Cross-Layer Hybrid Compression for Efficient Fundus Image Transmission in Machine Vision Systems

Original Research Article

Abstract

To address the trade-off between bandwidth bottlenecks and feature fidelity in RESTful image transmission for tele-ophthalmology IoT systems, we propose a machine-vision-oriented cross-layer collaborative compression framework. This scheme integrates lossy dimensionality reduction (JPEG) at the semantic layer with entropy coding optimization (Zlib) at the transport layer. Utilizing the Messidor-2 dataset, we evaluated the feature consistency of ResNet-18 and MobileNetV2. The experiments identified an "Empirical Inflection Point" at $Q = 0.4$, where data volume was reduced by 57.3% while feature space cosine similarity remained above 0.920. Although classification accuracy experienced a minor 4% decline, the system maintained high inference stability. Furthermore, the Zlib strategy successfully eliminated the 33% encoding redundancy introduced by the Base64 protocol. This framework demonstrates that by co-optimizing the semantic and transport layers, low-latency transmission can be balanced with AI reliability, offering a scalable reference for broader medical IoT systems operating in bandwidth-constrained environments.

Keywords: Tele-ophthalmology; Image compression; Feature consistency; RESTful API; Cross-layer
2010 Mathematics Subject Classification: 68T45; 94A08; 92C55; 68T01

1 Introduction

Diabetic Retinopathy (DR) constitutes one of the primary causes of blindness within the global working-age population Sun et al. (2022); Magliano and Boyko (2021). Given the persistent escalation in diabetes prevalence Gong et al. (2025), automated screening for the condition is contingent upon efficient image acquisition and transmission systems. Recent meta-analyses have substantiated the significant efficacy of telemedicine in providing diabetic eye care to underserved populations Chidi and Akubue (2024); Khan et al. (2024). A systematic review by Farahat et al. Farahat et al. (2024) further underscores the pivotal role of AI algorithms in these screening processes. Nevertheless,

in practical telemedicine deployments, systems utilizing RESTful architectures confront a formidable engineering challenge: achieving low-latency transmission of high-resolution fundus images from edge devices (such as fundus cameras and smartphones) to cloud servers.

The primary bottleneck in this process is bandwidth. Although more efficient binary transmission protocols exist, such as gRPC, which demonstrates superior performance in certain high-concurrency scenarios Ain et al. (2025), RESTful API remains the preferred architecture for medical IoT systems, particularly when interconnecting heterogeneous edge devices Singh and Yadav (2024), due to its exceptional cross-platform interoperability and extensive support for Web standards Devalla (2018). Mabotha et al. Mabotha et al. (2025) explored the feasibility of dynamic RESTful API implementation within IoT environments, while Dwiyanto et al.'s Dwiyanto et al. (2025) performance analysis of real-time monitoring systems indicates that although REST APIs excel in interoperability, their text-based transmission mechanism imposes significant performance overheads in scenarios demanding high concurrency and low latency. However, standard fundus images typically range from 2 MB to 10 MB in size. In weak network environments (e.g., 3G/4G), RESTful protocols typically require images to be Base64-encoded within the JSON body, which artificially inflates the data size by approximately 33% Josefsson (2006), thereby exacerbating transmission latency.

We focus on the engineering optimization and signal processing analysis of transmission systems. Existing research has rarely analyzed the *joint impact* of compression artifacts on deep learning feature extraction from the perspective of "machine vision coding". The main contributions of this paper are as follows:

1. We provide a quantitative analysis of the relationship between image compression intensity (quality factor) and the Feature Space Stability of deep neural networks.
2. We propose a transmission stack optimization strategy based on information entropy principles, which effectively eliminates the inherent Base64 redundancy in RESTful architectures.
3. We verify the feasibility of this framework in edge computing scenarios through t-SNE visualization and computational complexity analysis.

The remainder of this paper is organized as follows: Section 2 systematically reviews existing work and technical challenges in the fields of medical image compression and Coding for Machines (CfM). Section 3 elaborates on the system architecture, mathematical metrics, and the rationale for deep learning model selection within the proposed two-stage hybrid compression framework. Section 4 details the experimental implementation and dataset preprocessing workflow, followed by a quantitative analysis of results across three dimensions: signal fidelity, feature space consistency, and end-to-end transmission latency. Section 5 discusses the mechanisms of transport layer optimization from an information-theoretic perspective and conducts an in-depth analysis of the clinical impact of precision loss alongside engineering limitations. Finally, Section 6 summarizes the key findings of this paper and outlines directions for future research.

2 Related Work

The application of image compression in telemedicine has been extensively investigated. Bourai et al. Bourai et al. (2024) systematically reviewed the challenges associated with deep learning-assisted medical image compression, identifying the trade-off between diagnostic quality and compression ratios under bandwidth constraints as a core difficulty. Traditional medical image compression standards (e.g., JPEG-LS or JPEG 2000 within DICOM) are primarily Human Visual System (HVS)-centric, aiming to optimize perceptual metrics such as PSNR or SSIM. To further enhance efficiency, Min et al. Min et al. (2022) explored anatomical information-based lossless compression, while Baiee et al. Baiee and AL-Assadi (2024) proposed hybrid models integrating deep learning with lossless techniques. However, with the proliferation of AI-aided diagnostic systems, the research focus is progressively shifting towards Coding for Machines (CfM). Against this backdrop, this study explores

”machine-friendly” hybrid compression strategies specifically targeting RESTful transmission links in ophthalmic scenarios.

