A Survey on Deep Learning–Based Image Compression: Methods, Efficiency, and Open Challenges

Ayalneh Bitew Wondmagegn, Ton That Tam Dinh, Thwe Thwe Win, Dongwook Won, and Sungrae Cho
School of Computer Science and Engineering, Chung-Ang University, Seoul 06974, South Korea
Email: {ayalneh, tttdinh, ttwin, dwwon}@uclab.re.kr, srcho@cau.ac.kr

***Abstract*—The rapid growth of image data in communication, storage, and multimedia applications has created an urgent demand for efficient and adaptive compression techniques. While conventional codecs rely on handcrafted transforms and fixed statistical models, deep learning–based image compression (DLIC) has emerged as a powerful alternative by jointly optimizing feature extraction, quantization, and entropy modeling in an end-to-end manner. This survey presents a structured overview of recent advances in DLIC, covering major architectural paradigms such as autoencoder-based models, hyperprior and autoregressive entropy models, attention- and transformer-based designs, as well as variable-rate and slimmable frameworks for adaptive bitrate and complexity control. Key challenges related to entropy modeling, computational efficiency, perception–distortion trade-offs, and practical deployment are discussed, along with a comparative analysis of representative methods and a brief review of standardization efforts and real-world deployment considerations. Finally, open research directions are outlined to guide the development of scalable and deployment-ready learned image compression systems.**

***Index Terms*—Learned image compression, autoencoder, entropy model, hyperprior, autoregressive context, mixture likelihood, perceptual quality, variable-rate, slimmable networks.**

I. Introduction

Image compression is a fundamental component of modern multimedia systems, enabling efficient storage, transmission, and retrieval of visual data across diverse applications. With the exponential growth of image-centric content on digital platforms and resource-constrained networks, the demand for advanced compression techniques has never been greater [1]. Compression techniques are broadly classified into two categories: lossy and lossless [2] [3]. Lossy compression achieves high compression ratios by discarding perceptually less significant information, which may slightly degrade visual quality [4]. Lossless compression, on the other hand, preserves the original image data perfectly but typically achieves lower compression efficiency [5]. Both approaches aim to minimize the number of bits required to represent an image while maintaining acceptable quality, thereby reducing storage requirements and improving transmission efficiency [6] [7].

Traditional codecs such as JPEG [8], JPEG2000 [9], and BPG [10] have been widely used for decades. These methods rely on handcrafted linear transforms (e.g., Discrete Cosine Transform and Discrete Wavelet Transform) followed by quantization and entropy coding to eliminate spatial redundancy. However, their fixed transform designs and manually tuned components often fail to fully exploit the complex spatial structures and semantic correlations present in natural images, especially under diverse visual conditions and application requirements.

The advent of deep learning (DL) has brought a paradigm shift to image compression by replacing traditional hand-engineered pipelines with end-to-end trainable models [11]–[16]. DLIC leverages powerful architectures such as convolutional neural networks, recurrent neural networks (RNNs), and transformers to jointly optimize the trade-off between bitrate and visual quality. These models learn nonlinear feature representations directly from data, exploit long-range spatial dependencies, and preserve perceptual details more effectively than conventional codecs, particularly at low bitrates.

Recent advancements in DLIC have evolved along several key directions. Autoencoder-based frameworks [11] [12] [17] learn compact latent representations that are quantized and entropy-coded. Advanced entropy models [14]–[16] [18] leverage hyperpriors, Gaussian mixtures, or autoregressive priors to improve probability estimation and compression efficiency. Additionally, attention mechanisms and transformer-based architectures [19] have enhanced contextual modeling by capturing long-range spatial relationships, further boosting compression performance.

Despite these advances, several challenges remain before DLIC can achieve widespread deployment. Deep models often entail high computational costs, limiting their use in real-time or resource-constrained environments. Moreover, adapting a single model to support multiple bitrates and diverse hardware platforms remains a nontrivial problem. Emerging solutions, including slimmable networks, modulation-based architectures, and progressive compression strategies, aim to overcome these limitations by enabling adaptive complexity, dynamic bitrate control, and hardware-aware deployment.

