Slimmable compressive autoencoders for practical neural image compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4998–5007

[398] Yang J, Wang X, Li Q, et al (2025) Subset-selection weight post-training quantization method for learned image compression task. IEEE Access 13:5145–5153

[399] Yang R, Mandt S (2023) Lossy image compression with conditional diffusion models. Advances in Neural Information Processing Systems 36:64971–64995

[400] Yang R, Liu D, Wu F, et al (2025) Learned image compression with efficient cross-platform entropy coding. IEEE Journal on Emerging and Selected Topics in Circuits and Systems 15(1):72–82

[401] Yang S, Hu Y, Yang W, et al (2021) Towards coding for human and machine vision: Scalable face image coding. IEEE Transactions on Multimedia 23:2957–2971

[402] Yang W, Huang H, Hu Y, et al (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

[403] Yang Y, Mandt S (2023) Computationally-efficient neural image compression with shallow decoders. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 530–540

[404] Yang Y, Bamler R, Mandt S (2020) Improving inference for neural image compression. Advances in Neural Information Processing Systems 33:573–584

[405] Yang Y, Gao R, Wang X, et al (2024) Mma-diffusion: Multimodal attack on diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7737–7746

[406] Yang Z, Wang Y, Xu C, et al (2020) Discernible image compression. In: Proceedings of the 28th ACM International Conference on Multimedia, pp 1561–1569

[407] Yao Z, Gholami A, Shen S, et al (2021) Adahessian: An adaptive second order optimizer for machine learning. In: proceedings of the AAAI conference on artificial intelligence, pp 10665–10673

[408] Ye F, Li L, Liu D (2025) Variable-rate learned image compression with integer-arithmetic-only inference. Journal of Visual Communication and Image Representation p 104634

[409] Ye F, Liu B, Li L, et al (2025) Rate-distortion-optimized deep preprocessing for jpeg compression. IEEE Transactions on Circuits and Systems for Video Technology

[410] Yu J, Mai S, Zhang P, et al (2025) Activation and weight distribution balancing for optimal post-training quantization in learned image compression. In: Proceedings of the 33rd ACM International Conference on Multimedia, pp 7959–7967

[411] Yu J, Mai S, Zhang P, et al (2025) Mixed-precision post-training quantization for learned image compression. IEEE Internet of Things Journal

[412] Yu Q, Zhang Z, Zhu R, et al (2025) Dapo: An open-source llm reinforcement learning system at scale. arXiv preprint arXiv:250314476

[413] Yu Y, Wang Y, Yang W, et al (2023) Backdoor attacks against deep image compression via adaptive frequency trigger. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 12250–12259

[414] Yu Y, Wang Y, Yang W, et al (2024) Robust and transferable backdoor attacks against deep image compression with selective frequency prior. IEEE Transactions on Pattern Analysis and Machine Intelligence 47(3):1674–1693

[415] Zamir R, Feder M (1992) On universal quantization by randomized uniform/lattice quantizers. IEEE Transactions on Information Theory 38(2):428–436

[416] Zeng F, Tang H, Shao Y, et al (2025) Mambaic: State space models for high-performance learned image compression. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 18041–18050

[417] Zhang C, Gao W (2024) Learned rate control for frame-level adaptive neural video compression via dynamic neural network. In: European Conference on Computer Vision, Springer, pp 239–255

[418] Zhang D, Li F, Liu M, et al (2023) Exploring resolution fields for scalable image compression with uncertainty guidance. IEEE Transactions on Circuits and Systems for Video Technology 34(4):2934–2948

[419] Zhang G, Li H, Cai Y, et al (2025) Progressive learned image transmission for semantic communication using hierarchical vae. IEEE Transactions on Cognitive Communications and Networking

[420] Zhang H, Li L, Liu D (2023) On uniform scalar quantization for learned image compression. arXiv preprint arXiv:230917051

[421] Zhang H, Koyuncu AB, Ahonen JI, et al (2025) Learned image codec with progressive multi-scale probability model for streaming in unreliable communication channels. In: 2025 IEEE International Workshop on Multimedia Signal Processing (MMSP), IEEE, pp 298–303