2.1 Machine-Oriented Coding and Transmission Optimization

With the established dominance of deep learning in computer vision, traditional human-centric compression standards have begun to reveal their limitations. The Video Coding for Machines (VCM) working group, recently established by the MPEG organization, is dedicated to formulating new standards that seek a balance between compression efficiency and machine task performance. In this domain, Zhang et al. Zhang et al. (2024) proposed a perceptual video coding method based on the ”Satisfied Machine Ratio”, which optimizes coding parameters by modeling the satisfaction of machine vision tasks rather than solely pursuing pixel-level fidelity. Meanwhile, Lorkiewicz et al. Lorkiewicz et al. (2025) explored the application of neural network-based chroma synthesis techniques within VCM. Concurrently, research by Urbaniak Urbaniak (2024) indicates that utilizing compressed JPEG images in deep learning training requires a careful assessment of artifact impact, while Adzic Adzic (2023) compared the effects of various image encoders on machine task performance, further supporting the necessity of optimizing coding parameters specifically for machine vision.

At the transmission layer of medical IoT, bandwidth limitations and high latency constitute the primary challenges. De et al. De et al. (2025) reviewed multimedia transmission strategies over constrained networks such as LoRa, pointing out that cross-layer optimization tailored to specific application scenarios is pivotal for resolving bandwidth bottlenecks. Addressing the computational constraints of edge devices, Blesswin et al. Blesswin et al. (2025) proposed a lightweight semantic compression scheme. Furthermore, regarding the architectural performance of Web-based medical systems, Dwiyanto et al. Dwiyanto et al. (2025) noted that the text-based transmission overhead of REST APIs in high-concurrency scenarios cannot be ignored, thereby further emphasizing the necessity of efficient entropy coding at the transport layer.

2.2 Robustness of Deep Learning to Image Quality

Deep Neural Networks (DNNs) are generally considered to possess a certain degree of non-linear robustness against Gaussian noise; however, their response to structured artifacts, such as JPEG blocking effects, is more complex. The seminal work by Gulshan et al. Gulshan et al. (2016) demonstrated the expert-level performance of deep learning in DR detection, subsequent to which Doshi et al. Doshi et al. (2016) and Zhang et al. Zhang et al. (2022) further explored the adaptability of different network architectures. Targeting mobile applications, Intaraprasit et al. Intaraprasit et al. (2023) validated the effectiveness of lightweight networks in retinal disease classification, while Mubeena et al. Mubeena and Jawahar (2025) focused on lightweight compression and encrypted transmission within blockchain environments. This study investigates the signal fidelity of ResNet and MobileNet architectures under ophthalmic-specific data distributions, aiming to quantify this drift boundary.

3 Methodology

3.1 System Architecture

We adopted a typical client-server architecture utilized in modern telemedicine applications, as shown in Fig. 1. The workflow consists of three stages:

1. **Semantic Layer (Client):** The raw fundus image I_{raw} is downsampled and compressed using the JPEG standard with a controllable quality factor ($Q \in [0, 1]$). Although modern formats such as WebP offer superior compression ratios, considering that existing fundus

camera hardware primarily outputs JPEG streams and that transcoding at the edge introduces additional computational latency, we optimize the universal JPEG standard to ensure maximum hardware compatibility and low-latency processing.

2. **Transport Layer (Transmission):** The compressed binary stream is encoded as a Base64 string. Although this introduces approximately 33% volume redundancy, this design is intended to ensure "Transactional Atomicity". In medical scenarios, images must be strictly bound to patient metadata (e.g., ID, diagnosis time). Encapsulating the image within a JSON payload guarantees that both are transmitted as a single atomic unit. While multipart transmission (Multipart/form-data) avoids Base64 encoding overhead, it introduces the risk of data association misalignment in weak network environments. To counteract the volume inflation, the JSON payload is processed using lossless compression algorithms (such as Zlib or LZ4 Liu et al. (2018)) prior to transmission via HTTP.
3. **Inference Layer (Server):** The server decompresses the payload and feeds the reconstructed image $\hat{I}$ into the AI model for inference.

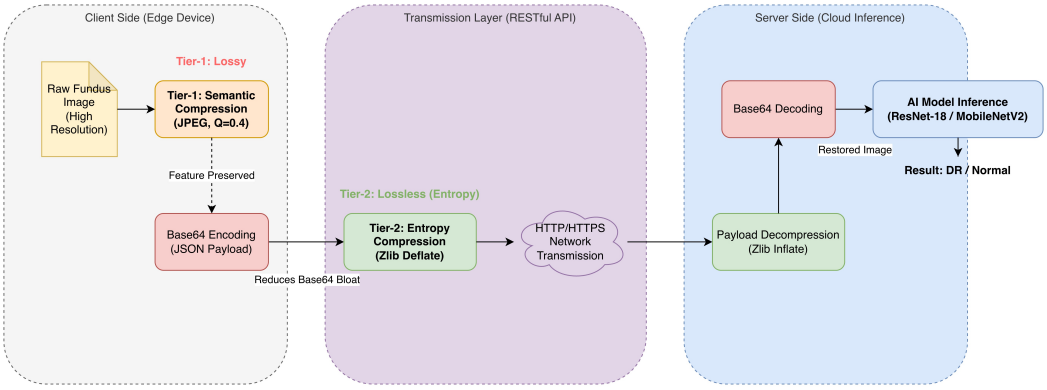

Figure 1: The proposed two-stage compression framework for RESTful tele-ophthalmology.

