SFTNet: Spatial-Frequency Transformer Network for Learned Image Compression

Meiqin Liu, Lingxue Li, Jiaming Liang, Chao Yao, *Member, IEEE*, Jian Jin, *Member, IEEE*, Tammam Tillo, *Senior Member, IEEE*, and Yao Zhao, *Fellow, IEEE*

***Abstract*—In learned image compression (LIC), most existing methods process spatial domain information through convolutional neural networks (CNNs) or Transformers. However, they struggle to model frequency-domain correlations, particularly high-frequency detail correlations across regions. They suffer from high-frequency detail redundancy and inefficient uniform bit allocation, neither of which adapts to regional complexity. To address these issues, we propose the Spatial-Frequency Transformer Network (SFTNet) for LIC, which models cross-domain correlations between spatial structures and frequency components to enable region-adaptive bit allocation. To model these correlations, the Frequency-aware Transformer Block (FATB) employs dual attention mechanisms to collaboratively process spatial and frequency components. Specifically, its intra-block frequency attention enhances high-frequency details within each block by using low-frequency components as local structural guidance. Meanwhile, its inter-block spatial attention captures global consistency by modeling cross-block dependencies among low-frequency components. Building on this, the Feature Reweighting Strategy (FRS) evaluates block importance via joint analysis of spatial dependencies and high-frequency energy. This enables dynamic alignment of the spatial-frequency features modeled by FATB for adaptive bit allocation, prioritizing structurally complex regions and compressing smooth areas to reduce redundancy. Experimental results demonstrate that our SFTNet outperforms VVC by −6.55% and −6.18% in BD-Rate on Kodak and CLIC datasets while achieving better reconstruction quality of the high-frequency details compared to recent LIC methods.**

***Index Terms*—Learned image compression, adaptive bit allocation, spatial-frequency features, dual attention mechanisms.**

I. Introduction

IMAGE compression aims to represent images with fewer bits and ensure acceptable reconstruction quality. The traditional image compression codecs, such as JPEG [1], JPEG2000 [2], BPG [3] and VVC Intra [4], involve artificially designed and independently optimized modules, including transform, quantization, entropy coding, and prediction. In contrast to the traditional codecs, learned image compression (LIC) methods [5], [6] optimize these modules in an end-to-end manner. LIC methods minimize the rate-distortion (R-D) cost to balance the bitrate (compression efficiency) and distortion (image quality). Some LIC methods [7], [8], [9], [10], [11], [12], [13], [14] even achieved superior R-D performance, and outperformed VVC Intra [4].

Nonlinear analysis and synthesis transforms in early LIC methods [12], [15], [16], [17], [18], [19], [20], [21] were typically achieved via convolutional neural networks (CNNs). While the local receptive fields of CNNs excel at capturing fine-grained details, their inability to model global relationships directly leads to redundant feature encoding in complex scenes [22]. In response, Transformers were introduced in recent works [7], [9], [23], [24], [25], [26] to model global spatial relationships for better R-D performance. Despite their strength in capturing global dependencies, Transformers often require heavy computational resources and overlook fine-grained local features.

To address this, hybrid CNN-Transformer models [7], [8] balanced efficiency and representation by integrating CNN local feature extraction and Transformer global modeling. A multi-reference entropy model was introduced in MLIC [10] to enhance image compression by extracting spatial multi-scale features. However, there was a lack of interpretation regarding the frequency features in natural images, which played a crucial role in image representation. Standard attention layers processed all features uniformly, failing to discriminate the semantic importance between low-frequency structures and high-frequency details. For example, wavelet analysis [27] decomposed images into multiscale subbands, revealing structural information and details across different frequency ranges. Moreover, most LIC methods [10], [11], [20] still adopt uniform bit allocation strategies, which fail to adapt to regional complexity by assigning equal bits to both complex regions and homogeneous regions. These strategies entail bit wastage in low-frequency regions and insufficient representation in high-frequency areas.

In this paper, we propose the Spatial-Frequency Transformer Network (SFTNet) for LIC, which enables adaptive bit allocation via spatial-frequency collaboration. The Frequency-aware Transformer Block (FATB) is designed to decompose features into low-frequency and high-frequency component

Received 14 July 2025; revised 6 November 2025; accepted 14 November 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62372036, Grant 62120106009, Grant 62332017, Grant U24B20179, and Grant U22A2022. *(Corresponding author: Chao Yao.)*

Meiqin Liu, Lingxue Li, Jiaming Liang, and Yao Zhao are with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China, and also with the Visual Intelligence +X International Cooperation Joint Laboratory of MOE, Beijing 100044, China (e-mail: mqliu@bjtu.edu.cn; lingxueli@bjtu.edu.cn; 20120305@bjtu.edu.cn; yzhao@bjtu.edu.cn).

Chao Yao is with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China (e-mail: yaochao@ustb.edu.cn).

Jian Jin is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: jian.jin@ntu.edu.sg).

Tammam Tillo is with the Institute of Digital Media and Communication, Taiyuan University of Science and Technology, Taiyuan 030024, China (e-mail: tammam.tillo2@gmail.com).

Digital Object Identifier 10.1109/TBC.2025.3635389

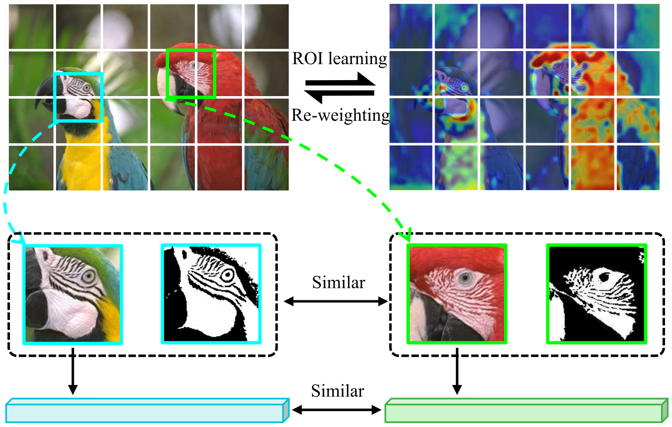

Fig. 1. An example diagram of image redundant information based on our SFTNet. The complex area encoded is depicted using a heat map in the upper right corner. A pair of patches with long-distance redundancy is shown at the bottom.

using the fast fourier transform. Within FATB, intra-block frequency attention and inter-block spatial attention are devised to enhance local high-frequency details and model cross-block low-frequency dependencies, respectively. As shown in Fig. 1, intra-block frequency attention extracts low-frequency contours (e.g., parrot heads marked by blue/green boxes) as structural guidance to selectively enhance high-frequency details (e.g., feather textures) while suppressing redundant background regions. Inter-block spatial attention identifies redundant low-frequency features across patches (e.g., bottom parrot head pairs) and reuses these features to maintain cross-block structural coherence, avoiding boundary artifacts.

To further optimize compression efficiency, the Feature Re-weighting Strategy (FRS) is introduced to optimize bit allocation by generating a content-adaptive weight matrix. This matrix prioritizes complex spatial-frequency regions (highlighted by the heat map in the upper right of Fig. 1), ensuring targeted bit allocation for detail preservation while enabling efficient compression. Experimental results show that our SFTNet outperforms both recent LIC methods and VVC Intra [4] on Kodak and CLIC datasets, validating our superior compression performance. Our contributions can be summarized as follows:

- We propose the Spatial-Frequency Transformer Network (SFTNet) for LIC to model spatial-frequency correlations and optimize bit allocation for the high-frequency detail reconstruction and low complexity.
- We design the Frequency-aware Transformer Block (FATB) to enhance high-frequency details via low-frequency guidance and model low-frequency cross-block dependencies for global structural consistency.
- We present the Feature Re-weighting Strategy (FRS) to dynamically weight features by regional complexity, prioritizing bit allocation to preserve texture details while compressing redundant areas.
- Experimental results demonstrate that our SFTNet outperforms VVC by -6.55% and -6.18% in BD-Rate on Kodak and CLIC datasets, and excels in high-frequency detail reconstruction quality compared to recent LIC methods.

The rest of this paper is structured as follows. Related works are reviewed in Section II. An architecture of SFTNet is detailed in Section III. Section IV describes the experimental settings and results. Section V draws the conclusion.

II. Related Works

A. Learned Image Compression

The mainstream architectures of learned image compression (LIC) can be classified into three categories: convolutional neural networks (CNNs), Transformer architectures, and the hybrid CNN-Transformer architectures. Early CNN frameworks, such as the variational autoencoder (VAE) [28], adapted for compression by Ballé et al. [15], enabled end-to-end optimization of nonlinear transformations. However, because these frameworks relied on convolutional operations, which had a local receptive field, they struggled to model global structural correlations in complex textures. This limitation led to issues such as detail loss and structural incoherence in reconstructed images [18]. Similarly, multiscale transformation networks [29] and decoder-side enhancement subnetworks [19] aimed to enhance multi-level feature representation. However, they relied on CNNs with inherently limited receptive fields, leaving them unable to maintain global structural consistency during reconstruction. Transformer architectures enhanced global modeling via self-attention mechanisms. For example, TIC [25] and Entroformer [9] leveraged these mechanisms to aggregate contextual information, thereby strengthening structural coherence. However, self-attention mechanisms split images into uniform blocks to generate tokens, leading to a uniform encoding strategy. This strategy applied the same processing scale to both complex textures and smooth backgrounds, increasing computational load for high-resolution inputs [30]. Meanwhile, uniform partitioning was unable to adapt to the fine-grained features and failed to capture subtle local differences as precisely as convolutional operations [8]. To mitigate this computational issue, LALIC [14] introduced a novel approach by applying linear attention-based RWKV models to address the high computational complexity of traditional Transformer-based LIC methods. Although existing adaptive bit allocation methods [31], [32] attempted to adjust coding resources based on spatial complexity, these methods failed to establish an effective collaborative mechanism between local details and global modeling.

