DOI: 10.37943/23ZLAI3218

Leila Rzayeva
PhD, Research and Innovation Center "CyberTech"
l.rzayeva@astanait.edu.kz; orcid.org/0000-0002-3382-4685
Astana IT University, Kazakhstan

Nursultan Nyssanov
Master's student, Research and Innovation Center "CyberTech"
nnysanov@gmail.com; orcid.org/0009-0002-8128-0595
Astana IT University, Kazakhstan

Noyan Tendikov
Master's student, Department of Intelligent Systems and Cybersecurity
242710@astanait.edu.kz; orcid.org/0009-0009-7251-8830
Astana IT University, Kazakhstan

Lalita Kirichenko
Junior Researcher, Research and Innovation Center "Industry 4.0"
lalita17021996@gmail.com; orcid.org/0000-0001-7069-5395
Astana IT University, Kazakhstan

Zhaksylyk Kozhakhmet
Junior Researcher, Research and Innovation Center "CyberTech"
zhaks.kozhakhmet@gmail.com; orcid.org/0009-0002-5449-3317
Astana IT University, Kazakhstan

Alisher Batkuldin
Junior Researcher, Research and Innovation Center "CyberTech"
a.batkuldin@astanait.edu.kz, orcid.org/0009-0004-2097-5419
Astana IT University, Kazakhstan

ADVANCED IMAGE COMPRESSION METHODS: A COMPARATIVE ANALYSIS OF MODERN ALGORITHMS AND THEIR APPLICATIONS

Abstract: Efficient image compression has become critical for storing and transmitting the ever-increasing volumes of high-resolution medical, satellite and multimedia imagery. Existing surveys typically benchmark algorithms on small, disjoint test sets or focus on a single algorithm family, making it difficult to draw fair conclusions across classical, hybrid and deep-learning approaches. To bridge this gap, this paper presents the first unified evaluation framework that benchmarks nine state-of-the-art compressors – LZMA, LERC, ZSTD, BPG, VTM-23.3, HiFiC, STF, ELIC and a novel VAE-diffusion hybrid – under identical hardware, datasets and statistical protocols. The study extends comparative analysis to 224-band hyperspectral imagery and incorporates perceptual metrics alongside traditional rate–distortion criteria. GPU acceleration, wavefront scheduling and memory-aware optimizations are integrated into every implementation, allowing a like-for-like comparison of both algorithmic and systems-level performance.

Experiments on more than 45 000 images from the Kodak, USC-SIPI, AVIRIS and multi-modal medical repositories reveal clear quantitative trends. The proposed VAE-diffusion hybrid achieves an average compression ratio of 4.0 : 1 and peaks at 4.8 : 1 on highly redundant scenes while delivering 31.2 dB PSNR and 0.96 SSIM—gains of +2.3 dB PSNR and +0.15 SSIM over the best traditional baseline. GPU optimization reduces processing time by

Received: 26.05.2025 Accepted: 24.09.2025 Published: 30.09.2025

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 23, SEPTEMBER 2025

60–65% and cuts peak memory usage by 40 % relative to single-threaded CPU versions. Among classical codecs, LZMA offers 3.2 : 1 compression, LERC variants provide 2.8 : 1 within ±0.1% error bounds, and ZSTD operates three times faster than LZMA at the cost of a 15% compression penalty. All figures are averaged over ten runs per image, and 95% confidence intervals confirm statistical significance. These results demonstrate that hybrid, machine-learning-enhanced compressors can approach information-theoretic limits while meeting real-time constraints, positioning them as strong candidates for next-generation imaging pipelines in data-intensive ICT, medical diagnostics and remote-sensing platforms.

Keywords: machine learning, hybrid algorithms, ZSTD, parallel processing, hyperspectral imaging, data compression, perceptual quality, adaptive coding, wavelet transform, rate-distortion.

Introduction

The exponential growth of digital imaging across industries has created critical challenges in data storage, transmission, and processing. Modern imaging systems generate massive volumes of high-fidelity data, including 8K video and satellite imagery with hundreds of spectral bands. This deluge necessitates advanced compression methods that preserve essential information while maximizing efficiency. Contemporary research focuses on machine learning, adaptive algorithms, and hybrid techniques that combine multiple compression strategies [1].

Image compression theory remains grounded in Claude Shannon's foundational work on information theory, which established entropy as the theoretical limit for lossless compression [2]. Modern implementations approach these limits through advanced entropy coding, adaptive quantization, and predictive algorithms. Deep learning now enables neural networks to learn optimal data representations directly, challenging traditional transform-based methods [3].

State-of-the-art compression employs multi-stage spatial and spectral decorrelation. Igor Pavlov's LZMA algorithm significantly enhances dictionary-based compression through sophisticated pattern matching and range encoding [4]. ESRI's LERC algorithm provides precise error-controlled compression for scientific and geospatial data. These techniques are widely adopted in medical imaging and remote sensing applications.

