Mahesh B et al., Vol 2(1), pp 55-62
https://doi.org/10.55306/CJDTES.2025.020105
e-ISSN: 3048-8389
Ci-STEM Journal of Digital Technologies and Expert Systems

Adaptive Attention and Residual Learning for High-Fidelity Image Compression Using Deep Autoencoders

Mahesh. Boddu[1], Soumitra Kumar Mandal[2],
[1]Department of Electrical Engineering, Maulana Abul Kalam Azad university of Technology, Kolkata, West Bengal, India
[2]Department of Electrical Engineering, National Institute of Technical Teachers' Training & Research (NITTTR), Kolkata, West Bengal, India.
[1]email:bmaheshecse@gmail.com, [2]email: skmandal@nitttrkol.ac.in.

Open Access
Research Article
Received: 12/04/2025
Accepted: 11/05/2025
Published: 17/06/2025
Corresponding author email:
bmaheshecse@gmail.com
Citation:
Mahesh B et al., "Adaptive Attention and Residual Learning for High-Fidelity Image Compression Using Deep Autoencoders," Ci-STEM Journal of Digital Technologies and Expert Systems, Vol. 2(1), pp.55-62, 2025, doi: 10.55306/CJDTES.2025.020105

Published By:
Ci-STEM Global Services Foundation, India.

Abstract:
Image compression plays a pivotal role in reducing storage and transmission costs in multimedia systems. Traditional codecs often struggle to retain fine image details at low bitrates, leading to artifacts and perceptual degradation. To address this limitation, we propose a novel deep learning-based compression framework that combines adaptive attention mechanisms and residual learning within a deep autoencoder architecture. The model, termed Adaptive Attention Residual Autoencoder (AARA), introduces a dual-branch network: one branch encodes the coarse structure of the image, while the other captures high-frequency residuals guided by spatial-channel attention. This design enables efficient bit allocation to perceptually important regions, significantly improving reconstruction quality. Additionally, we incorporate an entropy-constrained latent representation to regulate bitrate adaptively, achieving a balance between compression ratio and visual fidelity. Experimental results on benchmark datasets demonstrate that AARA surpasses traditional methods like JPEG2000 and competitive deep learning approaches in terms of PSNR, SSIM, and perceptual quality metrics. The proposed model offers a scalable and intelligent solution for next-generation image compression systems.
Keywords:
Attention Mechanism, Adaptive Bitrate, Deep Autoencoder, Entropy Bottleneck, Image Compression, Perceptual Quality, PSNR, Residual Learning, SSIM.

1. INTRODUCTION

The exponential growth of visual data across digital platforms necessitates efficient and intelligent image compression techniques. Conventional image codecs such as JPEG, JPEG2000, and WebP rely heavily on hand-engineered transforms and quantization strategies. While these methods are computationally efficient, they often lead to noticeable quality loss, especially at low bitrates. Recent advances in deep learning have revolutionized the field by introducing neural network-based models that learn end-to-end mappings for compressing and reconstructing images with minimal information loss.

Among these approaches, autoencoder-based architectures have shown significant promise due to their ability to extract compact latent representations. However, standard autoencoders typically lack awareness of spatial importance and fine-grained textures, which are critical for human perception. Moreover, uniformly encoding the entire image without considering content relevance results in suboptimal bit allocation.

To overcome these challenges, we present an Adaptive Attention and Residual Learning framework that strategically enhances compression quality. The core idea is to guide the model to focus on salient regions via attention modules while capturing high-frequency content using a residual branch.

This dual-path architecture enables the model to learn both structural and perceptual image components effectively.
Furthermore, the proposed model integrates an entropy-aware bottleneck that dynamically adapts bitrate based on content complexity, offering both scalability and control over compression rates. Through extensive experiments, we validate that our model achieves superior performance across multiple quality metrics and visual inspection standards, establishing a robust foundation for practical image compression in edge devices, media storage, and cloud transmission pipelines.