Liu et al. [11] proposed a hybrid CNN-Transformer block to balance the local efficiency of CNNs and the global modeling capability of Transformers, but high computational costs for high-resolution inputs remained a major challenge. To address this issue, Wang et al. [33] integrated Swin Transformer with convolutional layers into an end-to-end framework. Through hierarchical window attention, they demonstrated that hybrid architectures could balance efficiency and performance. Building on this idea, Jiang et al. [10] advanced the hybrid approach by embedding CNN-Transformer attention modules into the entropy model, dynamically fusing channel and spatial contexts to optimize bit allocation for texture regions. Additionally, Xu et al. [34] proposed a spatial-channel hybrid framework that combined residual blocks and window-based channel attention for local-global modeling, while using wavelet transform to expand receptive fields. Similarly, DCAE [35] innovatively employed a dictionary-based cross-attention entropy model to enhance prior information for entropy modeling. Existing hybrid models mostly focus on processing in the spatial domain and fail to leverage the property that different frequencies in the frequency domain correspond to different features, resulting in the aliasing of frequency information and impairing the efficiency of representation.

B. Frequency-Based Image Compression

The traditional image codecs (e.g., JPEG [1], JPEG2000 [2]) employed fixed transforms for frequency decomposition. Their basis functions and decomposition methods were preset and uniform, failing to adapt dynamically to image content and struggling to balance representation efficiency between complex and simple regions. Subsequently, LIC methods explored adaptive frequency processing to address such limitations. Ye et al. [36] proposed reparameterizing convolutions as linear combinations of DCT kernels to enhance frequency-domain representation. Dynamic textures exhibit rapidly changing details and frequencies, and fixed DCT bases struggle to handle these rapid changes. This is because they lack adaptability to the spatially varying frequency characteristics of such textures, resulting in detail loss and blurring during compression. Frequency-decomposition window attention and error-variance adaptive partitioning were introduced by Li et al. [37] and Rhee et al. [38], respectively, to optimize frequency component segmentation via dynamic mechanisms. These methods aligned more closely with content characteristics than early approaches relying on fixed-transform paradigms [1], [2]. However, it was difficult for static thresholding and directional decomposition to cope with the differences of frequency distribution in complex textures. Mishra et al. [39], Fu et al. [40], and Xu et al. [34] used wavelet transforms to achieve multi-scale frequency decomposition for more refined frequency-domain characterization. However, predefined wavelet bases struggle to dynamically adapt to the local texture features of images, and their inefficiency in high-frequency coding leads to suboptimal performance in complex scenes.

III. Proposed Method

This section details the architecture of the Spatial-Frequency Transformer Network for LIC (SFTNet). It is a frequency-aware encoder-decoder framework for spatial-frequency redundancy reduction. The Frequency-aware Transformer Block (FATB) is then presented, which employs intra-block frequency attention to enhance the local structure-guided details and inter-block spatial attention to model global cross-block dependency. Finally, the Feature Re-weighting Strategy (FRS) is proposed to adaptively allocate bits to complex blocks by modulating feature importance and prioritizing critical textures while suppressing background redundancies.

A. Overall Architecture

The overall architecture of our SFTNet is illustrated in Fig. 2. Given an input image x, the encoder network transforms x into a latent representation feature y to reduce the spatial redundancy among image features, which is formulated as:

$$y = Enc(x; \phi), \tag{1}$$

where $Enc(\cdot)$ represents the encoder network, ϕ represents the trainable parameters of the encoder network. In the encoder network, the Frequency-aware Transformer Block (FATB) is introduced to decompose features into low-frequency and high-frequency components using fast fourier transform (FFT). In FATB, a dual attention mechanism is employed to jointly model spatial-frequency dependencies. Intra-block frequency attention leverages low-frequency components as structural guidance to enhance high-frequency details within spatial blocks, while inter-block spatial attention models cross-block dependencies in low-frequency components to ensure the global structural consistency.

Before quantization, the importance of the latent feature y is adaptively adjusted by the Feature Re-weighting Strategy (FRS) through leveraging the frequency-domain characteristics. Since different frequency components in feature y contribute differently to image reconstruction, FRS re-weights features to prioritize critical ones, ensuring more quantization bits are allocated to important components. This optimizes bit allocation for the quantized latent representation $\hat{y}$, which is formalized as:

$$\hat{y} = Q(Rw(y)), \tag{2}$$

where $Q(\cdot)$ denotes the generic quantization function and $Rw(\cdot)$ denotes the re-weighting strategy of FRS. The quantized feature $\hat{y}$ is then compressed into a bitstream by an arithmetic encoder (AE).

In the decoder, the bitstream is first decompressed to recover $\hat{y}$ by an arithmetic decoder (AD). The decoder network then reconstructs the image $\hat{x}$ from $\hat{y}$, which is formulated as:

$$\hat{x} = Dec(\hat{y}; \theta), \tag{3}$$

where $Dec(\cdot)$ denotes the decoder network, θ denotes the trainable parameters. Within the decoder, FATBs reconstruct low-frequency structures and high-frequency details by fusing multi-scale contextual information and cross-block dependencies. This dual-path design in FATB ensures critical detail preservation through collaborative frequency-spatial integration for effective reconstruction of both global image structure and local texture details.

B. Frequency-Aware Transformer Block (FATB)

The Frequency-aware Transformer Block (FATB) achieves global-local collaborative compression via joint frequency-spatial domain modeling, as shown in Fig. 3. First, the input feature $f_{in} \in \mathbb{R}^{H \times W \times C}$ is transformed into the frequency domain via Fast Fourier Transform (FFT) [41]. Its inherent properties match the three core needs of FATB: (1) dynamically capturing cross-regional high-frequency correlations, (2) preserving high-frequency energy to guide low-frequency reconstruction, (3) ensuring artifact-free reversibility [42], [43], [44], [45]. To target the global structures and local details, the frequency-domain feature $f'_{in} \in \mathbb{C}^{H \times W \times C}$ is decomposed into low-frequency f_l and high-frequency components f_h.

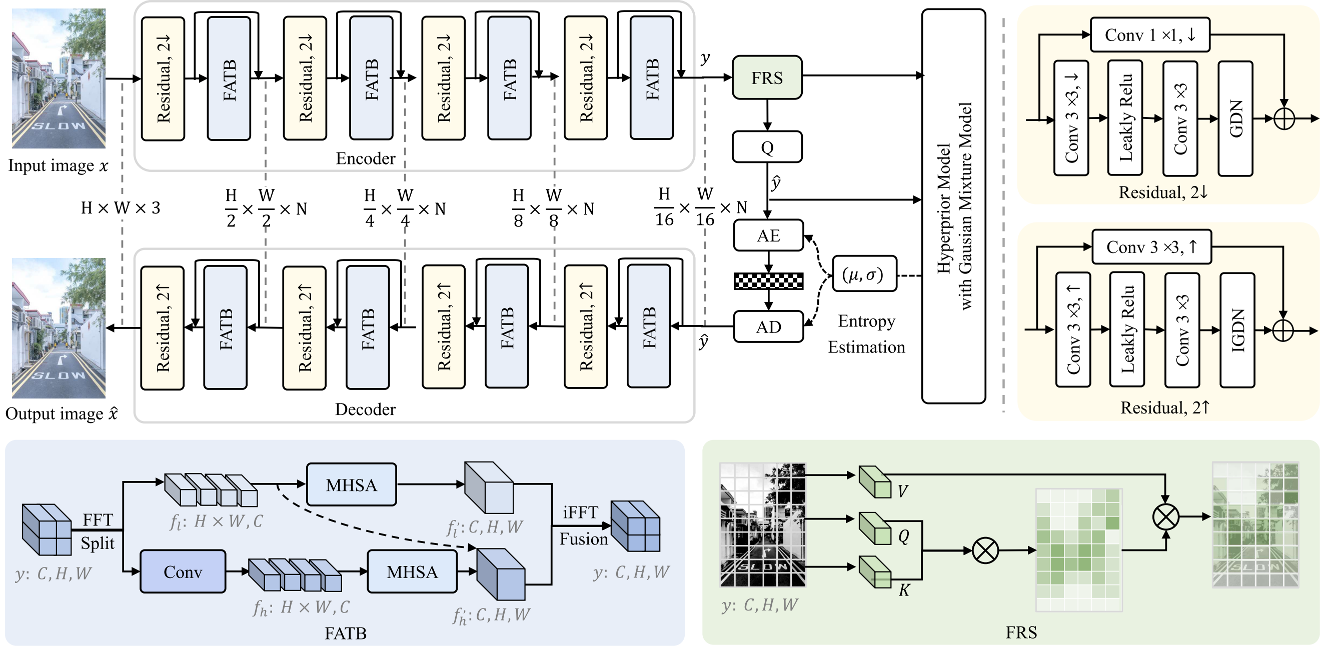

Fig. 2. The overview framework of our SFTNet. FRS represents the Feature Re-weighting Strategy. FATB represents the Frequency-aware Transformer Block. Q, AE, and AD represent the quantization, arithmetic encoder, and arithmetic decoder, respectively. ↑ and ↓ denote the up-sampling and down-sampling operations. The details of the residual blocks are depicted in the yellow box.

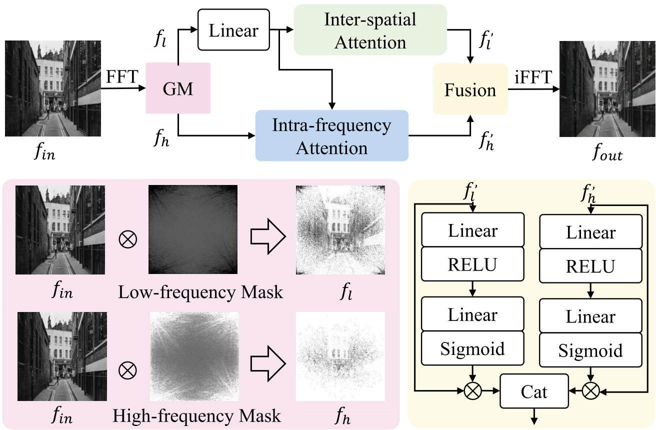

Fig. 3. The structure of the proposed Frequency-aware Transformer Block (FATB). The gating mask distinguishes high-frequency and low-frequency information by generating adaptive masks, as depicted in the pink box. The yellow block is the fusion module of high-frequency and low-frequency features.