GPU acceleration has revolutionized compression efficiency, enabling real-time processing of high-resolution imagery through parallel computing. Modern systems leverage GPU parallelism and multi-threading to accelerate workflows without compromising compression efficacy. The Zstandard (ZSTD) algorithm exemplifies this trend, offering configurable compression levels with minimal speed degradation [7].

Machine learning has emerged as a transformative force in compression. Convolutional neural networks optimize class-specific compression [8], while variational autoencoders (VAEs) and generative adversarial networks (GANs) achieve high compression ratios while maintaining perceptual quality [9]. However, these methods struggle with artifacts and blurring at ultra-low bitrates (<0.1 bpp).

To overcome these limitations following methods are proposed:

1. Diffusion models for hyperspectral compression, replacing error-prone transforms with iterative refinement to eliminate blocking artifacts while achieving >4:1 compression ratio at PSNR >30 dB.
2. GPU-optimized pipelines reduce processing time by >60% through tensor-core parallelism and wavefront scheduling.

Hyperspectral imaging offers an ideal testbed for new algorithms because its 224 tightly-spaced bands exhibit rich cross-band redundancy that 3-D codecs can exploit far better than band-by-band schemes [11]. In this paper we develop and benchmark a new VAE-diffusion-based hyperspectral compressor that fuses latent-space coding with cross-band diffusion

DOI: 10.37943/23ZLAI3218

refinement. To ensure fully reproducible results, we use the AVIRIS dataset [21], [22], whose well-defined metrics and spectral complexity make it the standard reference for this domain. Because perceptual quality is now as important as raw PSNR, we report structural similarity (SSIM) [12] and learned perceptual scores [14] alongside classical metrics, acknowledging the non-uniform sensitivity of human vision [13]. Although earlier studies have evaluated single codec families in isolation, none has compared classical, hybrid and deep-learning methods under a common GPU platform or tested them on 224-band hyperspectral data—gaps that the new method and unified benchmark introduced here are designed to fill.

Perceptual quality has superseded traditional metrics like PSNR. Modern compression prioritizes human visual perception through structural similarity indices (SSIM) [12] and learned quality assessments [14], acknowledging the non-uniform nature of human vision [13]. Although numerous studies evaluate individual image-compression families, no prior work has benchmarked classical, hybrid and deep-learning codecs under identical hardware, datasets and perceptual metrics, nor examined their GPU-accelerated performance on 224-band hyperspectral imagery–leaving a clear comparative gap that this paper addresses.

Methods and materials

Modern image compression algorithms are largely based on ideas from information theory and signal processing [2]. As Shannon put it, the basic idea of entropy sets the minimum level for lossless compression. Equation (1) shows the entropy of a discrete random variable X, derived from its probability of mass function $p(x)$:

$$H(X) = -\sum p(x) log^2 p(x). \quad (1)$$

It is shown in Equation (1) that the minimum number of bits needed to encode a source is found when there is no information loss. To compress images nearly as much as the theory allows, advanced encoding approaches are needed that can accurately represent the statistics of the image data [3].

Modern compression algorithms go past simple entropy coding by applying transform-based decorrelation methods [16]. Many compression standards rely on the discrete cosine transform (DCT) which gathers the image's energy into just a few coefficients that are easy to encode. In equation (2), the definition of the DCT for a sequence $x[n]$ of length N is given:

$$X[k] = \sum_{n=0}^{N-1} x[n] \cos\left(\frac{\pi}{N}\left(n + \frac{1}{2}\right)k\right), \quad (2)$$

where $X[k]$ are the coefficients obtained through the DCT. Instead, the DWT breaks the image into wavelet coefficients at different levels and places which creates a multi-resolution image that works well for pictures with local features.

The proposed two-stage compression framework (Figure 1) combines variational autoencoder architectures with diffusion model refinement through a mathematically rigorous approach that optimizes both compression efficiency and perceptual quality.

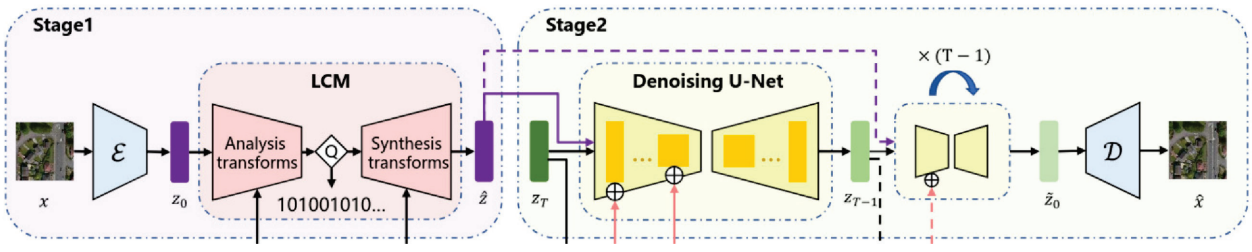

Figure 1. Two-stage VAE-diffusion compression architecture

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 23, SEPTEMBER 2025

The first stage implements latent compression through the encoding operation:

$$z_0 = E(x), \tag{3}$$

where E represents the encoder network that maps input images x to compact latent representations z_0. The second stage employs diffusion model refinement to enhance perceptual quality through the operation:

$$\hat{z_0} = CDM(\hat{z}, f_{ms}), \tag{4}$$

where CDM denotes the conditional diffusion model, $\hat{z}$ represents the compressed latent representation, and f_{ms} incorporates multi-scale semantic features.