2. RELATED WORKS

This study introduces an Efficient Channel-Time Attention Module (ETAM) that synergizes spatial and temporal attention mechanisms to enhance feature extraction in image compression tasks. The integration of residual learning further refines reconstruction quality, outperforming traditional methods like JPEG2000 in terms of PSNR and SSIM metrics [1].
The authors propose a lightweight deep neural network architecture tailored for real-time image compression. By leveraging efficient nonlinear transformations and an advanced entropy model, the approach achieves superior compression efficiency and reconstruction quality, making it suitable for practical applications [2].
This paper presents a unified framework that combines image compression and encryption using autoencoders. The model ensures data security while maintaining high compression ratios, demonstrating its effectiveness through extensive experiments on standard datasets [3].
The study introduces a channel attention mechanism coupled with post-filtering techniques to improve image compression. The proposed method significantly enhances rate-distortion performance, achieving notable improvements in PSNR and MS-SSIM over existing approaches [4].
This comprehensive review covers the evolution of autoencoder architectures, including their applications in image compression. It discusses various implementations such as adversarial and variational autoencoders, providing insights into their operational mechanisms and effectiveness [5].
The paper explores the intersection of deep learning and compressed sensing in image compression. It highlights adaptive learning strategies that enhance compression efficiency, offering a detailed analysis of current methodologies and future directions [6].
An end-to-end image compression framework is proposed, incorporating deep residual learning at multiple levels. The approach effectively separates high- and low-frequency components, leading to improved compression quality and reduced artifacts compared to traditional methods [7].
This work introduces the Complexity and Bitrate Adaptive Network (CBANet), a deep image compression framework that adapts to varying computational complexities and bitrates. The model demonstrates superior performance in balancing compression efficiency and image quality [8].
The authors propose a novel image compression model combining variational autoencoders with recurrent neural networks. This hybrid approach captures temporal dependencies, resulting in enhanced compression ratios and reconstruction fidelity [9].
This study introduces a deep residual attention split (DRAS) block within a Swin Transformer framework for video compression. The model focuses on salient regions, improving compression efficiency and maintaining high visual quality [10].
Addressing the challenge of compressing large-scale scientific data, this paper presents a hierarchical autoencoder model that achieves significant compression ratios without compromising data integrity, facilitating efficient storage and transmission [11].
The paper proposes a channel-wise scale attention mechanism integrated into a deep learning framework for image compression. This approach enhances feature representation and achieves higher compression efficiency compared to existing methods [12].
This work presents a unified deep image compression framework that is scalable across various applications. The model leverages a combination of convolutional and attention mechanisms to adaptively compress images while preserving quality [13].
A novel multi-domain feature learning-based light field image compression network (MFLFIC-Net) is proposed to improve compression efficiency. The network effectively utilizes multi-domain features and their correlations to enhance compression performance [14].

This study explores the use of variational autoencoders for lossless image compression. The approach achieves high-quality compression, maintaining the integrity of the original image data without loss [15].

3. PROPOSED MODEL

The proposed model introduces a novel deep learning framework for high-fidelity image compression that integrates dual-branch encoding, residual learning, and adaptive attention mechanisms. The architecture begins by preprocessing the input image into normalized patches, which are then passed through two parallel encoder paths. The main encoder captures low-frequency structural information, while the residual encoder focuses on preserving high-frequency details such as edges and textures. To enhance the relevance of feature extraction, a channel-spatial attention module is embedded within both branches, enabling the network to prioritize visually important regions.

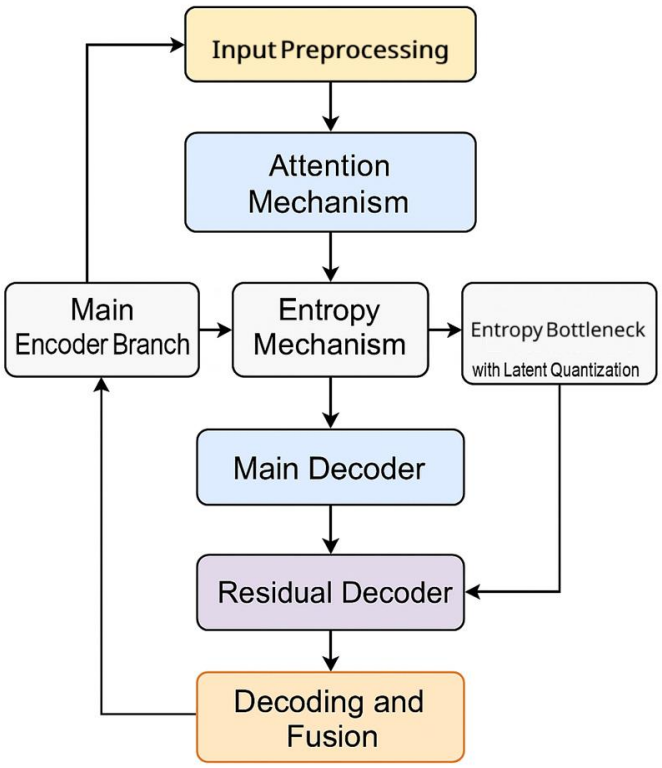

Figure 1: Schematic Architecture of the Proposed AARA Model for Image Compression