3.2 Compression and Image Quality Metrics

The visual information loss in the first stage is governed by the JPEG quantization table. To evaluate image fidelity from a signal processing perspective, we compute the Mean Squared Error (MSE) and the Peak Signal-to-Noise Ratio (PSNR):

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i,j) - K(i,j)]^2 \tag{3.1}$$

$$PSNR = 10 \cdot \log_{10} \left(\frac{MAX_I^2}{MSE} \right) \tag{3.2}$$

Furthermore, to quantify the preservation of structural information, the Structural Similarity Index Measure (SSIM) Wang et al. (2004) is employed:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{3.3}$$

Secondly, to quantify the consistency of images in the deep feature space before and after compression, we compute the Cosine Similarity between feature vectors. Let v_{raw} and v_{cmp} denote the feature vectors output by the penultimate layer of the DNN for the original and compressed images, respectively; the similarity is defined as:

$$\text{CosSim}(v_{raw}, v_{cmp}) = \frac{v_{raw} \cdot v_{cmp}}{\|v_{raw}\|\|v_{cmp}\|} \tag{3.4}$$

This metric ranges from $[-1, 1]$, with values approaching 1 indicating greater consistency in feature orientation and more complete preservation of semantic information. Furthermore, we employ t-SNE van der Maaten and Hinton (2008) to conduct visual dimensionality reduction analysis on the high-dimensional feature space.

Finally, the File Reduction Rate (η) is defined as:

$$\eta = \left(1 - \frac{S_{compressed}}{S_{original}}\right) \times 100\% \tag{3.5}$$

3.3 Deep Learning Model Architectures and Selection

To comprehensively evaluate the generalizability of the compression framework across diverse computational environments, this study selected two representative Convolutional Neural Network (CNN) architectures: ResNet-18 and MobileNetV2.

3.3.1 ResNet-18: High-Fidelity Benchmark for Cloud-Side Inference

Serving as a general-purpose benchmark for server-side inference, ResNet-18 He et al. (2016) employs a Residual Learning Framework to address the degradation problem inherent in deep networks. Its fundamental unit, the Residual Block, facilitates identity mapping through the introduction of "skip connections":

$$y = \mathcal{F}(x, \{W_i\}) + x \tag{3.6}$$

Where x and y denote the input and output vectors, respectively, and $\mathcal{F}$ represents the residual function. This architectural configuration renders the network highly sensitive to the loss of high-frequency details in input images, thereby making it an ideal candidate for assessing the impact of compression artifacts on the integrity of feature extraction.

3.3.2 MobileNetV2: Lightweight Benchmark for Edge Computing

MobileNetV2 Sandler et al. (2018) is explicitly designed for resource-constrained mobile devices, incorporating "Inverted Residuals" and "Linear Bottlenecks". Distinct from standard convolution, it leverages Depthwise Separable Convolutions to significantly reduce computational costs. For a feature map with input dimensions $D_F \times D_F \times M$, the computational cost of depthwise separable convolution is calculated as follows:

$$\text{Cost}_{dw} = D_K^2 \cdot M \cdot D_F^2 + M \cdot N \cdot D_F^2 \tag{3.7}$$

Here, D_K represents the convolution kernel size, and N denotes the number of output channels. Compared to the $D_K^2 \cdot M \cdot N \cdot D_F^2$ of standard convolution, the computational complexity is reduced by a factor of approximately 8 to 9. This enables MobileNetV2 to operate efficiently on portable fundus cameras, yet it also renders it less robust to input noise than ResNet, thereby constituting the "lower bound of robustness" for this study.

3.3.3 Rationale for Model Selection

The selection of these specific architectures is primarily predicated on the following considerations:

1. **Coverage of Typical Application Scenarios:** ResNet-18 serves as a proxy for high-performance cloud server environments, prioritizing the semantic fidelity of features; conversely, MobileNetV2 represents resource-constrained Edge Computing scenarios, emphasizing the equilibrium between transmission bandwidth and computational efficiency. This dichotomy encompasses the two most prevalent deployment modalities within telemedicine.
2. **Dataset Adaptability:** Given the moderate scale of the Messidor-2 dataset (1,744 images), architectures such as Vision Transformers (ViT) often struggle to converge due to a lack of inductive bias, whereas VGG networks exhibit excessive parameter redundancy and low inference efficiency. ResNet-18 and MobileNetV2 have demonstrated superior convergence and generalization capabilities on small-to-medium-scale medical imaging datasets.