[422] Zhang H, Li L, Liu D (2025) Generalized gaussian model for learned image compression. IEEE Transactions on Image Processing

[423] Zhang H, Li Y, Li L, et al (2025) Learning switchable priors for neural image compression. IEEE Transactions on Circuits and Systems for Video Technology

[424] Zhang J, Pan X, Chen Z, et al (2025) Efficient jpeg-ai image coding for remote sensing semantic segmentation. IEEE Geoscience and Remote Sensing Letters

[425] Zhang L, Zhang L, Mou X, et al (2011) Fsim: A feature similarity index for image quality assessment. IEEE transactions on Image Processing 20(8):2378–2386

[426] ZHANG Q, MEI J, GUAN T, et al (2024) Recent advances in video coding for machines standard and technologies. ZTE Communications 22(1):62

[427] Zhang R, Isola P, Efros AA, et al (2018) The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 586–595

[428] Zhang S, Wang L, Mao X, et al (2022) Rate controllable learned image compression based on rfl model. In: 2022 IEEE International Conference on Visual Communications and Image Processing (VCIP), IEEE, pp 1–5

[429] Zhang T, Luo X, Li L, et al (2025) Stablecodec: Taming one-step diffusion for extreme image compression. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 17379–17389

[430] Zhang W, Li D, Min X, et al (2022) Perceptual attacks of no-reference image quality models with human-in-the-loop. Advances in Neural Information Processing Systems 35:2916–2929

[431] Zhang X, Wu X (2023) Lvqac: Lattice vector quantization coupled with spatially adaptive companding for efficient learned image compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 10239–10248

[432] Zhang X, Guo P, Lu M, et al (2024) All-in-one image coding for joint human-machine vision with multi-path aggregation. Advances in Neural Information Processing Systems 37:71465–71503

[433] Zhang X, Lu M, Chen Y, et al (2025) Perception-oriented latent coding for high-performance compressed domain semantic inference. In: IEEE ICME

[434] Zhang Y, Zhu F (2026) Leveraging second-order curvature for efficient learned image compression: Theory and empirical evidence. arXiv preprint arXiv:260120769

[435] Zhang Y, Lu G, Chen Y, et al (2023) Neural rate control for learned video compression. In: The Twelfth International Conference on Learning Representations

[436] Zhang Y, Zhu L, Jiang G, et al (2023) A survey on perceptually optimized video coding. ACM Computing Surveys 55(12):1–37

[437] Zhang Y, Duan Z, Lu M, et al (2024) Another way to the top: Exploit contextual clustering in learned image coding. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 9377–9386

[438] Zhang Y, Duan Z, Huang Y, et al (2025) Accelerating learned image compression through modeling neural training dynamics. Transactions on machine learning research 2025:4249

[439] Zhang Y, Duan Z, Huang Y, et al (2025) Balanced rate-distortion optimization in learned image compression. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 2428–2438

[440] Zhang Y, Huang Y, Zhu F (2026) Qarv++: An improved hierarchical vae for learned image compression. IEEE Transactions on Circuits and Systems for Video Technology

[441] Zhang Z, Chen Z, Lin J, et al (2019) Learned scalable image compression with bidirectional context disentanglement network. In: IEEE International Conference on Multimedia and Expo (ICME), pp 1438–1443, https://doi.org/10.1109/ICME.2019.00249

[442] Zhang Z, Esenlik S, Wu Y, et al (2023) End-to-end learning-based image compression with a decoupled framework. IEEE transactions on circuits and systems for video technology 34(5):3067–3081

[443] Zhang Z, Xia Y, Chen B, et al (2026) Autoregressive-based progressive coding for ultra-low bitrate image compression. In: The Fourteenth International Conference on Learning Representations

[444] Zheng J, Meister M (2025) The unbearable slowness of being: Why do we live at 10 bits/s? Neuron 113(2):192–204

[445] Zheng Y, Chen Y, Qian B, et al (2025) A review on edge large language models: Design, execution, and applications. ACM Computing Surveys 57(8):1–35