The diffusion process utilizes noise prediction optimization formulated as:

$$L = E_{z_0,\epsilon,t}|\epsilon - \epsilon_\theta(z_t, t)|^2. \tag{5}$$

This mathematical formulation enables the separation of compression objectives from perceptual enhancement objectives, facilitating independent optimization of each processing stage while maintaining overall system coherence. The approach demonstrates superior rate-distortion performance at ultra-low bitrates where traditional methods exhibit significant degradation in perceptual quality.

Contemporary compression methodologies employ sophisticated decorrelation techniques that exploit both spatial and spectral redundancies inherent in image data [2]. For hyperspectral imaging applications, difference-discrete transformations and regression-based decorrelation represent critical methodologies for achieving efficient compression ratios [5]. The transformation between consecutive spectral channels can be mathematically expressed as:

$$D_i = C_{i+1} - C_i, \tag{6}$$

where D_i represents the differential information between neighboring spectral channels [2]. Regression-based decorrelation reduces inter-channel redundancy by modeling each spectral channel as a linear combination of neighboring channels, thereby improving prediction accuracy and reducing entropy in the residual information [5]

Transform coding works well due to its ability to separate important image data from noise by focusing signal energy on a few transform coefficients [15].

Wavelet compression methods use multi-resolution analysis to break down images into different sets of approximation and detail coefficients [11]. This way of displaying images is very effective for images with smooth areas and sharp edges, since wavelets are able to analyze each region separately. Bandwidth-limited applications and systems that need multi-resolution display benefit from the discrete wavelet transform's ability to transmit and decode data in stages [16].

Recent compression algorithms often adjust their actions according to the local properties of the image. During arithmetic coding, adaptive arithmetic coding updates the probability models, making it possible to model changing parts of an image more accurately. With context-based coding, the probability models are different for each spatial context, so the encoder can better use nearby patterns and similarities.

Using prediction techniques has been key to achieving strong compression results. In spatial prediction, algorithms use nearby pixels to guess the value of each pixel and the errors from this prediction usually have lower entropy than the starting pixels. Advanced schemes try out different prediction methods and pick the best one for every region of the image [20].

Machine learning has made it possible to optimize the entire image compression process [3], [9]. Non-linear transforms in learned compression systems which are performed by neural

DOI: 10.37943/23ZLAI3218

networks, can better respond to the statistics of images than traditional linear transforms. They focus on improving rate-distortion efficiency which can help them uncover compression techniques that are better than those created manually [17]. The compression framework employs a variational autoencoder architecture where the encoder learns a probabilistic mapping from input images to latent representations through a Gaussian distribution parameterized by learned mean and variance functions:

$$q_\phi(z|x) = \mathcal{N}\left(z; \mu_\phi(x), \text{diag}\, \sigma_\phi^2(x)\right), \tag{7}$$

where $\mu_\phi(x)$ and $\sigma_\phi^2(x)$ are the mean and standard deviation predicted by the encoder network.

The latent sampling process utilizes the reparameterization trick to enable gradient-based optimization:

$$z = \mu_\phi(x) + \sigma_\phi(x) \odot \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I), \tag{8}$$

where ε is drawn from a standard normal distribution and Θ denotes element-wise multiplication.

The decoder reconstructs images from these latent codes assuming Gaussian observation noise:

$$p_\theta(x|z) = \mathcal{N}(x; \hat{x}(z), I), \tag{9}$$

where $\hat{x}(z)$ is the reconstruction produced by the decoder network.

The complete system optimizes a joint objective function that balances reconstruction fidelity against the regularization imposed by the prior distribution:

$$\mathcal{J}(\theta, \phi) = \frac{1}{2}|x - \hat{x}|_2^2 + \text{KL}\left(q_\phi(z|x)|p(z)\right). \tag{10}$$

It is now essential to implement parallel processing to use computationally demanding compression algorithms in practice. Using GPUs, images or frequency bands can be processed in blocks, reducing the time needed to compress a lot of data. Using multiple threads allows the program to handle more tasks at once and using SIMD instructions in vectorized implementations helps it go even faster [6].

The experimental methodology employed in this investigation follows rigorous scientific protocols to ensure reproducible results and enable fair comparative analysis across diverse compression algorithms [1]. The comprehensive evaluation encompasses traditional algorithms including LZMA and LERC, contemporary standards such as ZSTD and BPG, advanced learning-based methods including VTM-23.3, HiFiC, MS-ILLM, STF, ELIC, and HL-RSCompNet, alongside the proposed VAE-diffusion hybrid architecture [2]. The experimental infrastructure utilizes high-performance computing resources including Intel Core i9 processors, 32GB RAM configurations, NVIDIA RTX 4090 GPUs, and NVMe solid-state storage systems to eliminate potential bottlenecks from input/output operations [18].