4 Experiments

4.1 Implementation Details

All experiments were conducted in a Python 3.14 environment, primarily utilizing the PyTorch 2.9.1 deep learning framework and the CUDA 12.8 acceleration library. Image preprocessing employed Pillow 12.1.0, while t-SNE visualization and evaluation metric calculations relied on Scikit-learn 1.7.2. The hardware platform was equipped with an NVIDIA RTX 3080 GPU. The deep learning models were implemented using PyTorch. We utilized Transfer Learning by initializing both ResNet-18 and MobileNetV2 with weights pre-trained on ImageNet. The final fully connected layers were replaced to match the binary classification task. During the training phase, data augmentation techniques were applied to prevent overfitting, including random horizontal flips and random rotations ($\pm 10^{\circ}$), followed by normalization using standard ImageNet mean and standard deviation. The network training employed the Adam optimizer with an initial learning rate of 1×10^{-5} and a batch size of 16. The training process was set to a maximum of 100 epochs, incorporating an Early Stopping mechanism that monitored the F1-score on the validation set with a patience of 10 epochs to ensure the retrieval of the optimal model checkpoint.

4.2 Dataset and Preprocessing Workflow

We utilized the Messidor-2 dataset Decencière et al. (2014); Krause et al. (2018), which is widely adopted in diabetic retinopathy screening. Specifically, to ensure experimental consistency and eliminate extraneous background noise, we adopted the preprocessed version published by Herrero on Kaggle Herrero (2020), in which the superfluous black fundus background has been cropped. Although the official documentation indicates a total of 1,748 images, subsequent to deduplication and annotation alignment verification, the effective dataset comprises 1,744 high-resolution color retinal images valid for experimentation.

4.2.1 Data Partitioning and Standardization

We partitioned the corpus of 1,744 images using a Stratified Sampling strategy to ensure consistency in class distribution across all subsets. First, we reserved 20% of the data as an independent test set (349 images). Subsequently, 10% of the remaining data was allocated as a validation set (140 images) for hyperparameter tuning, with the remainder serving as the training set (1,255 images).

To eliminate potential biases arising from inconsistent raw data formats and to establish a unified signal baseline, all images were converted to Quality=100 JPEG format (with chroma subsampling

disabled, subsampling=0) via a Python script (utilizing the PIL library) prior to partitioning. The high-quality baseline JPEG images generated in this step are defined as the "High-Quality Baseline" ($Q = 1.0$) for this study.

4.2.2 Test Set Gradient Generation Strategy

To evaluate the dynamic impact of compression artifacts on model inference, we constructed a graduated compression scheme specifically for the test set. The specific procedure is outlined as follows:

1. **Baseline Establishment:** The test set with a quality factor of $Q = 1.0$ is retained to serve as the control group.
2. **Distortion Injection:** Utilizing standard JPEG quantization tables, each image in the baseline test set was compressed at a step interval of 0.1 to generate nine subsets of varying compression intensities, corresponding to quality factors $Q \in \{0.1, 0.2, \ldots, 0.9\}$.
3. **Isolated Storage:** Test data associated with distinct Q values were stored in isolated directories.

During the inference phase, the trained models (trained on $Q = 1.0$ data) were evaluated separately across these ten test subsets (inclusive of the baseline). This experimental design ensures that the results reflect the intrinsic impact of the compression algorithm rather than discrepancies in data distribution.

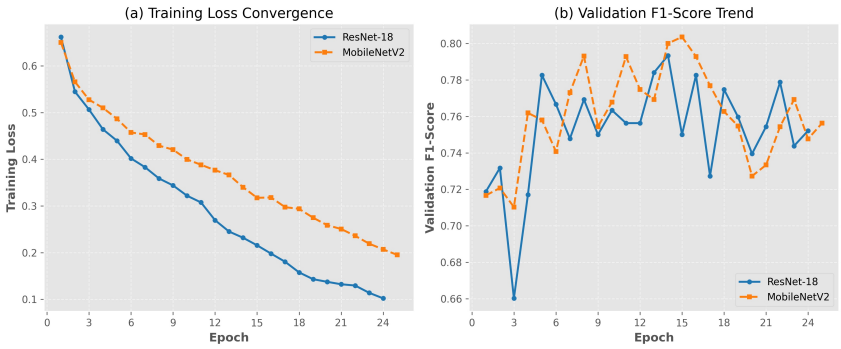

Figure 2: Training convergence curves for ResNet-18 (blue line) and MobileNetV2 (orange line).

4.3 Results and Analysis

4.3.1 Model Complexity and Baseline Performance

The training convergence curves for both model architectures are illustrated in Fig. 2; despite possessing fewer parameters, MobileNetV2 exhibited a convergence rate comparable to that of ResNet-18. Table 1 further presents a comparative analysis of computational complexity, indicating that MobileNetV2 achieves an order-of-magnitude reduction in FLOPs, thereby better aligning with the requirements of telemedicine terminals.

Table 1: Comparison of Deep Learning Model Complexity (Input Size: 224×224)

Model	Parameters (M)	FLOPs (G)	Baseline Accuracy
ResNet-18	11.18	1.82	75.9%
MobileNetV2	2.23	0.33	75.1%

4.3.2 Impact of Semantic Layer Compression on Feature Consistency

Table 2 reveals the non-linear trade-off between compression ratios and model performance. Although high Q values (0.6–0.8) maintained high classification accuracy, their volume reduction rates were marginal (ranging only from 9.4% to 18.5%), failing to yield substantial latency optimization in bandwidth-constrained mobile network environments. In contrast, $Q = 0.4$ exhibited the optimal engineering Cost-Benefit Ratio:

1. **Significant Bandwidth Savings:** The data volume reduction rate reached 57.3%, significantly surpassing the 32.9% achieved at $Q = 0.5$, thereby realizing a substantial reduction in data quantity.
2. **Controlled Accuracy Loss:** Although the accuracy of ResNet-18 decreased from the baseline of 0.759 to 0.719, the feature cosine similarity remained at a high level of 0.920. This indicates that $Q = 0.4$ effectively eliminated visual redundancy within the images while preserving the core semantic structures critical for machine inference.