Guided by adaptive masks, the decomposition is formulated as:

$$\begin{aligned} f_l &= f_{in}' \odot m_l, \\ f_h &= f_{in}' \odot m_h, \end{aligned} \tag{4}$$

where m_l and m_h denote adaptive masks for low-frequency components and high-frequency components, respectively. "$\odot$" denotes element-wise multiplication.

The key to these masks lies in a content-adaptive energy threshold, which ensures their alignment with the frequency distribution of specific images. The threshold f_{th} is utilized to separate the low and high frequencies, and is adaptively set based on the energy distribution of the FFT-transformed input feature fin', which is formulated as:

$$f_{\text{th}} = \arg\min_f \left\{ \sum_{\substack{0\le u\le f\\ 0\le v\le f}} |f_{\text{in}}'(u,v)|^2 \ge r_e \times \sum_{u=0}^{H-1}\sum_{v=0}^{W-1} |f_{\text{in}}'(u,v)|^2 \right\}, \tag{5}$$

where arg $\min_f$ represents the operator to find the f that minimizes the condition in the curly braces, (u,v) represents horizontal and vertical coordinates in the frequency domain of f_{in}', $|f_{\text{in}}'(u,v)|^2$ represents the energy of the frequency component at (u,v), r_e represents the energy ratio that balances the low-frequency structure and high-frequency details. The masks are formulated as follows:

$$m_l(u,v) = \begin{cases} 1, & u^2+v^2 \le f_{\text{th}}^2 \\ 0, & \text{otherwise} \end{cases}, \tag{6}$$

$$m_h(u,v) = 1 - m_l(u,v), \tag{7}$$

where u^2+v^2 represents the frequency magnitude of f_{in}' at coordinate (u,v), which ensures m_l covers low-frequency regions and m_h covers high-frequency regions.

After decomposing features into the low-frequency components f_l and high-frequency components f_h, FATB computes high-frequency energy map E_h and low-frequency dependency map D_l. For each patch, $E_h = \sum_{u,v} |f_h(u,v)|^2$ (sum of squared magnitudes of high-frequency coefficients), which quantifies the detail richness of the patch. For each pair of patches (i,j), $D_l(i,j) = \text{cosine similarity}(f_l^i, f_l^j)$, which measures the structural consistency between distant low-frequency patches (avoiding redundant bit allocation for similar structures).

1) Intra-Block Frequency Attention: Within each spatial block, the low-frequency component guides the high-frequency processing to address the inherent limitations of

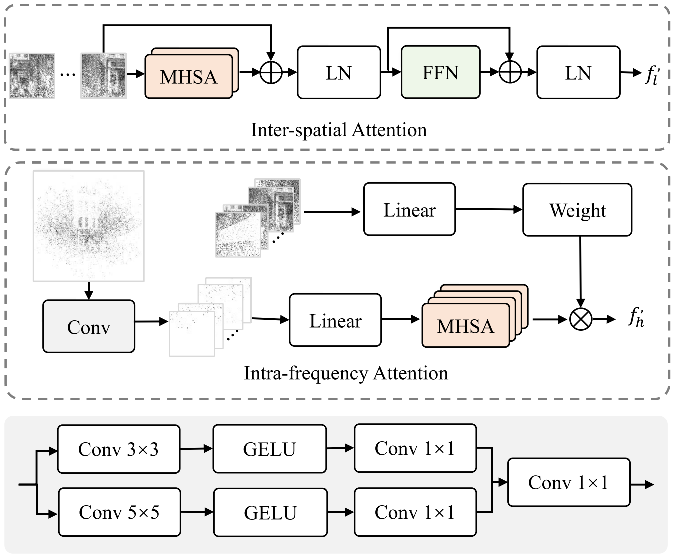

Fig. 4. The structure of the intra-block frequency attention and the inter-block spatial attention in FATB.

high-frequency details in isolation, as shown in Fig. 4. The high-frequency features, though rich in local textures and edges, lack awareness of the global structural context, which can lead to inconsistencies with the overall image structure. First, the input feature map ($f_{\text{in}} \in \mathbb{R}^{H\times W\times C}$) is split into non-overlapping 8×8 patches to balance detail preservation and computational efficiency. The number of patches along the height is $N_h = H/8$, and the number along the width is $N_w = W/8$, resulting in a total of $K = N_h \times N_w$ patches. The spatial indices of the global patch $k \in [0, K-1]$ are defined as $k_x = k \mod N_h$ and $k_y = k/N_h$, where $k_x \in [0, N_h - 1]$ and $k_y \in [0, N_w - 1]$. The spatial coordinates of the k-th patch are $y_{\text{start}} = k_x \times 8$, $y_{\text{end}} = (k_x + 1) \times 8$, $x_{\text{start}} = k_y \times 8$, $x_{\text{end}} = (k_y + 1) \times 8$. This patch is extracted as $f_{\text{in}}[y_{\text{start}} : y_{\text{end}}, x_{\text{start}} : x_{\text{end}}, :] \in \mathbb{R}^{8\times 8\times C}$ and then fed into the intra-block frequency attention to enhance the high-frequency details. The high-frequency feature f_h is processed by parallel multi-scale convolutional layers to preserve key high-frequency patterns. The smaller kernel captures fine details by focusing on local areas, while the larger kernel models broader patterns by integrating neighboring context [46]. Next, the multi-scale representation $f_{h,multi}$ is divided into patches and embedded into a sequence $f_{h,seq}$, with positional information preserved to retain spatial location information. Multi-Head Self-Attention (MHSA) is then applied to $f_{h,seq}$ to model dependencies, generating the feature $f_{h,\text{mhsa}}$, which is formulated as:

$$f_{h,mhsa} = MHSA(f_{h,seq}), \tag{8}$$

where $f_{h,seq}$ denotes the embedded high-frequency block sequence, $MHSA(\cdot)$ represents the multi-head attention operation. By using multiple attention heads, this step learns diverse dependencies across the sequence, reinforcing connections between semantically related blocks.

To align with low-frequency guidance, the low-frequency component f_l is first projected through a linear layer to form a preliminary guidance weight map. This map is adjusted via a 1×1 convolution and sigmoid function to match the dimension and value range of $f_{h,\text{mhsa}}$, yielding the final guidance weight map W.

The refined high-frequency component f_h' is obtained by element-wise multiplication of $f_{h,\text{mhsa}}$ with the guidance weight map W, adaptively emphasizing critical high-frequency details, which is formulated as:

$$f_h' = f_{h,mhsa} \odot W, \tag{9}$$

where "$\odot$" denotes element-wise multiplication. This ensures high-frequency details are reinforced where they align with the global structure and suppressed where they conflict, balancing local detail richness with global structural coherence.

2) Inter-Block Spatial Attention: Inter-block spatial attention resolves cross-block discontinuities by modeling long-range dependencies among low-frequency components, as shown in Fig. 4. The low-frequency feature f_l is tokenized into continuous spatial blocks, and frequency-domain positional encoding $Pos(u, v)$ (where (u, v) are spatial frequency coordinates) is added to form the sequence $f_{l,\text{seq}}$. This process preserves low-frequency correlations critical for capturing distant structural dependencies. The positional encoding $Pos(u, v)$ also provides frequency-based spatial cues, letting the attention mechanism distinguish spatial block positions clearly.

The sparse global MHSA is applied to $f_{l,\text{seq}}$ to model long-range dependencies. The sparse masks are learned via a gating network that predicts a subset $\mathcal{K}_i$ for each block i. Instead of global all-to-all attention, each block attends to a learnable subset $\mathcal{K}_i$, generating the block-specific low-frequency feature $f^i_{l,\text{mhsa}}$, which is formulated as:

$$f^i_{l,mhsa} = \sum_{j\in\mathcal{K}_i} A_{i,j} \cdot V(f^j_{l,seq}), \tag{10}$$

where $A_{i,j}$ represents the attention weights quantifying relevance between blocks i and j, and $V(\cdot)$ denotes a value projection layer that operates on magnitude and phase components of $f^j_{l,\text{seq}}$, $\mathcal{K}_i$ denotes the key blocks subset. This sparse selection reduces computational complexity while retaining critical long-range dependencies. A residual connection integrates the FFN output $f_{l,\text{ffn}}$ with the sparse MHSA output $f_{l,\text{mhsa}}$ to yield the refined low-frequency component f_l', formulated as:

$$f_l' = f_{l,ffn} + f_{l,mhsa}, \tag{11}$$

where $f_{l,ffn}$ denotes the FFN output feature, $f_{l,\text{mhsa}}$ denotes the attention-enhanced feature capturing long-range dependencies. This residual connection retains the original low-frequency structural information while integrating learned long-range dependencies, ensuring global consistency in the frequency domain.