This survey aims to provide a comprehensive overview of the current state of DLIC. We trace the evolution from traditional codecs to neural compression frameworks, categorize existing methods based on their architectural principles and entropy modeling strategies, and examine emerging trends such as variable-rate coding, transformer-based compression, and perceptual optimization. Furthermore, we discuss benchmark results, analyze current limitations, and highlight open research challenges, offering future directions for advancing compression efficiency, scalability, and real-world applicability.

II. Architectural Taxonomy

Deep learning–based image compression (DLIC) has evolved through successive architectural innovations that progressively replace handcrafted components with learned representations, enhance entropy modeling, and improve rate–distortion (R–D) efficiency. To reduce redundancy and improve clarity, this section organizes existing approaches into eight complementary categories, each introduced once and referenced consistently throughout the paper: (1) plain autoencoders, (2) hyperprior models, (3) autoregressive and hierarchical priors, (4) perceptual and generative decoders, (5) variable-rate compression, (6) slimmable and switchable networks, (7) attention-based designs, and (8) hybrid architectures.

A. Plain Autoencoders (AE/CAE)

Early DLIC systems are based on compressive autoencoders (CAEs) consisting of an encoder, quantizer, and decoder trained end-to-end under a rate–distortion objective. Theis et al. [12] introduced one of the first fully trainable frameworks by approximating quantization with additive uniform noise. Ballé *et al.* [11] further improved performance by incorporating generalized divisive normalization (GDN/IGDN) to decorrelate latent features. While these models demonstrated the feasibility of learned compression, their fixed bottleneck design limited adaptability across operating bitrates.

B. Hyperprior Models

Hyperprior models enhance entropy estimation by transmitting auxiliary side information that conditions the latent distribution. The scale hyperprior introduced by Ballé *et al.* [14] predicts spatially adaptive variance parameters, significantly reducing redundancy without altering the reconstruction path. Hyperpriors remain a cornerstone of modern DLIC systems due to their favorable trade-off between compression efficiency and decoding parallelism.

C. Autoregressive and Hierarchical Priors

To further exploit spatial dependencies, autoregressive context models are often combined with hyperpriors. Minnen *et al.* [15] fused hierarchical priors with causal context to achieve state-of-the-art R–D performance. However, the sequential nature of autoregressive decoding introduces latency and limited parallelism, motivating later efforts toward parallel entropy models.

D. Perceptual and Generative Decoders

Beyond pixel-wise fidelity, perceptual compression integrates feature-level losses and adversarial training to enhance visual realism. Agustsson *et al.* [20] proposed selective generative compression, where GAN-based decoders synthesize perceptually plausible details at ultra-low bitrates. While effective for human perception, such approaches often trade PSNR for realism and require careful balancing of fidelity and perceptual quality.

E. Variable-Rate Deep Image Compression

Variable-rate compression enables a single model to operate at multiple rate–distortion points. Conditional autoencoders [21] and modulated architectures [17] adjust latent scaling or network activations based on a target rate parameter. These approaches eliminate the need for multiple bitrate-specific models, improving practicality while preserving compression efficiency.

F. Slimmable and Switchable Designs

Slimmable designs address hardware heterogeneity by allowing dynamic adjustment of network width using shared parameters and switchable normalization [22]. In compression, SlimCAE [23] demonstrates that width scaling can jointly control computational complexity and bitrate, enabling efficient deployment across edge devices without retraining. Unlike variable-rate models, which primarily target bitrate adaptability, slimmable models emphasize resource-aware inference.

G. Transformers and Attention

Attention-based architectures capture long-range spatial dependencies that convolutional models often miss. Transformer-based designs [24] tokenize latent features and model global relationships, yielding strong compression performance for structured imagery such as remote sensing data. The primary drawback remains increased computational cost.