We identified an 'Empirical Inflection Point' at $Q = 0.4$. We define this point as the threshold where the second derivative of the feature consistency curve is maximized; beyond this point, further compression yields diminishing returns in file size reduction while causing a precipitate drop in cosine similarity. It is noteworthy that the threshold of $Q = 0.4$ reflects the signal characteristics specific to the Messidor-2 dataset and its acquisition conditions. In practical deployments, this optimal Operating Point should be regarded as a tunable parameter necessitating calibration according to the Signal-to-Noise Ratio (SNR) characteristics of different fundus camera models.

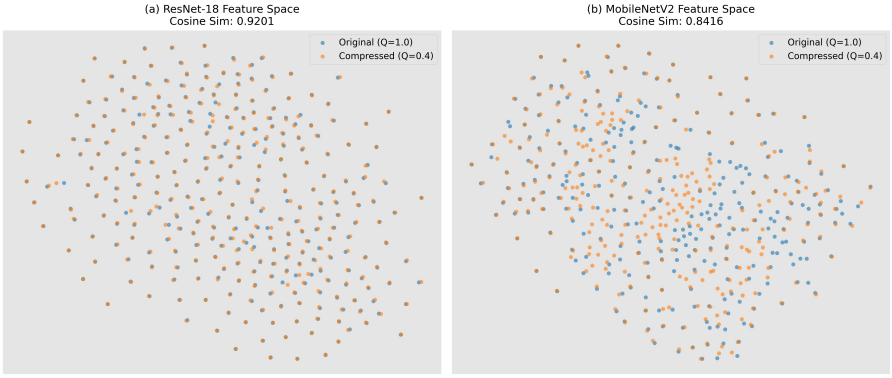

Figure 3: Visualization of the t-SNE feature space.

Quantitative analysis (Table 2) further confirms that at $Q = 0.4$, despite slight fluctuations in model accuracy, the feature cosine similarity consistently remains above 0.920, demonstrating the efficacy of this compression intensity in preserving critical semantic features. Furthermore, the dimensionality reduction visualization of the t-SNE feature space, as shown in Fig. 3, intuitively

reveals that the compressed samples at $Q = 0.4$ (orange) exhibit a high degree of overlap with the original samples (blue) within the manifold space. This overlap indicates that, despite substantial loss of pixel-level information, the topological structure and inter-cluster distances utilized for distinguishing lesions remain intact, thereby confirming the robustness of semantic features against compression artifacts.

Table 2: Variation of AI Model Performance and Feature Space Consistency with Compression Rate

Q	Volume Reduction	ResNet-18			MobileNetV2		
		Acc.	F1	Cos Sim.	Acc.	F1	Cos Sim.
0.1	81.8%	0.450	0.597	0.700	0.415	0.587	0.548
0.2	72.4%	0.564	0.635	0.823	0.415	0.587	0.670
0.3	61.2%	0.642	0.654	0.902	0.544	0.636	0.790
0.4	**57.3%**	**0.719**	**0.696**	**0.920**	**0.645**	**0.668**	**0.842**
0.5	32.9%	0.696	0.686	0.942	0.682	0.680	0.881
0.6	18.5%	0.736	0.703	0.966	0.682	0.663	0.924
0.7	9.4%	0.768	0.712	0.987	0.731	0.678	0.970
0.8	1.1%	0.751	0.667	0.991	0.739	0.643	0.982
0.9	-17.6%	0.754	0.677	0.996	0.759	0.684	0.991
1.0	0.0%	0.759	0.691	1.000	0.751	0.684	1.000

4.4 Transport Layer Efficiency and Latency Analysis

As shown in Table 3, the binary stream resulting from Tier-1 compression (Original Size: 2.79 MB) expanded significantly to 3.72 MB (Ratio 1.333) following Base64 encoding. However, the Zlib algorithm demonstrated superior entropy coding capabilities, re-compressing the data volume to 2.71 MB. This result not only completely eliminated the 33% transmission overhead introduced by Base64 but also reduced the final transmission payload to slightly below that of the original binary stream (System Reduction 2.87%). This ensures that the 57.3% bandwidth savings achieved by the semantic layer compression ($Q = 0.4$) are genuinely translated into end-to-end latency reductions, rather than being negated by the encoding losses of the application layer protocol. The detailed breakdown of end-to-end transmission latency is shown in Fig. 4; although Zlib introduces a minor computational overhead (green section), it significantly shortens the network transmission time (blue section), thereby optimizing the overall system response speed.

5 Discussion

To rigorously validate the proposed framework, we position this study not as a proposal for a novel CNN architecture, but as a system-level transmission model optimized for medical IoT. Unlike traditional single-layer compression approaches, our "Proofing Outcome" derived from the comparative analysis in Table 3 and Fig. 4 demonstrates that the proposed hybrid pipeline (JPEG $Q = 0.4$ + Zlib) constitutes the "Best Possible Model" for bandwidth-constrained scenarios. It outperforms standard Base64 transmission by strictly eliminating 33% redundancy and achieves a superior trade-off between latency and feature fidelity compared to pure dictionary-based algorithms like LZ4.