Finally, the low-frequency feature f_l' and the high-frequency feature f_h' are fused and transformed into the spatial domain to yield the output feature f_{out}, which is formulated as:

$$f_{out} = iFFT(fusion(f_l', f_h')), \tag{12}$$

where f_l' and f_h' represent the low-frequency and high-frequency component after branch processing, $fusion(\cdot)$ represents the feature fusion operation. $iFFT(\cdot)$ represents

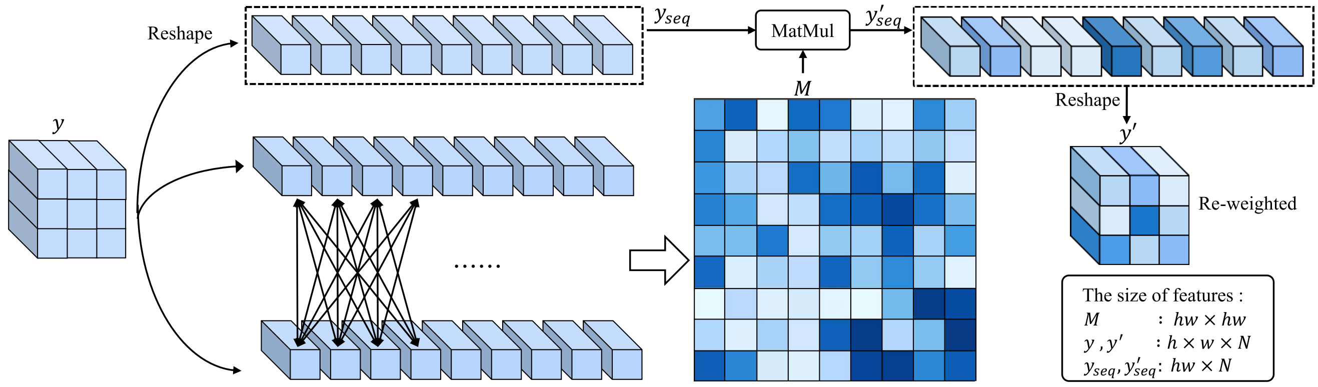

Fig. 5. The depiction of the feature re-weighting process, where M represents the spatial-frequency importance weights and the MatMul represents matrix multiplication. Different levels of blue indicate different weight values.

Inverse FFT (IFFT), transforming the features from the frequency domain to the spatial domain.

C. Feature Re-Weighting Strategy (FRS)

To address the uneven coding complexity of image regions, the Feature Re-weighting Strategy (FRS) adaptively sets compression priorities by integrating FATB spatial attention and latent feature high-frequency energy. As shown in Fig. 5, FRS processes the latent feature map $y \in \mathbb{R}^{H\times W\times C}$. First, the feature y is flattened into a 1-D sequence $y_{seq} \in \mathbb{R}^{HW\times C}$, which is formulated as:

$$y_{seq} = R_{2D\longrightarrow 1D}(y), \tag{13}$$

where $y_{seq} = (y_1, y_2, \cdots, y_{HW})$ represents a sequence of feature tokens, $R_{2D\longrightarrow 1D}(\cdot)$ reshapes the 2D feature y into 1D feature y_{seq}. Next, FRS incorporates E_h and D_l, ensuring patches with high E_h (rich details) and low D_l (unique structure) receive higher weights for bit allocation. A weight matrix $M \in \mathbb{R}^{HW\times HW}$ is computed via a self-attention mechanism, where the attention score $S_{i,j}$ between tokens i and j, which is formulated as:

$$S_{i,j} = \frac{\vec{q}_i \cdot \vec{k}_j}{\sqrt{d_k}} + B_{i,j} + \alpha \cdot E_h^j + \beta \cdot (1 - D_l(i,j)), \tag{14}$$

where $\vec{q}_i$ denotes the query vector corresponding to token i, $\vec{k}_j$ denotes the key vector corresponding to token j, d_k represents the dimension of the key vector, $B_{i,j}$ represents a learnable relative positional bias, and α and β represent the hyperparameters balancing spatial and frequency cues. $\alpha = 0.4$ and $\beta = 0.3$ represent the optimized values via ablation. The weight matrix $M_{i,j}$ is then updated with the enhanced score $S_{i,j}$ and directly guide the process of bit allocation, which is formulated as:

$$M_{i,j} = \frac{\exp(S_{i,j})}{\sum_{k=1}^{HW} \exp(S_{i,k})}, \tag{15}$$

where $\exp(\cdot)$ represents the exponential function for non-negativity, $S_{i,j}$ represents the attention score between token i and j, and HW represents the total number of tokens $H \times W$. The feature sequence $y'_{seq} \in \mathbb{R}^{HW\times C}$ is re-weighted by aggregating the value vectors v_j with the learned weights $M_{i,j}$, which is formulated as:

$$y'_{seq} = \sum_{j=1}^{HW} M_{i,j} v_j, \tag{16}$$

where $M_{i,j}$ represents the attention weight from token i to token j, and v_j represents the value vector of token j. After feature re-weighting, the 1-D sequence y'_{seq} is reshaped back into the 2-D format to generate the final latent feature $y' \in \mathbb{R}^{H\times W\times C}$, which is formulated as:

$$y' = \mathcal{R}_{1D\to 2D}(y'_{seq}), \tag{17}$$

where $\mathcal{R}_{1D\to 2D}(\cdot)$ represents the inverse operation of $\mathcal{R}_{2D\to 1D}$, restoring the spatial dimensions of the feature map. The re-weighted latent feature y' is quantized to $\hat{y}$ for arithmetic encoding, which is formulated as:

$$\hat{y} = Q(y'), \tag{18}$$

where $Q(\cdot)$ represents the quantization function.

D. Loss Function

The optimization goal of the network is to minimize both the bitrate and the distortion simultaneously. The bitrate represents the size of the bitstream, which is directly related to the storage and transmission cost. The distortion, on the other hand, measures the information loss between the original image x and the reconstructed image $\hat{x}$. To achieve this dual-objective optimization, we define the loss function as:

$$L = R + \lambda D(x, \hat{x}), \tag{19}$$

where R denotes the number of bits required for encoding latent feature $\hat{y}$, D represents the compression distortion, and λ denotes a hyperparameter to balance the rate and distortion.

IV. Experiments

A. Experimental Settings

1) Datasets: Our models are trained on a subset of the ImageNet dataset [47], comprising 13,600 images. All images are randomly cropped into 256×256 patches for training, with a batch size of 8. For evaluation, we use the Kodak dataset

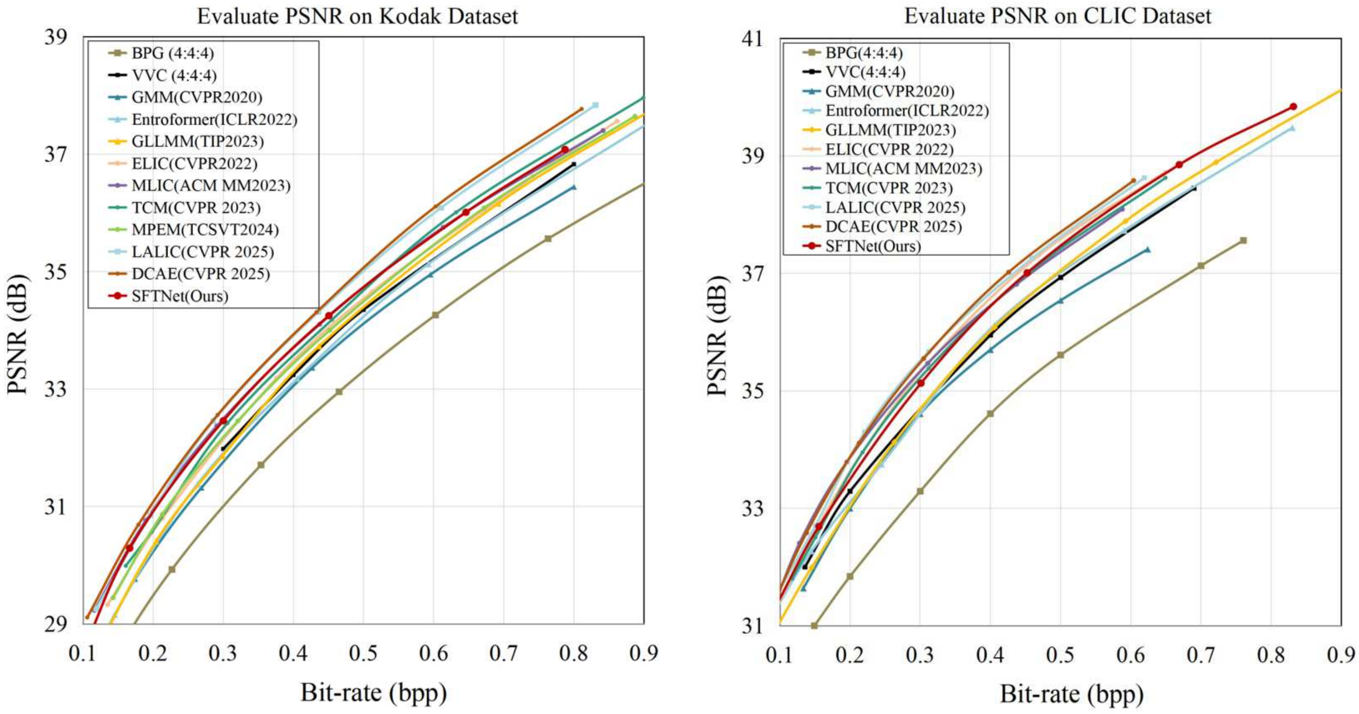

Fig. 6. Rate-Distortion performance with MSE-based loss on Kodak and CLIC datasets.

[48], which has 24 images at 512 × 768 resolution, and the CLIC professional validation dataset [49], which includes 41 images at 2048 × 1370 resolution.

2) Experimental Details: The Adam optimizer [50] with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\sigma = 1 \times 10^{-8}$ is utilized. All models are trained for 2,000,000 iterations with a batch size of 16. The initial learning rate is set to 1×10^{-4} and linearly decayed to 1×10^{-5} over the final 100,000 iterations. Two distortion terms are used as the loss function, i.e., mean square error (MSE) and multi-scale structure similarity (MS-SSIM). The Lagrangian multiplier λ is selected from {0.0016, 0.0032, 0.0075, 0.015, 0.03, 0.045} for MSE and $\{3, 12, 40, 120\}$ for MS-SSIM. Different λ values correspond to the model with different bitrates, which balance between the rate and distortion. All models have 192 filters and are trained on a single NVIDIA GeForce RTX 3090 Ti GPU.

B. Performance Comparison Results

1) Rate-Distortion Performance: We compare the rate-distortion (R-D) performance against the traditional compression codecs such as BPG [3] and VVC [4], as well as LIC methods like GMM [20], GLLMM [12], Entroformer [9], ELIC [51], MLIC [10], TCM [11], MPEM [52], DCAE [35] and LALIC [14]. The evaluation results are sourced as follows: (a) the results of GMM [20], ELIC [51], MLIC [10] and TCM [11] are obtained through independent training and testing by our team. (b) the results of Entroformer [9], DCAE [35] and LALIC [14] are derived from their pre-trained models. (c) due to the lack of open-source code, the results of MPEM [52] and GLLMM [12] are extracted from the original paper.