[446] Zhou J, Nakagawa A, Kato K, et al (2020) Variable rate image compression method with dead-zone quantizer. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition workshops, pp 162–163

[447] Zhou L, Sun Z, Wu X, et al (2019) End-to-end optimized image compression with attention mechanism. In: CVPR workshops, p 0

[448] Zhu M, Gupta S (2017) To prune, or not to prune: exploring the efficacy of pruning for model compression. arXiv preprint arXiv:171001878

[449] Zhu T, Sun H, Xiong X, et al (2024) Attack and defense analysis of learned image compression. arXiv preprint arXiv:240110345

[450] Zhu X, Song J, Gao L, et al (2022) Unified multivariate gaussian mixture for efficient neural image compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 17612–17621

[451] Zhu Y, Yang Y, Cohen T (2022) Transformer-based transform coding. In: International conference on learning representations

[452] Zou N, Zhang H, Cricri F, et al (2021) Adaptation and attention for neural video coding. In: 2021 IEEE International Symposium on Multimedia (ISM), IEEE, pp 240–244

[453] Zou R, Song C, Zhang Z (2022) 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, pp 17492–17501

Table 4 Evidence and deployment-readiness matrix for learned image coding.

Direction	Evidence base and common evaluation	Main deployment gap	Readiness
Fidelity-oriented R–D coding	Strong evidence on Kodak, Tecnick, CLIC, and JPEG AI test material using PSNR/MS-SSIM BD-rate against BPG (H.265/HEVC image-profile anchor) and VTM Intra (H.266/VVC image-profile anchor).	Training data, complexity reporting, entropy-coder runtime, and CTC are not yet uniformly specified across papers.	High
Transform and entropy modeling	Mature empirical base covering hyperpriors, autoregressive context models, checkerboard contexts, channel-wise contexts, and attention-based transforms.	Stronger probability models often introduce serial decoding, memory pressure, or hardware-unfriendly operators.	High–Medium
Variable-rate and rate control	Growing evidence from recurrent refinement, latent scaling, feature modulation, quality maps, and adaptive quantization.	Continuous rate control, region-of-interest (ROI) control, and rate-error guarantees remain less standardized than quantization-parameter control in conventional codecs.	Medium
Perceptual and generative compression	Increasing use of LPIPS, Frechet inception distance (FID), DISTS, and human preference studies for rate–distortion–perception analysis.	Hallucination, semantic fidelity, safety, and perceptual-metric reliability remain unresolved for high-stakes or archival use.	Medium–Low
Coding for machines and semantic communication	Promising results on classification, detection, segmentation, and task-aware latent-domain inference.	Evidence is task-, dataset-, and model-dependent; interoperability between human viewing and machine inference is still fragile.	Medium–Low
Integer inference and hardware deployment	Emerging evidence on mixed precision, 8-bit integer (INT8) inference, cross-platform consistency, latency, memory, and energy.	Platform-specific kernels, numerical mismatch, and incomplete hardware-aware training still limit broad reproducibility.	Medium
Robustness and transmission resilience	Initial studies cover distribution shift, adversarial perturbation, repeated coding, packet loss, and burst-error simulation.	No widely accepted CTC yet exists for robustness or lossy-network operation; robustness gains often trade against R–D efficiency.	Low–Medium
Standardization and interoperability	JPEG AI and IEEE 1857.11 provide common test material, reference software directions, and standardization momentum.	Bitstream conformance, decoder complexity, tooling maturity, and long-term ecosystem support are still evolving.	Medium

Table 5 Performance of TinyLIC under different complexity budgets and practical tool configurations. Fidelity-optimized models are evaluated by PSNR BD-rate using VTM Intra as the H.266/VVC image-profile anchor, while realism-optimized models are evaluated by LPIPS and FID BD-rate using MS-ILLM as the anchor. Lower values indicate better performance.