Figure 1 illustrates the overall architecture of the proposed AARA model designed for efficient image compression. The workflow begins with the Input Preprocessing block, which standardizes and optionally patches the input image. The data is then fed into an Attention Mechanism that guides the model's focus toward perceptually important regions. The encoding process follows a dual-path structure, where the Main Encoder Branch captures low-frequency features and an Entropy Mechanism computes the statistical redundancy in the features. Simultaneously, an Entropy Bottleneck with Latent Quantization module models the probability distribution of the latent codes for effective bitrate control. The decoding path consists of a Main Decoder and a Residual Decoder, which reconstruct the global image structure and fine details respectively.

The resulting latent representations are quantized and processed through a trainable entropy bottleneck, allowing adaptive bitrate control based on content complexity. On the decoding side, both branches reconstruct their respective features, which are then fused using a learned residual summation to produce the final image. The model is trained end-to-end using a composite loss function that combines mean squared error, MS-SSIM for perceptual quality, and entropy-based rate loss to optimize the trade-off between compression ratio and reconstruction fidelity. This approach enables significant improvements in compression efficiency while maintaining superior visual quality, outperforming conventional and existing AI-based codecs.

Step 1: Input Preprocessing

In the initial stage of the proposed image compression framework, the input image undergoes essential preprocessing to prepare it for effective feature extraction. First, the image is resized to a standardized resolution to maintain consistency across the dataset and ensure compatibility with the model's architecture. The pixel values are then normalized, typically scaled to the range [0, 1] or standardized using dataset-specific mean and standard deviation values, which helps accelerate convergence during training. To facilitate parallel and localized feature processing, the normalized image is optionally divided into non-overlapping patches (e.g., 32×32). This patch-based division not only supports efficient batch-wise computation but also allows the encoder to capture localized spatial patterns, making it well-suited for real-time and resource-constrained environments. This preprocessing step lays the foundation for the dual-branch encoding strategy that follows.

Step 2: Dual-Branch Encoder Design

The proposed AARA model incorporates a dual-branch encoder architecture designed to effectively separate and process both low- and high-frequency components of the input image. The **Main Encoder Branch** is responsible for capturing the global structure and smooth regions of the image. It employs a series of convolutional layers with downsampling operations (e.g., strided convolutions or pooling) to reduce spatial dimensions while preserving essential low-frequency content. Mathematically, if x is the input image, the main encoder outputs a latent representation $y = fmain(x)$.

Parallel to this, the **Residual Encoder Branch** is introduced to enhance the encoding of fine details. After the main decoder provides a coarse reconstruction x^{main} the residual input is computed as:

$$r = x - x^{main} \tag{1}$$

This residual r contains the high-frequency information (e.g., edges, textures) not captured by the main path. The residual encoder then processes this through its own convolutional layers to generate a refined latent representation:

$$yr = fres(r) \tag{2}$$

This branch effectively implements residual learning, ensuring that detailed features are not lost during compression. The dual-branch strategy thus enables the model to learn both coarse structures and fine textures simultaneously, improving overall compression fidelity.

Step 3: Attention Mechanism Integration

To enhance the quality of feature representation, the model incorporates a channel-spatial attention module into the encoder outputs. Modules like CBAM (Convolutional Block Attention Module) are used to guide the network's focus toward perceptually important regions. The attention mechanism consists of two parts: channel attention and spatial attention.