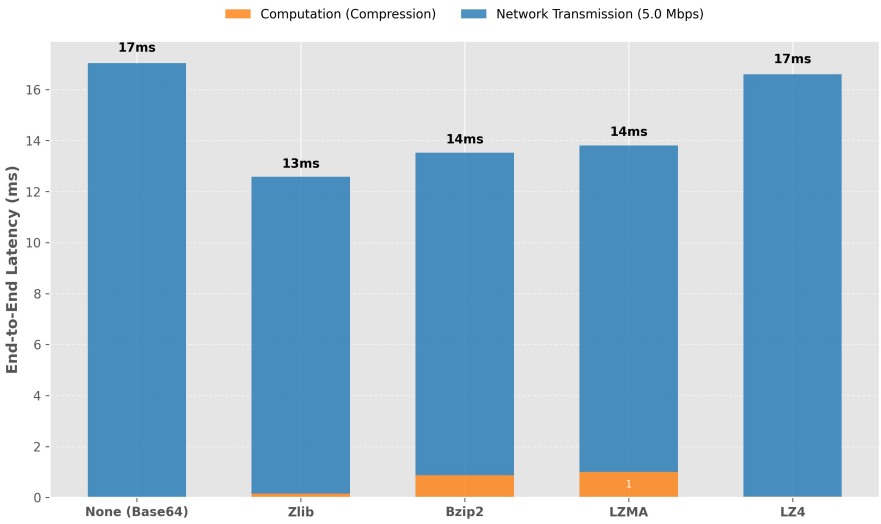

Figure 4: Comparison of End-to-End Transmission Latency Breakdown (4G Network, 5Mbps Upload).

Table 3: Comparison of Transport Layer Compression Efficiency (Base64 Encoded Stream)

Algorithm	**Orig. (MB)**	**Comp. (MB)**	**Ratio**	**Red. (%)**	**Mechanism**
None (Base64)	2.79	3.72	1.333	-33.33%	-
Zlib	2.79	**2.71**	**0.971**	**2.87%**	**Entropy + Dict**
Bzip2	2.79	2.76	0.989	1.08%	BWT + Entropy
LZMA	2.79	2.79	1.002	-0.16%	Dict + Range Coding
LZ4	2.79	3.65	1.308	-30.82%	Pure Dict (LZ77)

Note: Orig. = Original size; Comp. = Compressed size; Ratio = Compression ratio; Red. = Space reduction percentage.

5.1 Mechanism of Transmission Optimization Based on Information Entropy

Experimental results demonstrate that Zlib is extremely effective in compressing Base64 data. Analyzing from the perspective of Shannon information theory Shannon (1948); Verdú (1998), Base64 encoding maps the binary stream to a constrained subset of 64 characters; this mapping disrupts the local correlations of the original byte stream, leading to reduced efficiency in dictionary-matching-based algorithms (such as LZ4). However, the Base64 character stream exhibits an extremely non-uniform probability distribution (low-entropy characteristic); the Huffman coding stage built into Zlib Deutsch and Gailly (1996) is able to effectively exploit this statistical skewness, thereby completely eliminating the 33% redundancy introduced by the encoding protocol at the transport layer.

5.2 Analysis of Accuracy Loss and Error Patterns

Although the feature cosine similarity remained at a high level of 0.920 at $Q = 0.4$, the classification accuracy of ResNet-18 experienced a decline of approximately 4%. This suggests that certain "Hard Samples" located near the decision boundaries are relatively sensitive to the loss of high-frequency information. This minor loss of precision is acceptable within "Human-in-the-loop" workflows: the AI system primarily assumes the role of initial screening and load triage, while the low-latency advantage conferred by high compression significantly expands screening coverage, yielding public health benefits that far outweigh the precision loss on marginal samples.

5.3 Limitations Analysis

Although our framework achieved significant bandwidth gains at $Q = 0.4$, this strategy is essentially a globally uniform compression scheme. It does not account for the anatomical characteristics of fundus images, specifically, that subtle lesions within the Optic Disc and Macula regions may be more sensitive to compression artifacts. In contrast, Region of Interest (ROI)-based adaptive compression might offer advantages in preserving critical lesion details, yet this necessitates more complex edge-side preprocessing algorithms. Balancing the computational load on edge devices, this study opted for the global compression strategy due to its lower computational cost.

Furthermore, while modern formats like WebP or AVIF offer superior compression ratios compared to JPEG, this study prioritized JPEG ($Q = 0.4$) to ensure maximum compatibility with legacy fundus cameras and PACS systems widely used in low-resource settings, which often lack native support for newer codecs.

Finally, we acknowledge the limitation of using a single experimental run in this study due to computational constraints. While we identified a clear empirical inflection point, the absence of repeated trials and confidence intervals limits the statistical robustness of our findings. Future work will necessitate large-scale cross-validation to further verify the stability of these results.

5.4 Future Work

Building upon the baselines established in this study, future work will aim to further push the boundaries between transmission bandwidth and inference accuracy. The specific directions are outlined as follows:

1. **Network-Aware Adaptive Transmission:** The current $Q = 0.4$ represents a global optimum determined based on static datasets. Future research will explore dynamic control algorithms based on Deep Reinforcement Learning (DRL). By treating real-time network conditions (e.g., RTT, packet loss rate) as environmental states, this algorithm will dynamically adjust the compression quality factor Q. It will automatically downgrade quality during severe network

fluctuations to ensure real-time performance, and upgrade quality when bandwidth is ample to enhance diagnostic confidence, thereby achieving a closed-loop ”transmission-inference” feedback control mechanism.