H. Hybrid Architectures

Hybrid approaches integrate multiple principles to meet specialized requirements:

- Lossy + residual coding for near-lossless reconstruction [25];
- Frequency-decoupled coding using wavelet priors [26];
- Lossless coding guided by lossy priors [3].

These designs demonstrate how learned compression can be tailored to domain-specific constraints.

III. Methods and Efficiency Enhancements

Modern DLIC systems employ a unified learning-based pipeline consisting of encoding, quantization, entropy coding, and decoding. Advanced entropy models—including hyperpriors and hierarchical priors—significantly improve probability estimation and rate efficiency [14], [15]. To reduce computational overhead, lightweight architectures (e.g., depthwise convolutions, neural architecture search, and knowledge distillation) are increasingly adopted for edge and mobile deployment.

Recent work emphasizes scalability and latency reduction, including slimmable inference, progressive bitstreams, and parallel entropy decoding. These techniques mitigate the practical limitations of autoregressive models and enable real-time operation in bandwidth- and compute-constrained environments. Table I presents a comparative overview of

TABLE I
COMPARATIVE SUMMARY OF REPRESENTATIVE DLIC METHODS

Ref.	Core Architecture	Entropy Model	Variable Rate	Key Advantage	Complexity
[11]	CAE + GDN	Factorized Gaussian	No	End-to-end optimization	Low
[14]	CAE + Hyperprior	Scale Hyperprior	No	Improved RD efficiency	Medium
[15]	CAE	Hyperprior + AR	No	State-of-the-art RD	High (slow decode)
[16]	CAE + Attention	GMM + Context	No	Better spatial modeling	High
[21]	Conditional AE	Hyperprior	Yes	Single-model multi-rate	Medium
[23]	Slimmable CAE	Hyperprior	Yes	Hardware adaptability	Medium–Low
[20]	GAN-based Decoder	Learned Prior	No	Perceptual realism	High
[24]	Tokenized Encoder	Attention-based	Partial	Long-range modeling	Very High

representative DLIC approaches, detailing their core architectures, entropy modeling techniques (e.g., hyperpriors and autoregression), support for variable-rate operation, and typical performance–complexity trade-offs.

IV. TRAINING OBJECTIVES AND LOSSES

DLIC optimization is fundamentally guided by the rate–distortion trade-off:

$$\mathcal{L} = R + \lambda D, \tag{1}$$

where bitrate R is estimated via learned entropy models and distortion D measures reconstruction quality. Traditional distortion metrics include MSE and PSNR, while perceptually aligned metrics such as MS-SSIM and LPIPS better reflect human visual perception.

For perceptual applications, adversarial and feature-based losses are incorporated to enhance realism [20], [27]. Recent research further explores task-aware objectives, jointly optimizing compression and downstream vision tasks (e.g., detection or segmentation) to preserve task-relevant semantics [28]. Such hybrid objectives improve utility in real-world intelligent systems.

V. STANDARDIZATION EFFORTS

To bridge the gap between research and deployment, several standardization initiatives have emerged. JPEG AI aims to define interoperable standards for end-to-end neural image compression, including neural encoders, decoders, and entropy models. Similarly, MPEG Neural Network-based Representation (MPEG-NNR) investigates standardized neural representations for immersive and intelligent media. These efforts highlight growing interest in portability, decoder complexity constraints, and hardware compatibility, signaling a transition of DLIC toward real-world adoption.

VI. PRACTICAL DEPLOYMENT AND CASE STUDIES

In practical deployments, DLIC systems must balance compression efficiency with inference latency, memory footprint, and hardware constraints. Slimmable and variable-rate designs enable adaptive operation on heterogeneous platforms, including mobile devices and satellites. In remote sensing and satellite imaging, learned compression has demonstrated superior semantic preservation at low bitrates, making it suitable for bandwidth-limited downlinks. These studies emphasize that future DLIC research must jointly consider algorithmic performance and system-level constraints.