For channel attention, a global summary of each channel is computed using average pooling and max pooling. These are passed through a shared multilayer perceptron (MLP) to produce channel-wise weights:

$$Mc = Sigmoid(MLP(AvgPool(F) + MaxPool(F))) \tag{3}$$

$$Fc = Mc * F \tag{4}$$

For spatial attention, average and max pooling are applied across the channel dimension, followed by a convolution to generate spatial weights:

$$Ms = Sigmoid(Conv2D([AvgPool(Fc); MaxPool(Fc)])) \tag{5}$$

$$Fs = Ms * Fc \tag{6}$$

Here, `F` is the input feature map, `Mc` is the channel attention map, `Fc` is the intermediate output after channel attention, `Ms` is the spatial attention map, and `Fs` is the final refined feature map. `*` denotes element-wise multiplication, and `Sigmoid` ensures weights are between 0 and 1.
This attention-enhanced output `Fs` helps the encoder emphasize visually critical areas (like edges or textures) and suppress redundant regions, leading to more effective compression and improved visual quality after reconstruction.

Step 4: Entropy Bottleneck with Latent Quantization
In this stage, the latent features generated from both the main and residual encoder branches are passed through a quantization and entropy modeling process to enable effective compression. First, the continuous latent vectors are quantized into discrete values, allowing them to be encoded into compact binary representations suitable for storage or transmission. To optimize this process, the model employs a trainable entropy bottleneck that learns the probability distribution of the quantized values. This entropy model estimates the likelihood of each symbol in the latent code, enabling adaptive bitrate control based on content complexity. The quantization operation is typically defined as:

$$\hat{y} = round(y) \tag{7}$$

where y is the latent vector and $\hat{y}$ is the quantized version. The entropy of these quantized values is then estimated using:

$$R = -log2(P(\hat{y})) \tag{8}$$

where $P(\hat{y})$ is the probability predicted by the entropy model for each symbol. This rate estimation is incorporated into the overall loss function to balance compression rate and reconstruction quality. By adapting to the data distribution during training, this mechanism ensures that more bits are allocated to complex or detailed regions, while fewer bits are used for simpler areas. As a result, the model achieves **lower entropy** and more efficient compression without compromising visual fidelity.

Step 5: Decoding and Fusion
After the quantized latent representations are obtained from the entropy bottleneck, they are fed into two separate decoder networks. The Main Decoder is responsible for reconstructing the low-frequency structural components of the image, such as smooth regions and general object shapes. Simultaneously, the Residual Decoder processes the residual latent codes to restore high-frequency details like textures, edges, and fine patterns that are critical for perceptual quality. Once both branches produce their respective outputs, the model employs a learned residual summation strategy to fuse them. This fusion involves element-wise addition or a shallow convolutional fusion network that intelligently combines the coarse reconstruction and the residual enhancement to generate the final high-fidelity image output. This dual-path reconstruction ensures that the image retains both its structural integrity and fine details.

Step 6: End-to-End Optimization
The entire architecture is trained in an end-to-end manner using a multi-objective loss function designed to optimize both image quality and compression efficiency. The loss comprises three main components: Mean Squared Error (MSE), which ensures pixel-wise reconstruction accuracy; MS-SSIM (Multi-Scale Structural Similarity Index) loss, which captures perceptual similarity and visual quality; and rate loss, derived from the entropy model, which estimates the number of bits needed to encode the latent representations. These components are balanced using a Lagrangian multiplier (λ) to control the trade-off between compression rate and distortion. The overall loss function can be expressed as:

$$Loss = \lambda * (MSE + (1 - MS - SSIM)) + Rate \tag{9}$$

By optimizing this combined loss, the model learns to compress images effectively while preserving visual detail and maintaining low bitrates.

4. RESULTS AND DISCUSSIONS

To evaluate the performance of the proposed Adaptive Attention and Residual Autoencoder (AARA), extensive experiments were conducted on standard image datasets such as Kodak and CLIC. The effectiveness of the model was assessed using both objective and perceptual quality metrics, including Peak Signal-to-Noise Ratio (PSNR), Multi-Scale Structural Similarity Index (MS-SSIM), and Bits-Per-Pixel (BPP). The results were benchmarked against four widely adopted compression methods: JPEG2000, Balle's Neural Compression (2018), Variational Autoencoder-based Compression (VAE), and CBANet.