2. **Next-Generation Neural Compression for Machines:** With the evolution of the MPEG-VCM (Video Coding for Machines) standard and standardization efforts regarding learning-based visual data coding technologies Lorkiewicz et al. (2025); Liu et al. (2024), we will move beyond traditional JPEG encoding to explore end-to-end Learnable Image Compression frameworks. Drawing inspiration from the work of Zhang et al. Zhang et al. (2024) and recent advances in coarse-to-fine Transformers Yang et al. (2024) and cross-attention mechanisms Dai (2025), the objective is to design a loss function that directly optimizes “feature map fidelity” rather than pixel-level PSNR. Furthermore, by incorporating Semantic Communication principles Zhang et al. (2025); Yuan et al. (2024) and Variational Autoencoders (VAE) Xiang et al. (2024), the encoder can proactively discard visual background information that is irrelevant to AI inference, thereby achieving higher inference accuracy at extremely low bitrates.
3. **Clinical Validation and Subjective Quality Assessment:** Although this study has verified the robustness of machine vision, ”Human-in-the-loop” workflows remain indispensable in medical scenarios. Future work will involve conducting clinical pilot studies, inviting ophthalmologists to perform Turing Tests and Mean Opinion Score (MOS) evaluations on compressed images. This will ensure that the proposed compression scheme, while satisfying AI diagnostic requirements, does not introduce misleading visual artifacts during the physician review process.

6 Conclusion

We present a Cross-Layer Synergistic Optimization framework tailored for RESTful tele-ophthalmology. Distinct from traditional single-dimensional compression strategies, this framework leverages lossy compression at the Tier-1 semantic layer to precisely eliminate ”high-frequency visual redundancy” irrelevant to machine inference, while integrating entropy coding at the Tier-2 transport layer to effectively remove the ”encoding redundancy” introduced by the Base64 protocol.

Experimental results indicate that this dual-pronged strategy achieves a 57.3% data reduction while maintaining the cosine similarity within the deep feature space above 0.920. This finding provides strong empirical support for the theory of ”Coding for Machines” (CfM), demonstrating that within bandwidth-constrained edge medical scenarios, deep neural networks can tolerate significant visual signal loss without sacrificing semantic consistency. Furthermore, Zlib’s complete negation of the entropy increase at the transport layer fundamentally resolves the protocol overhead bottleneck that has long constrained Web-based medical systems. This work not only provides a ”plug-and-play” engineering solution compatible with legacy hardware for existing tele-ophthalmology systems but also demonstrates the feasibility of deploying low-latency, high-precision edge medical screening services in weak 3G/4G network environments, thereby establishing a critical reference benchmark for the future construction of high-throughput medical edge intelligence networks.

References

Adzic, V. (2023). Comparative analysis of image encoders and compression effects on machine task performance. In *2023 International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 1–6. IEEE.

Ain, M. Z., Ardiansyah, R., Pratama, S. A., Akbar, M., and Lapatta, N. T. (2025). Comparative performance analysis of grpc and rest api under various traffic conditions and data sizes using a quantitative approach. *J. Appl. Inform. Comput.*, 9:450–457.

Baiee, S. H. and AL-Assadi, T. A. (2024). Deep learning-based model for medical image compression. *J. Intell. Syst. Internet Things*, 13.

Blesswin, A. J., Selva Mary, G., Sankaranarayanan, S., Khan, A. R., Sait, A. R. W., and Lorenz, P. (2025). Lightweight semantic compression visual cryptography for secure medical image transmission in iot systems. *Sci. Rep.*, 15:21234.

Bourai, N. E. H., Merouani, H. F., and Djebbar, A. (2024). Deep learning-assisted medical image compression challenges and opportunities: systematic review. *Neural Comput. Appl.*, 36:10067–10108.

Chidi, R. and Akubue, I. (2024). Telemedicine in diabetic eye care: A meta-analysis of its effectiveness in underserved populations. *World J. Biol. Pharm. Health Sci.*, 17:131–139.

Dai, F. (2025). Deep learning based medical image compression using cross attention learning and wavelet transform. *Scientific Reports*, 15(1):40008.

De, S., Jalajamony, H. M., Adhinarayanan, S., Joshi, S., Upadhyay, H., and Fernandez, R. (2025). Multimedia transmission over lora networks for iot applications: A survey of strategies, deployments, and open challenges. *Sensors*, 25:7128.

Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., and Gain, P. e. a. (2014). Feedback on a publicly distributed image database: The messidor database. *Image Anal. Stereol.*, 33:231–234.

Deutsch, P. and Gailly, J.-L. (1996). Rfc 1950: Zlib compressed data format specification version 3.3. Network Working Group.

Devalla, S. (2018). Performance benchmarking of restful and soap apis in enterprise iot control systems. *J. Sci. Eng. Res.*, 5:376–390.

Doshi, D., Shenoy, A., Sidhpura, D., and Gharpure, P. (2016). Diabetic retinopathy detection using deep convolutional neural networks. In *2016 International Conference on Computing, Analytics and Security Trends (CAST)*, pages 261–266. IEEE.