VII. OPEN CHALLENGES AND FUTURE DIRECTIONS

Despite remarkable progress, deep learning-based image compression (DLIC) still faces several critical challenges that must be addressed to enable its practical deployment and widespread adoption. One of the most fundamental issues is the perception–distortion trade-off. Conventional metrics such as PSNR and MS-SSIM effectively capture pixel-level fidelity but often fail to reflect human perceptual quality. Conversely, perceptual metrics like LPIPS or VGG-based losses enhance realism but may compromise objective accuracy. Future research must therefore focus on hybrid optimization strategies that jointly balance these conflicting objectives to achieve high-fidelity and perceptually pleasing reconstructions.

Another major challenge concerns computational efficiency and deployability. Current DLIC models are often too complex for deployment in resource-constrained environments such as edge devices, IoT platforms, or real-time video streaming systems. This limitation calls for lightweight model design techniques—such as pruning, quantization, knowledge distillation, and hardware-aware neural architecture search (NAS)—to significantly reduce inference cost without sacrificing compression performance. Additionally, most existing approaches are optimized for fixed bitrates, which limits their adaptability under dynamic network conditions. Future solutions should incorporate variable-rate and scalable compression mechanisms, such as conditional coding, modulation networks, or progressive multi-layer bitstreams, to deliver flexible and bandwidth-aware performance.

Beyond efficiency, entropy modeling and decoding speed remain pressing concerns. Although autoregressive and context-based entropy models achieve state-of-the-art performance, their sequential decoding introduces significant latency. Exploring parallelizable entropy modeling techniques, grouped decoding strategies, and transformer-based context modeling could dramatically accelerate inference and make DLIC more suitable for latency-sensitive applications. Finally, emerging demands in edge intelligence necessitate task-aware and semantic compression, where compression is jointly optimized with downstream tasks such as object detection or semantic segmentation to retain task-relevant features while minimizing data transmission.

Addressing these challenges will pave the way for the next generation of DLIC systems—solutions that are not only efficient, perceptually optimized, and scalable but also task-aware, privacy-preserving, and ready for deployment in diverse real-world multimedia applications.

VIII. Conclusion

This survey has presented a comprehensive overview of deep learning-based image compression, covering architectural evolution, entropy modeling, scalable designs, and training objectives. Beyond algorithmic performance, emerging trends in standardization and deployment reveal a growing emphasis on practicality, interoperability, and system-level efficiency. Continued advances in lightweight architectures, adaptive bitrate control, and standardized neural codecs will be essential for transitioning DLIC from laboratory benchmarks to widespread real-world adoption.

Acknowledgment

This work was supported in part by the IITP (Institute of Information & Communications Technology Planning & Evaluation) - ITRC (Information Technology Research Center) (IITP-2026-RS-2022-00156353, 50%) grant funded by the Korea government (Ministry of Science and ICT) and in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00453301).

References

[1] A. M. A. Brifcani and J. N. Al-Bamerny, "Image compression analysis using multistage vector quantization based on discrete wavelet transform," in *2010 International Conference on Methods and Models in Computer Science (ICM2CS-2010)*. IEEE, 2010, pp. 46–53.

[2] D. Venugopal, S. Mohan, and S. Raja, "An efficient block based lossless compression of medical images," *Optik*, vol. 127, no. 2, pp. 754–758, 2016.

[3] E. Gu, Y. Zhang, X. Wang, and X. Jiang, "Lossless compression framework using lossy prior for high-resolution remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

[4] H. M. Yasin and A. M. Abdulazeez, "Image compression based on deep learning: A review," *Asian Journal of Research in Computer Science*, vol. 8, no. 1, pp. 62–76, 2021.