The R-D performance in terms of PSNR on the Kodak and CLIC datasets is presented in Fig. 6. The results show that MPEM [52] achieves slightly lower PSNR values compared to our SFTNet, while Entroformer [9] -another Transformer-based image compression method-exhibits inferior performance. Although SFTNet demonstrates marginally lower PSNR values than DCAE [35] and LALIC [14] at certain bitrates, we further evaluate perceptual quality using the MS-SSIM metric as shown in Fig. 7, where our SFTNet significantly surpasses many counterparts in perceptual detail preservation. The MS-SSIM values range from 0 (the worst) to 1 (the best), and most of the results are above 0.9. Therefore, we convert the MS-SSIM values into decibels $(-10\log_{10}(1 - MS\text{-}SSIM))$ to clarify gaps between curves. The results indicate that our SFTNet significantly demonstrates competitive advantages in MS-SSIM, surpassing many counterparts in perceptual detail preservation.

To compare R-D performance more accurately in numerical terms, we further calculated the BD-PSNR and BD-rate, as shown in TABLE I, with all results benchmarked against VVC. MLIC [10] achieves a BD-PSNR of 0.41 dB and a BD-rate of −7.16% on the Kodak dataset, along with 0.43 dB and −10.11% on the CLIC dataset, showing superior R-D performance. TCM [11] also delivers competitive results on Kodak, with a BD-PSNR of 0.34 dB and a BD-rate of −6.54%, demonstrating notable gains in reconstruction quality and bitrate savings. In contrast, our SFTNet achieves a BD-PSNR of 0.31 dB and a BD-rate of −6.55% on Kodak, and 0.27 dB with −6.18% on CLIC. While MLIC [10] outperforms our SFTNet by approximately 0.1 dB in BD-PSNR and 4% in BD-rate on CLIC, and TCM [11] edges SFTNet slightly in BD-PSNR by 0.03 dB on Kodak, SFTNet compensates with significant advantages in computational efficiency. Specifically, our encoding time (116 ms) and decoding time (91 ms) are

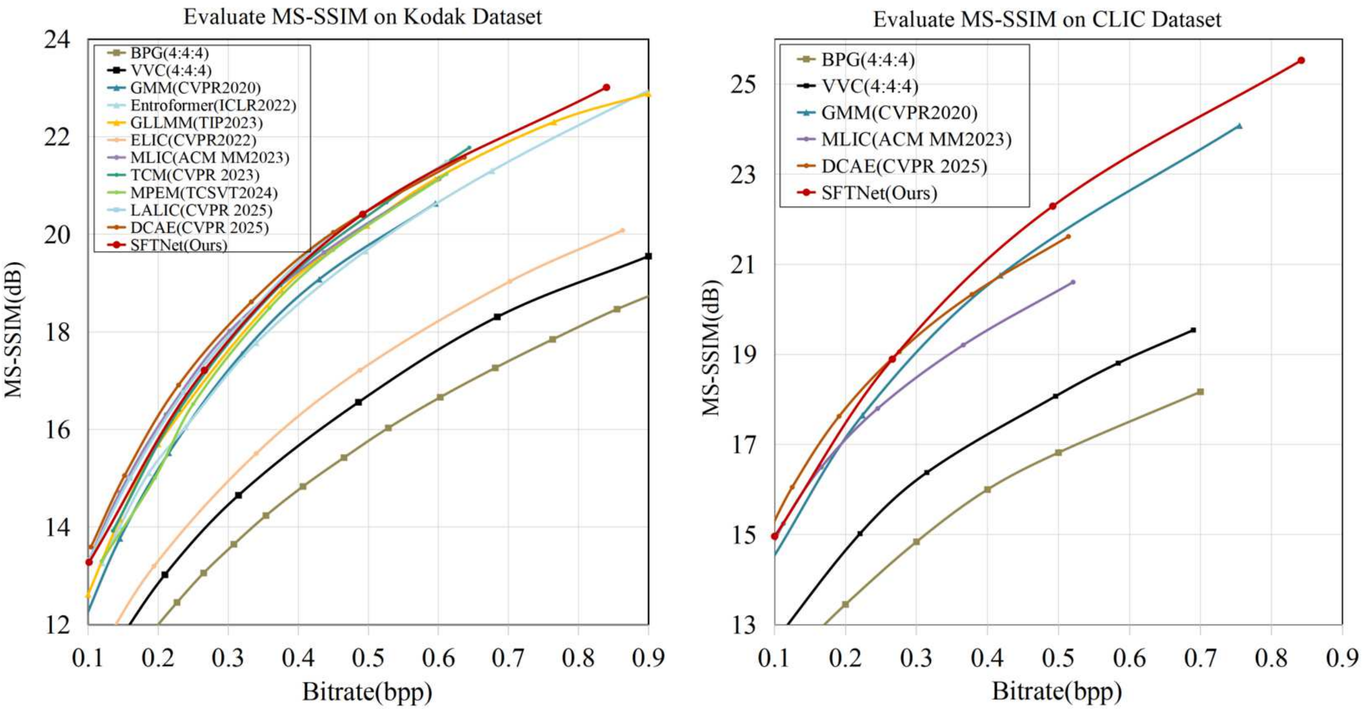

Fig. 7. Rate-Distortion performance with SSIM-based loss on Kodak and CLIC datasets.

TABLE I
THE RATE-DISTORTION (RD) PERFORMANCE AND COMPUTATIONAL COMPLEXITY COMPARISON OF LEARNED IMAGE COMPRESSION MODELS VS. VVC ON KODAK DATASET. LOWER BD-RATE VALUES INDICATE BETTER RD PERFORMANCE. ALL THE TESTS ARE CONDUCTED ON A SINGLE NVIDIA GEFORCE RTX 3090 TI GPU

Methods	Param. (M)	FLOPs (G)	Enc.Time. (ms)	Dec.Time. (ms)	BD-PSNR (dB)	PSNR-based BD-Rate (%)	MS-SSIM-based BD-Rate (%)	BD-MS-SSIM (dB)
GMM (CVPR ’20)	13.18	30	3708	8658	-0.21	4.94	-44.77	2.75
GLLMM (TIP ’23)	15.71	- -	385260	387620	-0.07	2.55	-49.36	3.29
ELIC (CVPR ’22)	41.9	332	91	119	0.23	-3.95	-11.62	0.52
Entroformer (ICLR ’22)	187.69	- -	4768	85919	-0.19	4.70	-44.95	2.97
MLIC (ACM MM ’23)	83.27	443	190	268	0.41	-7.16	-52.01	3.37
TCM (CVPR ’23)	75.9	415	154	107	0.34	-6.54	-50.61	3.41
MPEM (TCSVT ’24)	34.0	88	143	191	0.13	-1.88	-46.50	2.94
DCAE (CVPR ’25)	119.4	273	133	146	0.70	-13.57	-53.37	3.58
LALIC (CVPR ’25)	63.24	286	274	150	0.63	-12.16	-52.56	3.46
SFTNet (Ours)	45.78	208	116	91	0.31	-6.55	-52.88	3.55

substantially faster than MLIC [10] (190 ms encoding / 268 ms decoding) and TCM [11] (154 ms encoding / 107 ms decoding). SFTNet has only 45.78 M parameters, 208 GFLOPs, and 116 ms encoding time, demonstrating much lower complexity while achieving RD performance comparable to models like TCM [11].

While DCAE [35] and LALIC [14] exhibit superior overall BD-Rate performance, SFTNet maintains competitive high-frequency reconstruction capabilities with significant complexity advantages. Specifically, SFTNet achieves 61.6% and 27.6% fewer parameters, 23.8% and 27.3% lower FLOPs, and 12.8% and 57.7% shorter encoding times compared to DCAE [35] and LALIC [14], respectively. These results substantiate the core value of SFTNet in effectively balancing computational complexity with high-frequency detail reconstruction.

2) Subjective Performance: In order to compare the subjective performance, we reconstruct the images of our SFTNet,

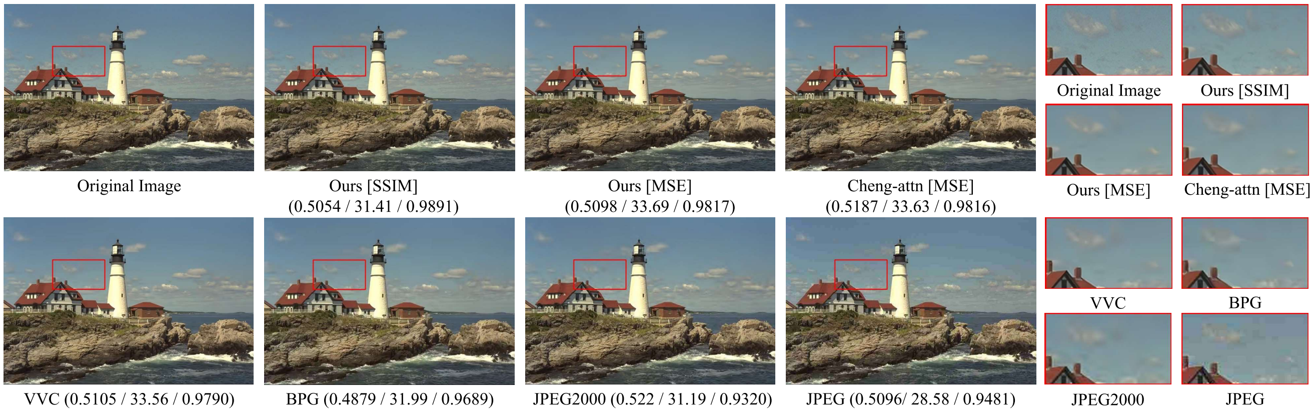
Original Image
Ours [SSIM] (0.5054 / 31.41 / 0.9891)
Ours [MSE] (0.5098 / 33.69 / 0.9817)
Cheng-attn [MSE] (0.5187 / 33.63 / 0.9816)
Original Image
Ours [SSIM]
Ours [MSE]
Cheng-attn [MSE]
VVC (0.5105 / 33.56 / 0.9790)
BPG (0.4879 / 31.99 / 0.9689)
JPEG2000 (0.522 / 31.19 / 0.9320)
JPEG (0.5096/ 28.58 / 0.9481)
VVC
BPG
JPEG2000
JPEG

Fig. 8. The subjective reconstructed images of different methods. The three values in each triplet under the image are bpp/PSNR (dB)/MS-SSIM. The enlarged details are displayed on the right side.