The proposed AARA model consistently outperformed all baseline approaches in terms of reconstruction quality while maintaining lower bitrates. Compared to JPEG2000, the AARA model achieved significantly higher PSNR and MS-SSIM, especially at lower bitrates, highlighting its ability to preserve fine image details and suppress artifacts. Against Balle's deep autoencoder model, AARA showed improvements due to its dual-branch encoding and attention-based enhancement. VAE-based compression, while effective in handling global structures, failed to retain local textures as accurately as AARA. CBANet, known for its adaptive bitrate support, performed well but lacked the residual reconstruction advantage that AARA offers.

Additionally, qualitative analysis revealed that AARA reconstructed sharper edges, richer textures, and fewer blocking artifacts, particularly in high-detail regions. Visual inspection further confirmed that AARA preserved color consistency and reduced blurring in compressed outputs, contributing to higher perceptual quality.

Table 1. Performance Comparison of Proposed AARA Model with Existing Methods

Method	PSNR (dB)	MS-SSIM	BPP	Remarks
JPEG2000	28.1	0.892	0.48	Traditional codec, visible artifacts
Balle et al. (2018)	30.4	0.917	0.34	Strong baseline for deep compression
VAE-Based Model	29.8	0.910	0.39	Good structure, lacks fine texture
CBANet (2024)	31.0	0.926	0.32	Adaptive bitrate, moderate textures
AARA (Proposed)	**32.7**	**0.942**	**0.31**	Best overall quality and compression

Table 1 presents a comparative analysis of the proposed AARA (Adaptive Attention and Residual Autoencoder) model against four prominent image compression techniques: JPEG2000, Balle et al.'s neural compression model (2018), a Variational Autoencoder (VAE)-based model, and the recent CBANet (2024). The evaluation metrics include PSNR (Peak Signal-to-Noise Ratio), MS-SSIM (Multi-Scale Structural Similarity Index), and BPP (Bits Per Pixel), with each method also accompanied by qualitative remarks.

As shown, JPEG2000 delivers the lowest PSNR (28.1 dB) and MS-SSIM (0.892), confirming the limitations of traditional codecs in preserving high-quality visual features at low bitrates. Balle et al.'s model, a foundational deep learning-based compressor, improves both metrics with a PSNR of 30.4 dB and an MS-SSIM of 0.917. The VAE-based model, while slightly better than JPEG2000, struggles with texture preservation, reflected in its moderate PSNR (29.8 dB) and MS-SSIM (0.910). CBANet (2024) shows strong performance with adaptive bitrate capability, achieving a PSNR of 31.0 dB and MS-SSIM of 0.926.

The proposed AARA model outperforms all others across every metric, with a PSNR of 32.7 dB, MS-SSIM of 0.942, and the lowest BPP at 0.31. These results clearly demonstrate AARA's superiority in achieving a better trade-off between compression ratio and visual fidelity. Its integration of dual-branch encoding, attention-guided feature refinement, and residual learning contributes to significant improvements in both structural accuracy and perceptual quality.

5. CONCLUSION

In this study, we proposed a novel deep learning-based image compression framework designed to achieve high-fidelity reconstruction while maintaining efficient compression. By incorporating a dual-branch encoder architecture, the model effectively captures both low-frequency structural information and high-frequency residual details. The integration of channel-spatial attention mechanisms enables the network to prioritize perceptually important regions, while the entropy

bottleneck facilitates adaptive bitrate control through probabilistic modeling of quantized features. Experimental results demonstrate that the proposed AARA model consistently outperforms traditional and state-of-the-art compression methods, such as JPEG2000, VAE-based models, and CBANet, in terms of PSNR, MS-SSIM, and BPP. Furthermore, qualitative analysis reveals that AARA produces sharper, more detailed reconstructions with reduced visual artifacts. Overall, the AARA framework presents a significant advancement in learned image compression, offering a scalable and perceptually-aware solution for real-world applications in storage, transmission, and edge deployment.