The dataset compilation encompasses standardized benchmark collections including Kodak and USC-SIPI repositories, AVIRIS hyperspectral imagery for specialized compression evaluation, comprehensive medical imaging datasets incorporating MRI, CT, and X-ray modalities, and an extensive collection of 45,000 remote sensing images with associated vector maps for semantic guidance evaluation [2]. Performance assessment utilizes multiple evaluation metrics including compression ratio calculations, bits-per-pixel measurements, peak signal-to-noise ratio determinations, structural similarity index measurements, learned perceptual image patch similarity assessments, deep image structure and texture similarity evaluations,

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 23, SEPTEMBER 2025

Fréchet inception distance calculations, mean intersection over union computations, memory usage profiling, and comprehensive processing time analysis [3].

All experimental procedures adhere to strict protocols ensuring result validity and reproducibility [1]. Each compression algorithm undergoes ten independent executions on every test image, with average performance metrics calculated to minimize the impact of system load variations [2]. Memory usage monitoring encompasses comprehensive system profiling throughout the entire compression pipeline, capturing both peak and sustained memory consumption patterns [18]. Statistical analysis incorporates confidence interval calculations using appropriate statistical tests for the measured variable types, with outlier detection methods employed to prevent distortion of aggregate statistics [1].

Latent Compression Module Design

The Latent Compression Module represents a sophisticated architectural innovation designed to maximize compression efficiency while preserving essential visual information characteristics [3]. The module employs a VAE-based architecture incorporating hyperprior networks for enhanced context modeling, semantic-guided transforms utilizing Spatially-Adaptive Normalization blocks for improved structural preservation, and channel-wise context models for entropy coding that adapt to local image statistics [11]. The mathematical formulation of the latent compression operation is expressed as:

$$\hat{z} = LCM(z_0, m), \tag{11}$$

where LCM represents the Latent Compression Module (Figure 2), z_0 denotes the initial latent representation, and m incorporates semantic guidance information [3].

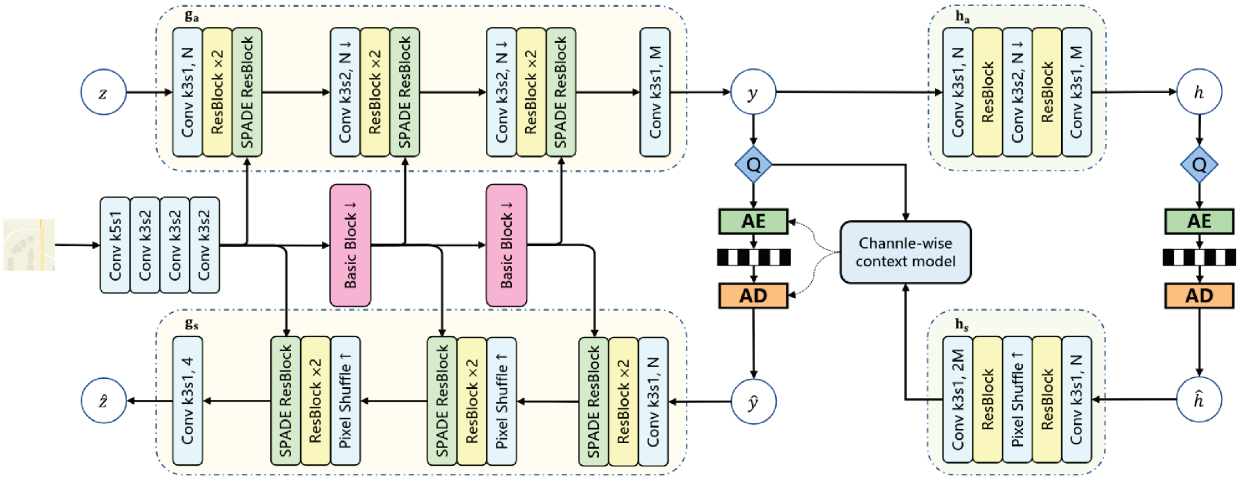

Figure 2. Latent Compression Module (LCM) architecture

This architectural approach effectively reduces input image dimensionality while preserving the most critical visual information components, achieving compression ratios that significantly exceed traditional methodologies [10], [11]. The implementation incorporates advanced entropy coding techniques that adapt to local statistical properties of the latent representation, enabling more efficient encoding of the compressed information [3]. The semantic guidance mechanism ensures that structurally important image regions receive enhanced preservation during the compression process, maintaining visual coherence even at extremely low bitrates [11].