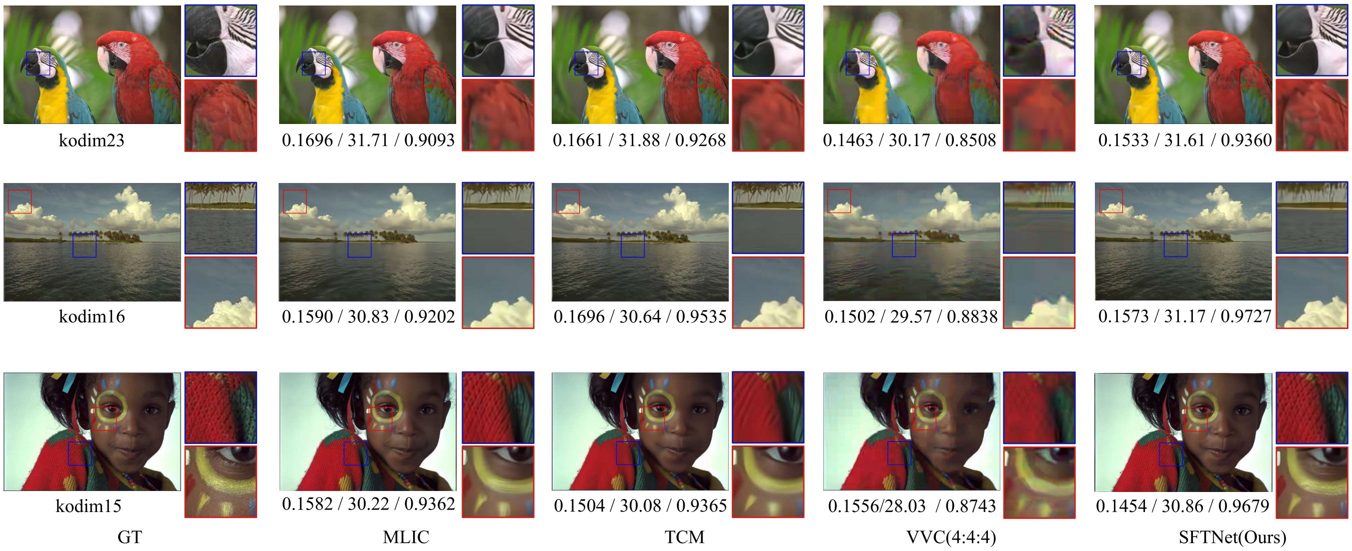

Fig. 9. The subjective reconstructed results of local magnification effects in edge/texture regions. The values of bpp, PSNR (dB) and MS-SSIM are shown under the images. The enlarged details are displayed on the right side.

LIC network Cheng-attn [20] and the traditional standards (i.e. JPEG [1], JPEG2000 [2], BPG [3] and VVC Intra [4]), as shown in Fig. 8. Images with approximately 0.5 bpp are selected for reconstruction comparison, as this bitrate effectively reflects performance under comparable compression levels. The reconstructed images of different methods are presented, and a representative region in the images is enlarged to observe the reconstruction details more meticulously.

VVC [4] introduces slight blur in cloud regions, while JPEG [1], JPEG2000 [2], and BPG [3] produce more pronounced blur with visible blocky or wavy artifacts. Cheng-attn [20], as a learned method, outperforms traditional standards but still exhibits relatively blurry cloud edges with some detail loss. In contrast, our SFTNet preserves image fidelity and clarity more effectively during reconstruction, excelling in both overall visual quality and fine details. When optimized with MSE loss, our SFTNet shows significant advantages in visual perception. Taking the edge part of the clouds in the figure as an example, the cloud edges reconstructed by our SFTNet are clearer and sharper. In addition, SFTNet delivers a more visually pleasing result when trained with MS-SSIM loss.

To further provide intuitive qualitative support for the high-frequency detail restoration advantage of SFTNet, we extend the subjective experiment by adding local magnification comparisons targeting high-frequency-dense regions (edge/texture) of the Kodak dataset. We select a lower bitrate of 0.15 bpp to compare SFTNet with MLIC [10], TCM [11] and VVC [4] as shown in Fig. 9.

For kodim23, MLIC and TCM blur the fine feather branches, and VVC introduces block artifacts in feather texture. Our SFTNet retains the clear details of feather. For kodim15, MLIC and TCM over-smooth skin texture while VVC introduces obvious artifacts, and SFTNet preserves clear skin and clothing details. For the cloud edges in kodim16, SFTNet also outperforms MLIC, TCM and VVC in edge sharpness and texture integrity. These results suggest the significant advantage of SFTNet in restoring fine details.

3) Feature Visualizations: To verify the capability of our SFTNet in learning high-frequency regions, the extracted features of SFTNet and MLIC [10] are visualized, as shown in Fig. 10. Both models are trained with the same number of iterations using MSE loss with the same λ. MLIC [10]

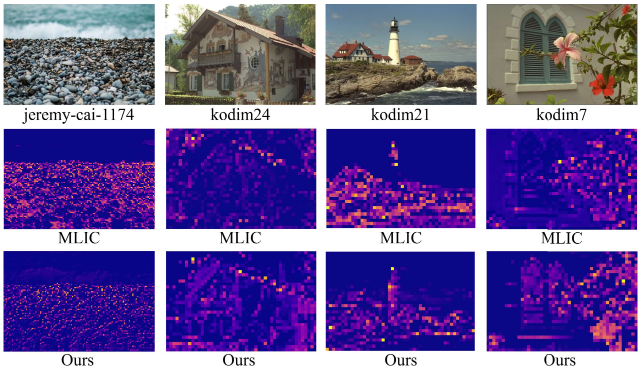

Fig. 10. The visualization comparison results of multi-scale features, where the yellow and brighter regions indicate the stronger signal and the purple and darker regions indicate the weaker signal.

TABLE II

Ablation Study Results of FATB and FRS Collaboration on the Kodak Dataset

Modules	FATB	FRS	BD-PSNR (dB)	BD-Rate (%)	Parameters (M)
Module 1	✓		0.25	-5.88	34.38
Module 2		✓	0.27	-5.92	28.04
Module 3 (SFTNet)	✓	✓	0.31	-6.55	45.78

has more extensive purple-darker regions (weaker signals) in feature maps, revealing its weaker focus on high-frequency details. This sparse yellow-brighter area distribution leads to inefficient encoding of image details, as key high-frequency info isn't fully captured. In contrast, our SFTNet presents prominent yellow-brighter regions in intermediate layers, focusing strongly on high-frequency areas while using fewer bits for final latent features. This confirms FATB effectively captures high-frequency details, yielding a compact representation where the retained rich high-frequency information is key to maintaining image quality during compression.

C. Ablation Studies

1) FATB & FRS: The ablation results on the Kodak dataset for Module 1, Module 2, and Module 3 are shown in TABLE II. Module 1 includes only the FATB module; Module 2 includes only the FRS module; Module 3 includes both FATB and FRS modules. Module 3 achieves a 0.06 dB higher BD-PSNR and 0.67% lower BD-Rate than Module 1, and a 0.04 dB higher BD-PSNR and 0.63% lower BD-Rate than Module 2. This demonstrates that the collaborative interaction between FATB (frequency-domain feature enhancement) and FRS (adaptive bit allocation) is synergistic, as neither module alone can achieve the optimal rate-distortion performance.

To further verify the specific contribution of each core module (FATB and FRS) to high-frequency detail preservation, we supplemented specialized ablation experiments targeting high-frequency-dense scenarios as shown in TABLE III. Three high-frequency-dense images from the Kodak dataset are chosen as the target images (kodim23: feather textures, kodim16: cloud edges, kodim15: skin texture details). These are representative of the fine details that SFTNet aims to preserve.

TABLE III

Ablation Study on High-Frequency Dense Images. Edge Similarity Refers to Structural Similarity Between Edge Maps of the Original and Reconstructed Images. The Ablation Study Follows the Original Configuration: 2 Million Training Iterations, Adam Optimizer With $\beta_1 = 0.9$ and $\beta_2 = 0.999$, 4λ Values (0.0032, 0.0075, 0.015, 0.045) for MSE Loss, and Testing on a Single NVIDIA RTX 3090 Ti

Models	BD-PSNR (dB)	BD-Rate (%)	Edge Similarity (%)
SFTNet w/o FATB	0.19	-4.0	71
SFTNet w/o FRS	0.25	-5.1	73
Full SFTNet	0.45	-8.4	83

TABLE IV

Ablation Study of Transformer Layer in FATB, With All Models Trained for Identical Iterations Using MSE Loss

Schemes	BD-PSNR (dB)	BD-Rate (%)	Parameters (M)
1-Layer	0.18	-5.21	22.68
2-Layers	0.22	-5.98	24.38
3-Layers	0.21	-5.85	28.04

We introduced BD-PSNR, BD-Rate, and the high-frequency-specific metric Edge Similarity (defined as the structural similarity between edge maps of original and reconstructed images) for quantitative evaluation. As shown in TABLE III, the full SFTNet achieves BD-PSNR of 0.45 dB and BD-Rate of −8.4%, with 83% Edge Similarity. Without FATB, the model's BD-PSNR and BD-Rate drop to 0.19 dB and −4.0%, which indicates the fine details reconstruction performance of FATB. Similarly, without FRS, the model achieves BD-PSNR of 0.25 dB and BD-Rate of −5.1%, which further demonstrates the strong detail recovery ability of FRS by allocating bits to high-energy components. These results explicitly validate that both FATB and FRS play irreplaceable roles in high-frequency detail preservation, and their synergy maximizes this advantage.