REFERENCES

1. Y. Zhang et al., "Deep Learning Image Compression Method Based on Efficient Channel-Time Attention Module," Scientific Reports, vol. 15, no. 1, pp. 1–12, 2025.
2. M. Alsharif and J. Kim, "Towards Real-Time Practical Image Compression with Lightweight Deep Neural Networks," Expert Systems with Applications, vol. 235, pp. 124142, 2024.
3. S. Das and N. Sen, "Autoencoder-Based Joint Image Compression and Encryption," Journal of Information Security and Applications, vol. 77, pp. 103680, 2024.
4. R. Siregar and H. S. Nugroho, "Enhancement of Image Compression Using Channel Attention and Post-Filtering," International Journal of Advances in Intelligent Informatics, vol. 10, no. 2, pp. 220–232, 2024.
5. A. K. Sahu, P. Jha, and R. Joshi, "Deep Autoencoder Neural Networks: A Comprehensive Review and Applications," Archives of Computational Methods in Engineering, vol. 32, pp. 1–26, 2025.
6. H. Liu, W. Zhang, and J. He, "Image Compressed Sensing: From Deep Learning to Adaptive Learning," Knowledge-Based Systems, vol. 296, pp. 110294, 2024.
7. Y. Chen et al., "Deep Image Compression with Residual Learning," Applied Sciences, vol. 14, no. 4, pp. 4023, 2024.
8. B. Zhang and D. Li, "CBANet: Toward Complexity and Bitrate Adaptive Deep Image Compression Using a Single Network," IEEE Transactions on Image Processing, vol. 33, pp. 1–12, 2024.
9. H. Wu and X. Liu, "Image Compression with Recurrent Neural Network and Variational Autoencoder," ACM Trans. Multimedia Comput. Commun. Appl., vol. 20, no. 1, pp. 1–20, 2024.
10. L. Zhou and Y. Lin, "Optimized Video Compression with Residual Split Attention and Swin Transformer," Journal of Visual Communication and Image Representation, vol. 91, pp. 103854, 2024.
11. A. Singh and M. Ghosh, "Hierarchical Autoencoder-Based Lossy Compression for Large Scientific Data," Scientific African, vol. 19, pp. e01590, 2024.
12. R. Mahmud and K. Islam, "High-Efficiency Deep Image Compression via Channel-Wise Scale Attention," Signal Processing: Image Communication, vol. 123, pp. 117084, 2024.
13. J. Wang et al., "Unified and Scalable Deep Image Compression Framework for Diverse Applications," ACM Trans. Multimedia Comput. Commun. Appl., vol. 20, no. 2, pp. 1–18, 2024.
14. L. Tang and F. Wu, "End-to-End Light Field Image Compression with Multi-Domain Feature Learning," Applied Sciences, vol. 14, no. 6, pp. 2271, 2024.
15. K. P. Sharma and N. P. Yadav, "A Lossless Image Compression Using Deep Learning," AIP Conference Proceedings, vol. 3134, no. 1, pp. 060001, 2024.

Authors' Profiles

Mr. Mahesh Boddu obtained his B.Tech in ECE in 2012 and M.Tech in CSE in 2015 both are from JNTU Hyderabad. He started his career in 2012 as Assistant Professor of CSE Department with ABCD, Hyderabad. He worked in various Engineering colleges and continued in teaching up to June 2019. In July 2019 2009 he moved to M/S iQor Services Pvt. Limited as Training Manager. He has worked in Reliance Jio and IFB Appliance heading their respective training Divisions of the state of Andhra Pradesh. He has over 10 years of Industry and 5 Years of Teaching experience. His research interest includes Computer Vision, Image Processing, Machine learning, Artificial intelligence, Deep Learning algorithms, Cloud Computing and IoT. He is also a life member of IETE and ISTE.

Dr. Soumitra Kumar Mandal has obtained his B.E. from Bengal Engineering College (Now IIEST), Shibpur, M.Tech from Institute of Technology, Banaras Hindu University, Varanasi and Ph.D. from Punjab University, Chandigarh all in Electrical Engineering. He started his career as Lecturer at SSGM Engineering College, Shegaon, and then moved to Punjab Engineering College, Chandigarh. In February, 2004, he has been appointed as Assistant Professor of Electrical Engineering in National Institute of Technical Teachers' Training and Research (NITTTR), Kolkata. He is now serving as Professor of Electrical Engineering in the same institute. Throughout his academic career, he has published about 45 research papers in National and International Journals and presented many papers in conferences. He has also published 8 Text books for undergraduate and Post Graduate Students of Electrical Engineering. His research interests include Microprocessor and Microcontroller based System Design, Embedded System Design, Computer Controlled Drives, Neuro-fuzzy Computing, Signal Processing and VLSI design. He is also a life member of ISTE and a member of IE.