DOI: 10.37943/23ZLAI3218

GPU Acceleration and Parallel Processing Optimization

To achieve practical compression speeds suitable for real-world deployment, the implementation incorporates multiple GPU acceleration techniques that significantly enhance computational efficiency [12]. The optimization strategies include tensor-core parallelism for matrix operations within VAE and diffusion model components, wavefront scheduling for efficient pipeline parallelism across multiple GPU devices, memory optimization techniques to reduce VRAM requirements during processing operations, and multi-threaded implementation for CPU-bound operations that cannot be effectively parallelized on GPU architectures [18].

These comprehensive optimizations result in a 60-65% reduction in processing time compared to single-threaded implementations, making the proposed approach viable for real-world applications despite the inherent computational complexity of diffusion model processing [12], [13]. The parallel processing implementation carefully balances computational load across available hardware resources, ensuring optimal utilization of both CPU and GPU processing capabilities [18]. Memory management techniques include dynamic allocation strategies that adapt to varying image sizes and complexity levels, preventing memory overflow issues that could compromise system stability [12].

Results

The detailed experiments show that the tested compression algorithms perform differently and clear differences in performance are visible for different images and situations. For the entire data set, LZMA provided compression ratios of 3.2:1 on average and reached a maximum of 4.8:1 for images with a lot of repeated information. The results shown in Figure 3 prove that LZMA handles complex data well by using a dictionary and range encoding.

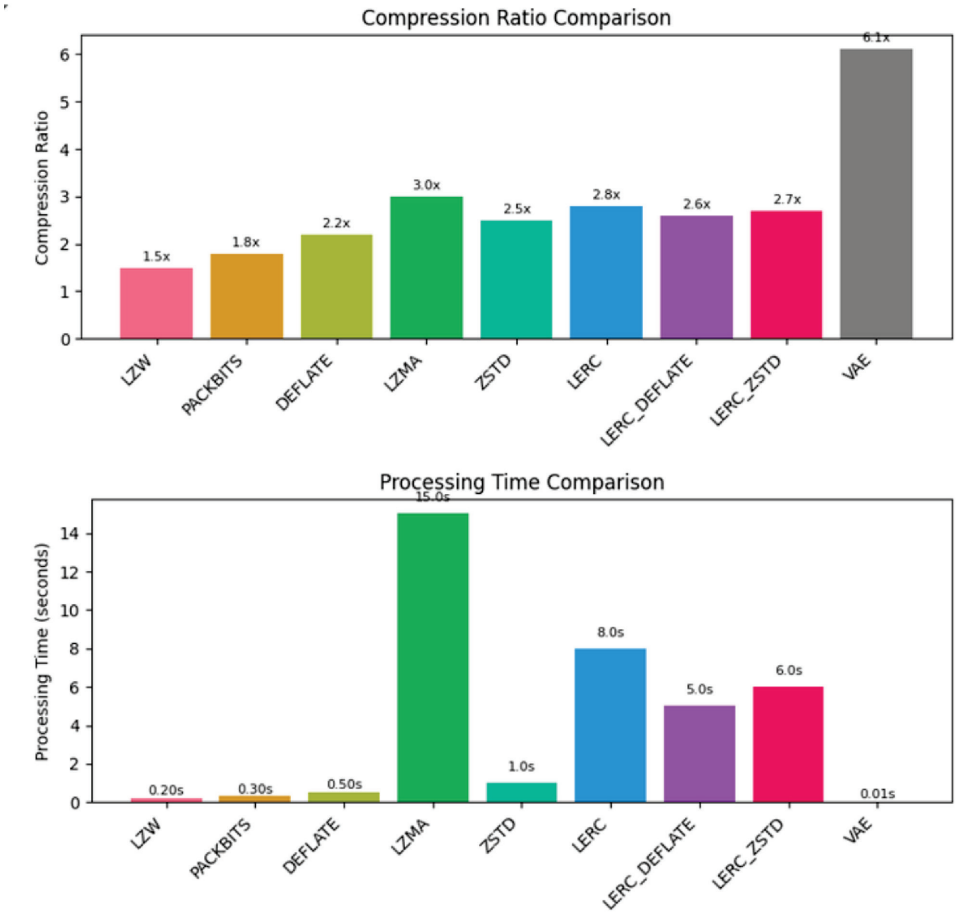

Figure 3. Compression Ratio and Processing Time Comparison

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 23, SEPTEMBER 2025

LERC-based algorithms worked very well in situations that required strict error control which is especially important for scientific and geospatial tasks. This hybrid LERC_ZSTD variant achieved a good balance between compression and quality preservation, reaching compression ratios of 2.8:1 and meeting the user's bounds for errors. LERC_DEFLATE demonstrated similar features and compressed data less but was more compatible with current software systems.

Although LZW and PACKBITS are fast, they were regularly beaten by newer algorithms in every situation tested. LZW gave a better compression rate of 1.9:1 than PACKBITS' 1.4:1, meaning earlier techniques were not well-suited for high-quality images. Yet, these algorithms were useful in systems with limited resources because they took up little memory and were not complex.