Dwiyanto, R. A., Mutiara, G. A., and Sari, M. I. (2025). Performance analysis of rest api in a real-time iot-based vehicle monitoring system. *Int. J. Reconfigurable Embedded Syst.*, 2089:4864.

Farahat, Z., Zrira, N., Souissi, N., Bennani, Y., Bencherif, S., Benamar, S., Belmekki, M., Ngote, M. N., and Megdiche, K. (2024). Diabetic retinopathy screening through artificial intelligence algorithms: A systematic review. *Surv. Ophthalmol.*, 69:707–721.

Gong, W., Pu, Y., Ning, T., Zhu, Y., Mu, G., and Li, J. (2025). Deep learning for enhanced prediction of diabetic retinopathy: a comparative study on the diabetes complications data set. *Front. Med.*, 12:1591832.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., and Venugopalan, S. e. a. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316:2402–2410.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Herrero, M. (2020). Messidor-2 (preprocessed). Kaggle.

Intaraprasit, P., Bui, T. H., and Paing, M. P. (2023). Mobilenetv2-based deep learning for retinal disease classification on a mobile application. In *2023 15th Biomedical Engineering International Conference (BMEiCON)*, pages 1–5. IEEE.

Josefsson, S. (2006). The base16, base32, and base64 data encodings. RFC 4648, RFC Editor.

Khan, I. A., Bashar, M. A., Tripathi, A., Priyanka, N., Bashar, M. A., and Priyanka II, N. (2024). The benefits and challenges of implementing teleophthalmology in low-resource settings: A systematic review. *Cureus*, 16.

Krause, J., Gulshan, V., Rahimy, E., Karth, P., Widner, K., Coram, G. S., Peng, L., and Webster, D. R. (2018). Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*, 125:1264–1272.

Liu, D., Liu, S., Ascenso, J., Tian, D., and Yu, L. (2024). Guest editorial special section on recent standardization efforts for learning-based visual data coding. *IEEE Trans. Circuits Syst. Video Technol.*, 34:3063–3066.

Liu, W., Mei, F., Wang, C., O'Neill, M., and Swartzlander, E. E. (2018). Data compression device based on modified lz4 algorithm. *IEEE Trans. Consum. Electron.*, 64:110–117.

Lorkiewicz, M., Różek, S., Stankiewicz, O., Grajek, T., Maćkowiak, S., and Domański, M. (2025). Video coding for machines with neural-network-based chroma synthesis. *IEEE Access*.

Mabotha, E., Mabunda, N. E., Ali, A., and Khan, B. (2025). Exploring dynamic restful api implementation in iot environments using docker. *Sci. Rep.*, 15:34267.

Magliano, D. J. and Boyko, E. J. e. (2021). *IDF Diabetes Atlas*. International Diabetes Federation, Brussels, 10th ed. edition.

Min, Q., Wang, X., Huang, B., and Zhou, Z. (2022). Lossless medical image compression based on anatomical information and deep neural networks. *Biomed. Signal Process. Control*, 74:103499.

Mubeena, S. and Jawahar, P. K. (2025). Lightweight compression and chaos-based encryption for secure iot healthcare data storage on blockchain. *Eng. Technol. Appl. Sci. Res.*, 15:29759–29769.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:379–423.

Singh, A. P. and Yadav, S. K. (2024). A secure data collection model in iot. *Int. J. Creat. Res. Thoughts*, 12:e151–e158.

Sun, H., Saeedi, P., Karuranga, S., Pinkepank, M., Ogurtsova, K., Duncan, B. B., and Stein, C. e. a. (2022). Idf diabetes atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res. Clin. Pract.*, 183:109119.

Urbaniak, I. A. (2024). Using compressed jpeg and jpeg2000 medical images in deep learning: A review. *Appl. Sci.*, 14:10524.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *J. Mach. Learn. Res.*, 9:2579–2605.

Verdú, S. (1998). Fifty years of shannon theory. *IEEE Trans. Inf. Theory*, 44:2057–2078.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.*, 13:600–612.

Xiang, Z., Wang, Z., Xie, J., Wang, W., and Xi, L. (2024). Research on medical image data enhancement based on variational autoencoders. In *2024 International Conference on Virtual Reality and Visualization (ICVRV)*, pages 78–81. IEEE.

Yang, X., Lu, G., Feng, D., Cheng, Z., Yu, G., and Song, L. (2024). Coarse-to-fine transformer for lossless 3d medical image compression. In *2024 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5. IEEE.

Yuan, C., Ye, H., and Miao, Y. (2024). Universal image semantic communication for edge network. In *2024 4th International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*, pages 316–320. IEEE.

Zhang, G., Sun, B., Zhang, Z., Pan, J., Yang, W., and Liu, Y. (2022). Multi-model domain adaptation for diabetic retinopathy classification. *Front. Physiol.*, 13:918929.

Zhang, J., Zhang, Y., Ji, B., Chen, A., Liu, A., and Xu, H. (2025). Feature-driven semantic communication for efficient image transmission. *Entropy*, 27(4):369.

Zhang, Q., Wang, S., Zhang, X., Jia, C., Wang, Z., Ma, S., and Gao, W. (2024). Perceptual video coding for machines via satisfied machine ratio modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46:7651–7668.