		Objective-Fidelity-Oriented				Perceptual-Realism-Oriented					
		PSNR BD-rate (%)				LPIPS BD-rate (%)				FID BD-rate (%)	
Methods	MACs/pixel	Kodak	CLIC	Tecnick	JPEG AI	Kodak	CLIC	Tecnick	JPEG AI	Tecnick	JPEG AI
Anchor											
VTM Intra 22.0 (H.266/VVC image-profile anchor) [35]		0.00	0.00	0.00	0.00	/	/	/	/	/	/
BPG (H.265/HEVC image-profile anchor) [28]		30.10	41.87	33.94	30.65	/	/	/	/	/	/
ELIC [139]	573k	-3.22	-3.89	-4.57	/	/	/	/	/	/	/
EVC-small [367]	52k	6.9	/	/	/	/	/	/	/	/	/
AsymLLIC [372]	130k	0.65	/	/	/	/	/	/	/	/	/
ShiftLIC-Small [26]	114k	15.56	11.18	10.67	/	/	/	/	/	/	/
CSLIC-3 [374]	125k	-0.94	3.48	2.75	/	/	/	/	/	/	/
CSLIC-1 [374]	77k	5.37	10.18	10.12	/	/	/	/	/	/	/
ShallowNTC [403]	20k	22.22	30.36	/	/	/	/	/	/	/	/
FastNIC [423]	10k	26.00	34.26	35.61	/	/	/	/	/	/	/
MS-ILLM [279]	600k	/	/	/	/	0.00	0.00	0.00	0.00	0.00	0.00
HiFiC [273]	600k	/	/	/	/	50.30	71.01	50.22	66.97	74.05	68.73
Fixed-Rate TinyLIC											
Nano	10k	14.78	14.68	12.76	13.86	30.92	97.47	83.49	87.98	91.57	40.62
Small	20k	5.43	4.63	2.73	6.33	20.96	78.66	60.03	63.81	76.21	41.86
Base	50k	-0.79	-1.76	-4.21	0.91	15.22	51.09	38.85	44.95	35.47	30.72
Variable-Rate TinyLIC											
VR-Adaptive Layer	50k	-1.50	-2.75	-5.02	0.29	14.75	49.17	41.30	40.02	28.35	26.47
VR-Dictionary	50k	-0.94	-2.21	-4.26	0.96	15.20	49.80	41.57	40.66	28.97	28.75
Variable-Rate (VR-Dictionary) + Consistent Decoding											
Quant-All (INT8)	50k	2.21	2.49	-0.22	4.74	17.16	60.02	43.20	51.52	32.59	28.60
Quant-Shared (INT8)	50k	-0.41	-1.45	-3.68	1.52	13.98	50.06	38.38	45.32	26.80	29.22
Variable-Rate (VR-Dictionary) + Consistent Decoding (Quant-Shared) + Progressive Coding / Attack-Resilient / Loss-Resilient											
PC-Variance Aware	50k	3.15	-1.04	-1.71	2.55	21.88	67.02	52.24	55.09	44.01	44.08
AR-Source	50k	0.75	-0.84	-2.69	2.67	15.45	61.89	42.41	51.76	37.63	34.43
AR-Noise	50k	-0.20	-1.36	-3.36	1.80	14.55	60.05	41.07	49.69	35.19	32.98
LR-GE (20%)	50k	9.78	8.52	6.22	8.73	33.85	81.91	72.88	58.90	56.35	41.32
LR-Random (20%)	50k	9.69	8.54	6.01	8.53	33.52	81.01	71.41	57.30	54.54	39.53

Table 6 Complexity and parameter distribution of TinyLIC variants.

TinyLIC	Complexity		Parameters (M)						
	Enc.	Dec.	g_a	g_s	h_a	h_s	CAR	LRP	Sum
Nano	11.92	9.99	0.81	0.53	0.63	0.63	0.42	0.15	3.16
Small	20.81	19.99	1.26	1.07	0.65	0.65	0.97	0.15	4.76
Base	53.00	50.76	3.45	3.05	1.38	1.38	1.17	0.15	10.57

Note: Complexity is measured in KMACs/pixel. Enc. and Dec. denote encoding and decoding complexity. g_a/g_s: main analysis/synthesis transforms; h_a/h_s: hyper analysis/synthesis transforms; CAR: channel autoregressive context model; LRP: latent residual prediction [275].