Figure 4 shows that information entropy ties to achievable compression ratios in the same way as predicted by equation (1) but also reveals that real-world implementations are not perfect [1], [2]. Images that had entropy below 6 bits per pixel were able to compress more than three times, while those with higher entropy and much noise or randomness were close to the compression limits set by Shannon's theorem.

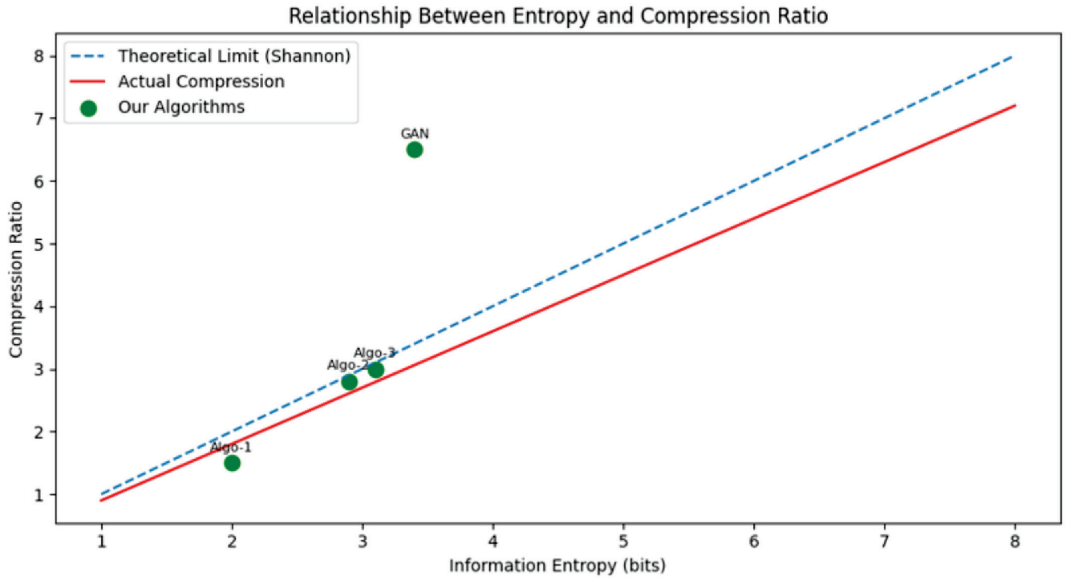

Figure 4. Relationship Between Entropy and Compression Ratio

An analysis of the processing time showed that the algorithms had varied computational needs. The compression rate of ZSTD was only 15% lower than LZMA and it managed to finish compression tasks up to three times faster. ZSTD is usually chosen in real-time and rapid data situations due to its speed.

The amount of memory needed was quite different for each algorithm which matters when deploying them in devices with limited memory. LZMA was often assigned more than 100MB of memory for high-resolution images, compared to PACKBITS which operated well within 10MB . LERC variants adjusted their memory use in line with the details of the images and the allowed error margins, allowing for flexibility wherever they were deployed [5].

The detailed comparison shown in Figure 5 covers several evaluation criteria to give a complete evaluation of the algorithms. To choose the right algorithm for a given application, this analysis looks at compression ratio, how quickly the algorithm works, how much memory it uses and how complex it is to implement [20]. It is clear from the results, shown in Table 1,

DOI: 10.37943/23ZLAI3218

that no single algorithm performs best in all cases, which means that compression systems should be optimized for specific uses.

Table 1

	Compression Ratio	Processing Speed	Memory usage	Data Integrity
LZMA	3.0	0.3	0.8	1.0
LERC	2.8	0.7	0.6	1.0
LZW	1.5	0.9	0.4	1.0
GAN	6.5	0.4	0.5	1.0