2) The Number of Transformer Layers in FATB: To analyze the impact of Transformer layer number K in FATB, we set the number of Transformer layers $K = 1, 2, 3$ and evaluated the models on the Kodak dataset, as shown in TABLE IV. When K is increased from 1 to 2, it results in a 0.04 dB gain in BD-PSNR and a 0.77% reduction in BD-Rate, accompanied by a slight increase in parameters. This demonstrates that a second layer enables FATB to more effectively model cross-block frequency dependencies without introducing excessive complexity. However, when the layer count is further raised to 3, BD-PSNR decreases by 0.01 dB and BD-Rate increases by 0.13%, while the parameters grow to 28.04 M. This indicates that three layers lead to redundant computations without any improvement in rate-distortion performance, confirming that two layers achieve the optimal balance between modeling capability and computational efficiency.

3) Apply Transformer to Hyperprior: Given that applying Transformer to both encoder and decoder yields significant improvements, we explore whether integrating Transformer into hyperprior can further boost performance, as shown in

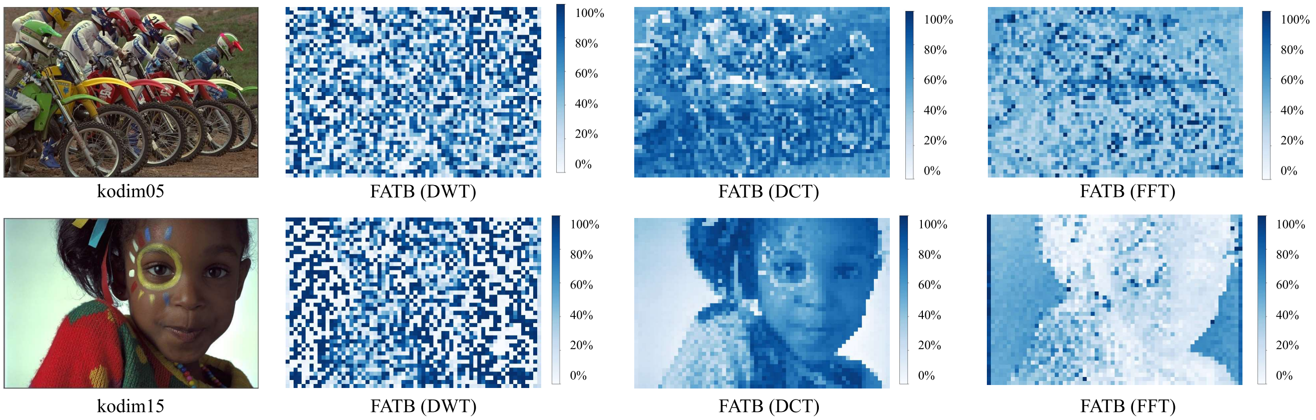

Fig. 11. The residual error maps generated by the Frequency-aware Transformer Block (FATB) under FFT, DCT, and DWT transformations, respectively. In the maps, light or white regions are identified as areas with smaller residuals after processing by the FATB module, indicating that more information from the original images is retained. Conversely, darker or blue regions are associated with larger residuals, suggesting a lower degree of information retention during the processing.

TABLE V
THE ABLATION RESULTS OF FATB IN DIFFERENT POSITIONS

Schemes	En/Decoder	Hyper.	BD-PSNR(dB)	Params(M)
Baseline	✗	✗	- -	15.71
SFTNet	✓	✗	0.52	34.38
SFTNet-Hyper	✓	✓	0.41	36.11

TABLE VI
ABLATION STUDY ON ENERGY RATIO THRESHOLDS (f_{TH}) FOR 64×64 FEATURE MAPS DERIVED FROM 256×256 INPUT IMAGES f_{IN}

Energy Ratio	f_{th}	BD-PSNR (dB)	BD-Rate (%)
70%	12	0.06	-0.05
80%	18	0.31	-6.55
90%	24	0.21	-3.55

TABLE VII
THE QUANTITATIVE RESIDUAL ANALYSIS OF BD-PSNR AND BD-RATE COMPARING FFT-, DCT- AND DWT- BASED FATB ON KODAK DATASET

Transform methods	BD-PSNR (dB)	BD-Rate (%)
FFT	0.31	-6.55
DCT	0.25	-5.28
DWT	0.27	-5.62

TABLE V. Baseline denotes a setup without any Transformer layers. SFTNet denotes a setup with Transformer layers only in the encoder and decoder. SFTNet-Hyper denotes a setup that adds Transformer layers in both the encoder-decoder and hyperprior encoder-decoder, with one layer at the end of the hyperprior encoder and another at the start of the hyperprior decoder to capture latent feature redundancy. The results of SFTNet-Hyper show performance degradation, indicating that redundancy in latent features is already efficiently reduced by the original hyperprior transformation (without extra Transformer layers). Thus, Transformer layers are omitted in the hyperprior to avoid unnecessary growth of parameters.

4) Frequency Decomposition Methods: To address the uncertainty in Eq. 4 regarding the determination of masks m_l and m_h, we focused on clarifying how to set the energy ratio r_e and its corresponding threshold f_{th} for separating low and high frequencies, and supplemented ablation experiments on the Kodak dataset. The lower the energy ratio for m_l, the more it violates the low-frequency-dominant energy distribution of natural images. It is difficult for the low-frequency component f_l to capture complete global structure information, which in turn makes it difficult for the intra-block frequency attention of FATB to obtain effective guidance. To validate the optimal threshold, we test three ratios (70%, 80% and 90%) on the Kodak dataset as shown in TABLE VI.

When the ratio is set to 80%, it achieves the highest BD-PSNR of 34.13 dB and the lowest BD-Rate of -6.55%. In contrast, the 70% ratio lacks sufficient low-frequency energy, while the 90% ratio wastes bits on redundant low-frequency structures. The results further confirm the optimality of the 80% energy ratio for mask generation.

To assess the efficacy of distinct transform methods in redundancy removal via joint frequency-spatial domain modeling in FATB, the FFT component is replaced with DCT and DWT transformations, as depicted in Fig. 11. Latent domain quantization loss is analyzed using scaled deviation $\epsilon = |\hat{y} - y| / \sum y$. The residual results in Fig. 11 originate from the 4th (final) layer of FATB in the encoder. The residuals of this final layer directly influence the latent feature y, which is the most relevant feature for rate-distortion (RD) performance. In contrast, the former layers (1)-(3) primarily focus on preliminary feature processing and are less directly related to the final RD performance. Experiments on the Kodak dataset reveal significantly lower latent deviation in the FFT-based approach compared to the DWT-based approach. The capacity of the FFT-based approach to handle high-frequency information and allocate bits efficiently is demonstrated, which contributes to improved compression performance. The optimal balance between detail preservation and data compactness achieved by the FFT makes it the preferred choice for enhancing compression efficiency.

To further quantify the impact of different frequency decomposition methods on compression performance, we conducted quantitative residual analysis and computed BD-PSNR and

BD-Rate for FFT, DCT, and DWT, with results shown in TABLE VII. FFT-based FATB achieves a BD-PSNR of 0.31 dB and a BD-Rate of −6.55%, outperforming DCT-based FATB (with BD-PSNR of 0.25 dB and BD-Rate of −5.28%) and DWT-based FATB (with BD-PSNR of 0.27 dB and BD-Rate of −5.62%). Consistent with the residual deviation analysis in Fig. 11, the quantitative results in TABLE VII further validate the superiority of FFT for frequency decomposition in FATB.

V. Conclusion

In this paper, we present the Spatial-Frequency Transformer Network (SFTNet) for learned image compression, addressing the limitations of the global-local feature modeling and coarse-grained bit allocation in existing methods. The Frequency-aware Transformer Block (FATB) decomposes features into low-frequency global structures and high-frequency local details via fast fourier transform and a gating mechanism. Through inter-block spatial attention for modeling cross-block low-frequency dependencies and intra-block frequency attention for refining high-frequency details guided by low-frequency structures, FATB enables collaborative spatial-frequency learning, thereby enhancing global consistency and preserving local texture. The Feature Re-weighting Strategy (FRS) is utilized to quantify block importance by integrating spatial attention and energy analysis, enabling more bits to be dynamically allocated to complex texture regions with rich high-frequency details. Experimental results demonstrate that SFTNet significantly outperforms existing methods in high-frequency detail reconstruction, validating the effectiveness of our SFTNet.

References

[1] G. K. Wallace, "The JPEG still picture compression standard," *Commun. ACM (CACM)*, vol. 34, no. 4, pp. 30–44, 1991.

[2] D. Taubman and M. W. Marcellin, "JPEG2000 image compression fundamentals, standards and practice," *J. Electron. Imag. (JEI)*, vol. 11, no. 2, p. 286, 2002.

[3] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[4] J.-R. Ohm and G. J. Sullivan, "Versatile video coding–towards the next generation of video compression," in *Proc. Picture Coding Symp. (PCS)*, vol. 2018, 2018.

[5] J. He, X. He, S. Xiong, and H. Chen, "Learned image coding for human–machine collaborative optimization," *IEEE Trans. Broadcast.*, vol. 71, no. 1, pp. 203–216, Mar. 2025.

[6] Z. Pan et al., "JND-LIC: Learned image compression via just noticeable difference for human visual perception," *IEEE Trans. Broadcast.*, vol. 71, no. 1, pp. 217–228, Mar. 2025.

[7] M. Shen, H. Gan, C. Ning, Y. Hua, and T. Zhang, "TransCS: A transformer-based hybrid architecture for image compressed sensing," *IEEE Trans. Image Process.*, vol. 31, pp. 6991–7005, 2022.