[5] D. A. Zebari, D. Q. Zeebaree, A. M. Abdulazeez, H. Haron, and H. N. A. Hamed, "Improved threshold based and trainable fully automated segmentation for breast cancer boundary and pectoral muscle in mammogram images," *Ieee Access*, vol. 8, pp. 203 097–203 116, 2020.

[6] B. Rusyn, O. Lutsyk, Y. Lysak, A. Lukenyuk, and L. Pohreliuk, "Lossless image compression in the remote sensing applications," in *2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP)*. IEEE, 2016, pp. 195–198.

[7] J. Moon, "Covert communications in a compress-and-forward relay system," *ICT Express*, vol. 10, no. 2, pp. 412–417, 2024.

[8] H. ZainEldin, M. A. Elhosseini, and H. A. Ali, "Image compression algorithms in wireless multimedia sensor networks: A survey," *Ain Shams engineering journal*, vol. 6, no. 2, pp. 481–490, 2015.

[9] M. Li, W. Zuo, S. Gu, D. Zhao, and D. Zhang, "Learning convolutional networks for content-weighted image compression," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3214–3223.

[10] U. Albalawi, S. P. Mohanty, and E. Kougianos, "A hardware architecture for better portable graphics (bpg) compression encoder," in *2015 IEEE international Symposium on Nanoelectronic and information systems*. IEEE, 2015, pp. 291–296.

[11] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," *arXiv preprint arXiv:1611.01704*, 2016.

[12] L. Theis, W. Shi, A. Cunningham, and F. Huszár, "Lossy image compression with compressive autoencoders," *arXiv preprint arXiv:1703.00395*, 2017.

[13] W. J. Yun, S. Park, J. Kim, M. Shin, S. Jung, D. A. Mohaisen, and J.-H. Kim, "Cooperative multiagent deep reinforcement learning for reliable surveillance via autonomous multi-uav control," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 10, pp. 7086–7096, 2022.

[14] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," *arXiv preprint arXiv:1802.01436*, 2018.

[15] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," *Advances in neural information processing systems*, vol. 31, 2018.

[16] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7939–7948.

[17] F. Yang, L. Herranz, J. Van De Weijer, J. A. I. Guitián, A. M. López, and M. G. Mozerov, "Variable rate deep image compression with modulated autoencoder," *IEEE Signal Processing Letters*, vol. 27, pp. 331–335, 2020.

[18] S. Park, C. Park, and J. Kim, "Learning-based cooperative mobility control for autonomous drone-delivery," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 4, pp. 4870–4885, 2023.

[19] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote sensing*, vol. 12, no. 10, p. 1662, 2020.

[20] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. V. Gool, "Generative adversarial networks for extreme learned image compression," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 221–231.

[21] Y. Choi, M. El-Khamy, and J. Lee, "Variable rate deep image compression with a conditional autoencoder," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3146–3154.

[22] J. Yu, L. Yang, N. Xu, J. Yang, and T. Huang, "Slimmable neural networks," *arXiv preprint arXiv:1812.08928*, 2018.

[23] F. Yang, L. Herranz, Y. Cheng, and M. G. Mozerov, "Slimmable compressive autoencoders for practical neural image compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4998–5007.

[24] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[25] Y. Bai, X. Liu, K. Wang, X. Ji, X. Wu, and W. Gao, "Deep lossy plus residual coding for lossless and near-lossless image compression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 3577–3594, 2024.

[26] S. Xiang and Q. Liang, "Remote sensing image compression based on high-frequency and low-frequency components," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[27] F. Mentzer, G. D. Toderici, M. Tschannen, and E. Agustsson, "High-fidelity generative image compression," *Advances in neural information processing systems*, vol. 33, pp. 11 913–11 924, 2020.

[28] Y. Hu, R. Liu, X. Li, D. Chen, and Q. Hu, "Task-sequencing meta learning for intelligent few-shot fault diagnosis with limited data," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 6, pp. 3894–3904, 2021.