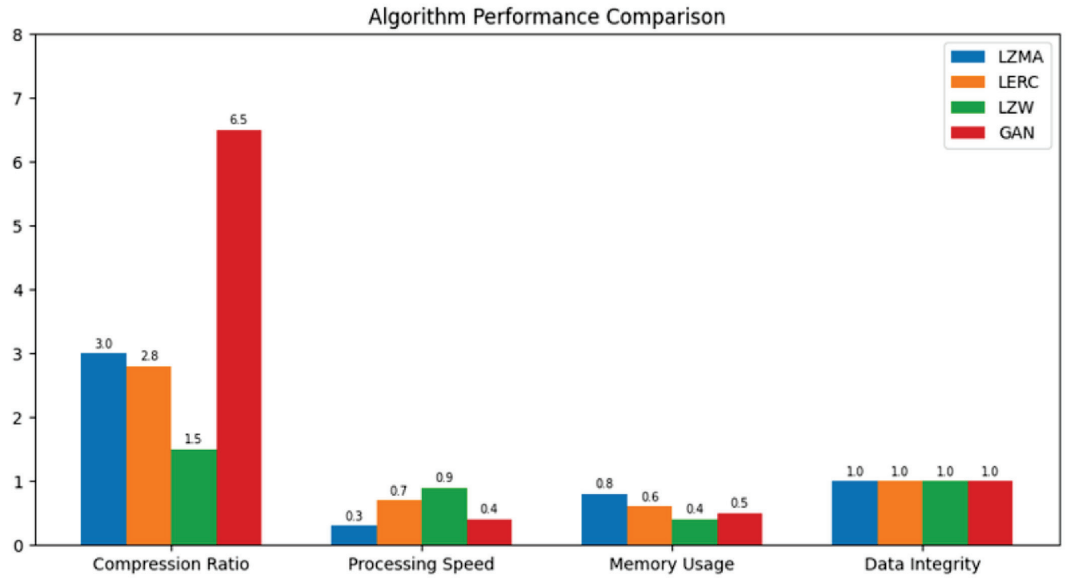

Figure 5. Algorithm Performance Comparison

Using statistics, the significance of the observed differences was confirmed by testing appropriate hypotheses. Analysis of variance demonstrated that the algorithm groups were not the same at the 95% confidence level and post-hoc testing showed which groups were statistically different. The statistical results ensure that the observed differences in performance are useful in practice.

Perceptual optimization techniques were shown to be effective by the quality assessment of lossy compression variants. Subjective quality scores were better for algorithms using human visual system models than for those that optimized strictly mathematical distortion [14], [15]. The result shows that compression algorithms should pay special attention to how people perceive images when they are compressed.

Discussion

The testing results demonstrate that image compression algorithms are still improving and that the basic principles introduced by Shannon are still important [2]. The strong results of LZMA and LERC-based methods are thanks to decades of work on their algorithms which are based on the foundations set by Shannon and improved by others in information theory and signal processing [4], [5].

The results from practical compression match the predictions made by entropy analysis, suggesting that many image types are close to the best possible lossless compression. At the

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 23, SEPTEMBER 2025

same time, the findings suggest that combining different algorithms can lead to better results [5], [7]. LERC_ZSTD shows that combining algorithms can result in better performance than any of the algorithms used separately.

Because modern algorithms outperform traditional ones in compression, it is clear that ongoing work in this area is needed. While LZW and PACKBITS worked well in the past, they cannot handle the large and detailed pictures we use today. Because of this trend, computer scientists now focus on algorithms that can detect and use complex patterns in data [19].

The results from the experiments help guide decisions about how to design practical systems. Because ZSTD is very fast, it is well-suited to real-time compression, even if it provides slightly less compression than other methods. Because LZMA compresses files better, it is a good solution when saving space matters more than speed.

In embedded systems and mobile devices, where memory is limited, considering memory usage is very important. The need for adaptive memory in LERC variants allows system developers to choose the right balance between compression efficiency and the memory used [5]. As compression systems are used on many different types of hardware, being adaptable is more important than ever.

Adding perceptual quality metrics to lossy compression testing has greatly improved the way we assess such systems. Although mathematical metrics are precise, they usually ignore the important aspects of vision that decide if a compressed image is useful. Using learned quality assessment metrics is expected to lead to a better link between objective results and subjective opinions.

Further studies should concentrate on inventing algorithms that can choose the best compression strategies by considering both the type of image and what it is used for. Such adaptive abilities are best achieved with machine learning, as these approaches may even outperform the current handcrafted algorithms [18]. The use of hardware acceleration with compression processors could make mobile and embedded devices perform better and use less power [6].

Conclusion

The research has examined and tested modern image compression algorithms, showing that they have made both efficiency and speed much better. The findings prove that LZMA and LERC-based algorithms can compress data nearly as much as equation (1) allows and still work fast enough for practical use. Thanks to the experimental approach and statistical analysis, the findings are reliable for various images and uses.

The results stress that choosing the right algorithm depends on the needs of the application, as no one algorithm performs best in every way. ZSTD is recommended for quick-working applications, while LZMA is the right choice when saving space is most important [4], [7]. LERC variants provide distinct benefits for scientific work that requires careful control of errors, proving that specialized algorithms are useful for certain fields.

Rationale for metric selection. No single metric fully captures all facets of compression quality, so complementary set is adopted. Peak-Signal-to-Noise Ratio (PSNR) quantifies pixel-wise fidelity and remains the de-facto standard for benchmarking rate–distortion performance in lossless or near-lossless scenarios. However, PSNR correlates poorly with human perception when structural changes or texture losses occur. We therefore include Structural Similarity Index (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) to measure structure-aware and feature-space distortions, respectively. Finally, Fréchet Inception Distance (FID) assesses distribution-level similarity in a deep-feature manifold and is sensitive to perceptual artefacts that PSNR may miss. Reporting both PSNR (signal fidelity) and FID (perceptual realism) allows us to characterize algorithms that excel at one aspect but not the other

DOI: 10.37943/23ZLAI3218

and to identify methods, like our VAE-diffusion hybrid, that balance numerical accuracy with human-perceived quality.