[8] A. B. Koyuncu, P. Jia, A. Boev, E. Alshina, and E. Steinbach, "Efficient contextformer: Spatio-channel window attention for fast context modeling in learned image compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 7498–7511, Aug. 2024.

[9] Y. Qian, M. Lin, X. Sun, Z. Tan, and R. Jin, "Entroformer: A transformer-based entropy model for learned image compression," 2022, *arXiv:2202.05492*.

[10] W. Jiang, J. Yang, Y. Zhai, P. Ning, F. Gao, and R. Wang, "MLIC: Multi-reference entropy model for learned image compression," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 7618–7627.

[11] J. Liu, H. Sun, and J. Katto, "Learned image compression with mixed transformer-CNN architectures," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14388–14397.

[12] H. Fu et al., "Learned image compression with Gaussian-Laplacian-logistic mixture model and concatenated residual modules," *IEEE Trans. Image Process.*, vol. 32, pp. 2063–2076, 2023.

[13] Y. Zhang, Z. Duan, Y. Huang, and F. Zhu, "Balanced rate-distortion optimization in learned image compression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 2428–2438.

[14] D. Feng et al., "Linear attention modeling for learned image compression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 7623–7632.

[15] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimization of nonlinear transform codes for perceptual quality," in *Proc. Picture Coding Symp. (PCS)*, 2016, pp. 1–5.

[16] J. Ballé, "Efficient nonlinear transforms for lossy image compression," in *Proc. Picture Coding Symp. (PCS)*, Jun. 2018, pp. 248–252.

[17] D. Minnen, J. Ballé, and G. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," 2018, *arXiv:1809.02736*.

[18] S. Ma, X. Zhang, C. Jia, Z. Zhao, S. Wang, and S. Wang, "Image and video compression with neural networks: A review," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1683–1698, Jun. 2020.

[19] J. Lee, S. Cho, and M. Kim, "An end-to-end joint learning scheme of image compression and quality enhancement with improved entropy minimization," 2019, *arXiv:1912.12817*.

[20] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized Gaussian mixture likelihoods and attention modules," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7936–7945.

[21] Z. Guo, Z. Zhang, R. Feng, and Z. Chen, "Causal contextual prediction for learned image compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2329–2341, Apr. 2022.

[22] J. Yue, M. Ye, L. Ji, H. Guo, and C. Zhu, "A survey of deep-learning-based compressed video quality enhancement," *IEEE Trans. Broadcast.*, early access, Jul. 29, 2025, doi: 10.1109/TBC.2025.3570871.

[23] A. Vaswani et al., "Attention is all you need," in *Proc. Neural Inf. Process. Syst. (NIPS)*, vol. 30, 2025, pp. 5998–6008.

[24] Y. Zhu, Y. Yang, and T. Cohen, "Transformer-based transform coding," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[25] M. Lu, P. Guo, H. Shi, C. Cao, and Z. Ma, "Transformer-based image compression," in *Proc. Data Compress. Conf. (DCC)*, Mar. 2022.

[26] Y. Liu, Y.-H. Wu, G. Sun, L. Zhang, A. Chhatkuli, and L. Van Gool, "Vision transformers with hierarchical attention," 2021, *arXiv:2106.03180*.

[27] D. Marpe, G. Blattermann, J. Ricke, and Maass, "A two-layered wavelet-based algorithm for efficient lossless and lossy image compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 7, pp. 1094–1102, Jul. 2000.

[28] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn.*, Jan. 2008, pp. 1096–1103.

[29] C. Cai, L. Chen, X. Zhang, and Z. Gao, "Efficient variable rate image compression with multi-scale decomposition network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3687–3700, Dec. 2019.

[30] L. Yu, W. Chang, S. Wu, and M. Gabbouj, "End-to-end transformer for compressed video quality enhancement," *IEEE Trans. Broadcast.*, vol. 70, no. 1, pp. 197–207, Mar. 2024.

[31] J. K. Wu and R. E. Burge, "Adaptive bit allocation for image compression," *Comput. Graph. Image Process.*, vol. 19, no. 4, pp. 392–400, May 1982.

[32] C. Hong and K. M. Lee, "AdaBM: On-the-fly adaptive bit mapping for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 2641–2650.

[33] M. Wang et al., "End-to-end image compression with swin-transformer," in *Proc. IEEE Int. Conf. Vis. Commun. Image Process. (VCIP)*, Dec. 2022, pp. 1–5.

[34] H. Xu, B. Hai, Y. Tang, and Z. He, "Window-based channel attention for wavelet-enhanced learned image compression," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2024.

[35] J. Lu, L. Zhang, X. Zhou, M. Li, W. Li, and S. Gu, "Learned image compression with dictionary-based entropy model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 12850–12859.

[36] Y. Ye, Y. Pan, Q. Jiang, M. Lu, X. Fang, and B. Xu, "Frequency-aware re-parameterization for over-fitting based image compression," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2023, pp. 2310–2314.

[37] H. Li, S. Li, W. Dai, C. Li, J. Zou, and H. Xiong, "Frequency-aware transformer for learned image compression," 2023, *arXiv:2310.16387*.

[38] H. Rhee, Y. I. Jang, S. Kim, and N. I. Cho, “LC-FDNet: Learned lossless image compression with frequency decomposition network,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6023–6032.

[39] D. Mishra, S. K. Singh, and R. K. Singh, “Wavelet-based deep auto encoder–decoder (WDAED)-based image compression,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1452–1462, Apr. 2021.

[40] H. Fu, J. Liang, Z. Fang, J. Han, F. Liang, and G. Zhang, “WeConvene: Learned image compression with wavelet-domain convolution and entropy model,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024.

[41] F. Li and J. Yang, “Joint demosaicing and denoising with frequency domain features,” in *Proc. Int. Conf. Artif. Neural Netw. (ICANN)*, 2023.

[42] J. W. Cooley, P. A. W. Lewis, and P. D. Welch, “The fast Fourier transform and its applications,” *IEEE Trans. Educ. (TE)*, vol. TE-12, no. 1, pp. 27–34, Jan. 1969.

[43] N. Ahmed, T. Natarajan, and K. R. Rao, “Discrete cosine transform,” *IEEE Trans. Comput. (TC)*, vols. C–23, no. 1, pp. 90–93, Jan. 1974.

[44] S. G. Mallat, “A theory for multiresolution signal decomposition: The wavelet representation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989.

[45] G. K. Wallace, “The JPEG still picture compression standard,” *IEEE Trans. Consum. Electron.*, vol. 38, no. 1, pp. xviii–xxxiv, Feb. 1992.

[46] J. Wan, H. Yin, Z. Liu, A. Chong, and Y. Liu, “Lightweight image super-resolution by multi-scale aggregation,” *IEEE Trans. Broadcast.*, vol. 67, no. 2, pp. 372–382, Jun. 2021.

[47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[48] (2013). *Kodak Photocd Dataset*. [Online]. Available: http://r0k.us/graphics/kodak/

[49] (2019). *CLIC Dataset*. [Online]. Available: http://www.compression.cc/

[50] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, *arXiv:1412.6980*.

[51] D. He, Z. Yang, W. Peng, R. Ma, H. Qin, and Y. Wang, “ELIC: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5708–5717.

[52] C. Li, S. Yin, C. Jia, F. Meng, Y. Tian, and Y. Liang, “Multirate progressive entropy model for learned image compression,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 7725–7741, Aug. 2024.

Meiqin Liu received the M.E. and Ph.D. degrees from Beijing Jiaotong University (BJTU), China, in 2007 and 2018, respectively. From 2014 to 2015, she was a Visiting Scholar at Simon Fraser University (SFU), Canada. She is currently a Professor at the Institute of Information and Science, BJTU. Her research interests include image/video compression and video processing.

Lingxue Li is currently pursuing the degree with the School of Computer Science and Technology, Beijing Jiaotong University (BJTU), China. Her research interests focus on image compression, computer vision, and deep learning.

Jiaming Liang received the M.E. degree in information and communication engineering from Beijing Jiaotong University (BJTU), China, in 2023. His research interests include image compression, computer vision, and deep learning.

Chao Yao (Member, IEEE) received the M.E. and Ph.D. degrees from Beijing Jiaotong University (BJTU) in 2010 and 2016, respectively. From 2014 to 2015, he was a visiting Ph.D. Student with the LTS4 Group, Institute of the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. He is currently a Professor with the University of Science and Technology Beijing (USTB). His research interests include image/video compression, computer vision, and human–computer interaction.

Jian Jin (Member, IEEE) received the Ph.D. degree in signal and information processing from Beijing Jiaotong University, China, in 2019. From 2016 to 2018, he was a joint Ph.D. Student at Simon Fraser University, Canada. He is currently a Senior Research Fellow at the College of Computing and Data Science, Nanyang Technological University, Singapore. His research interests include visual perceptual modeling and visual quality assessment.

Tammam Tillo (Senior Member, IEEE) received the Diploma degree from Damascus University, Damascus, Syria, in 1994, and the Ph.D. Diploma degree from the Politecnico di Torino, Italy, in 2005. In 1996, he completed his military service in Syria. From 2005 to 2008, he was with the Image Processing Laboratory, Politecnico di Torino. In 2008, he joined Xi’an Jiaotong–Liverpool University, China. In 2017, he joined the Free University of Bozen-Bolzano, Italy. In 2021, he joined the Indraprastha Institute of Information Technology Delhi, Delhi, India. In 2025, he joined Taiyuan University of Science and Technology, Taiyuan, China.

Yao Zhao (Fellow, IEEE) received the B.S. degree from the Radio Engineering Department, Fuzhou University, Fuzhou, China, in 1989, the M.E. degree from the Radio Engineering Department, South East University, Nanjing, China, in 1992, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing, China, in 1996. He became an Associate Professor with BJTU in 1998 and became a Professor in 2001. His current research interests include image/video coding, digital watermarking and forensics, and video analysis and understanding. He was named a Distinguished Young Scholar by the National Science Foundation of China in 2010 and was elected as a Chang Jiang Scholar of the Ministry of Education of China in 2013.