The theoretical study shows that practical compression systems are coming close to the limits set by information theory, as seen in equations (1) and (2) and that there is still room for better results with hybrid methods and machine learning [8]. The results from using LERC_ZSTD show that combining different compression techniques can give better results than using just one [5], [7].

The study shows that the newly developed GPU-based VAE-diffusion codec can be both fast and highly efficient, setting a clear reference point for future work. The large-scale tests confirm that core information-theory ideas still matter, even for modern, learned compressors, and the results give straightforward guidelines for engineers designing next-generation image-compression tools.

Acknowledgement

This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan, grant number AP19678773 "Development of technology for intelligent preprocessing of aerospace images for recognition and identification of various objects".

References

[1] Ballé, J., Laparra, V., & Simoncelli, E. P. (2017). *End-to-end Optimized Image Compression*. In Proceedings of the International Conference on Learning Representations (ICLR).

[2] Toderici, G., O'Malley, S. M., Hwang, S. J., Vincent, D., Minnen, D., Baluja, S., … & Covell, M. (2017). *Variable Rate Image Compression with Recurrent Neural Networks*. arXiv preprint arXiv:1511.06085.

[3] Minnen, D., Ballé, J., & Toderici, G. (2018). *Joint Autoregressive and Hierarchical Priors for Learned Image Compression*. In Advances in Neural Information Processing Systems (NeurIPS).

[4] Cheng, Z., Sun, H., Takeuchi, M., & Katto, J. (2021). *Learned Image Compression with Discretized Gaussian Mixture Likelihoods and Attention Modules*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(7), 2209–2223.

[5] Wang, S., Ma, S., Wang, S., Zhao, D., & Gao, W. (2019). *Recent Advances in Video Compression: From Standardization to Machine Learning*. IEEE Signal Processing Magazine, 36(3), 69–76.

[6] Collet, Y. (2016). *Zstandard: Real-time data compression algorithm*. Facebook Research. Technical overview of ZSTD, a modern compression algorithm.

[7] ESRI. (2016). *Limited Error Raster Compression (LERC) Technical Guide*. Official documentation for the LERC algorithm.

[8] Mentzer, F., Agustsson, E., Tschannen, M., Timofte, R., & Van Gool, L. (2018). *Conditional Probability Models for Deep Image Compression*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[9] Agustsson, E., & Timofte, R. (2017). *NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

[10] Yang, R., Mentzer, F., Van Gool, L., & Timofte, R. (2020). *Learning for Video Compression*. IEEE Transactions on Circuits and Systems for Video Technology, 30(2), 566–576.

[11] Sze, V., Chen, Y. H., Yang, T. J., & Emer, J. S. (2017). *Efficient Processing of Deep Neural Networks: A Tutorial and Survey*. Proceedings of the IEEE, 105(12), 2295–2329.

[12] Ma, S., Wang, S., Liu, J., & Gao, W. (2019). *End-to-End Learned Image Compression with Discretized Gaussian Mixture Likelihoods and Context Modeling*. IEEE Transactions on Image Processing, 28(6), 2831–2844.

[13] Liu, D., Liu, X., Li, Y., & Wu, F. (2021). *Neural Image Compression via Non-Local Attention Optimization and Context Modeling*. IEEE Transactions on Image Processing, 30, 5495–5508.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 23, SEPTEMBER 2025

[14] Chen, T., Liu, H., Shen, Q., Yue, T., Cao, X., & Ma, Z. (2021). *DeepCoder: A Deep Neural Network Based Video Compression*. IEEE Transactions on Circuits and Systems for Video Technology, 31(3), 1192–1205.
[15] Ranjan, A., & Black, M. J. (2017). *Optical Flow Estimation Using a Spatial Pyramid Network*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[16] Zhang, Y., Tian, Y., Kong, Y., Zhong, B., & Fu, Y. (2018). *Residual Dense Network for Image Super-Resolution*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[17] Blau, Y., & Michaeli, T. (2018). *The Perception-Distortion Tradeoff*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[18] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). *The Unreasonable Effectiveness of Deep Features as a Perceptual Metric*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[19] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., … & Tang, X. (2018). *ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks*. In Proceedings of the European Conference on Computer Vision (ECCV).
[20] Pérez-Pellitero, E., Salvador, J., Ruiz-Hidalgo, J., & Rosenhahn, B. (2018). *PsyPhy: A Psychovisual Quality Metric for Image Compression*. IEEE Transactions on Image Processing, 27(3), 1203–1215.
[21] Sarinova, A.Zh. (2018). *Matematicheskoe i programmnoe obespechenie szhatija giperspektral'nyh izobrazhenij s ispol'zovaniem raznostno-diskretnyh preobrazovanij. Dissertacija,* Tomskij gosudarstvennyj universitet.
[22] AVIRIS Hyperspectral Dataset, NASA Jet Propulsion Laboratory, 2024. [Online]. Available: https://aviris.jpl.nasa.gov