A Feature Dynamic Enhancement and Global Collaboration Guidance Network for Remote Sensing Image Compression

Qizhi FANG [1, 2], Shibo GU [2], Jingang WANG [2], Lili ZHANG [2]

[1] Liaoning General Aviation Academy, Shenyang 110136, China
[2] College of Electronic Information Engineering, Shenyang Aerospace University, Shenyang 110136, China

{arinc2006, 20052727}@sau.edu.cn, {gushibo, wangjingang}@stu.sau.edu.cn

Submitted March 7, 2025 / Accepted March 26, 2025 / Online first May 19, 2025

Abstract. *Deep learning-based remote sensing image compression methods show great potential, but traditional convolutional networks mainly focus on local feature extraction and show obvious limitations in dynamic feature learning and global context modeling. Remote sensing images contain multiscale local features and global low-frequency information, which are challenging to extract and fuse efficiently. To address this, we propose a Feature Dynamic Enhancement and Global Collaboration Guidance Network (FDEGCNet). First, we propose an Omni-Dimensional Attention Model (ODAM), which dynamically captures the key salient features in the image content by adaptively adjusting the feature extraction strategy to enhance the model's sensitivity to key information. Second, a Hyperprior Efficient Attention Model (HEAM) is designed to combine multi-directional convolution and pooling operations to efficiently capture cross-dimensional contextual information and facilitate the interaction and fusion of multi-scale features. Finally, the Multi-Kernel Convolutional Attention Model (MCAM) integrates global branching to extract frequency domain context and enhance local feature representation through multi-scale convolutions. The experimental results show that FDEGCNet achieves significant improvement and maintains low computational complexity regarding image quality evaluation metrics (PSNR, MS-SSIM, LPIPS, and VIFp) on the three datasets compared to the advanced compression models. Code is available at https://github.com/shiboGu12/FDEGCNet*

Keywords

Remote sensing image compression, convolutional networks, multiscale convolution, attention model, multiscale local features, global low-frequency information

1. Introduction

Remote sensing images can reflect a wealth of information about features, such as surface types, vegetation cover, water bodies, buildings, and so on. Therefore, remote sensing images are widely used in many fields [1], [2] such as earth science, geological exploration, environmental monitoring, agriculture, urban planning, and management. With the upgrading of platforms such as satellites, airplanes, and drones, along with the widespread use of high-resolution sensors [3], the volume and complexity of remotely sensed imagery continue to rise. The exponential growth in data volume may lead to serious transmission and storage challenges for remote sensing satellites and users. Consequently, the development of effective compression techniques for remote sensing images is of paramount importance. Compared with natural images, remote sensing images are affected by imaging angles, atmospheric conditions, lighting conditions, and other factors. Remote sensing images are characterized by rich feature information, delicate texture details, and a mixture of high-frequency and low-frequency features, which makes it difficult for traditional image compression methods to effectively compress remote sensing images [4].

Recent advancements in traditional remote sensing image compression methods have yielded notable research outcomes. For instance, Báscones et al. [5] have proposed a method that integrates Principal Component Analysis (PCA) with JPEG2000 [6] to compress hyperspectral image data. This underscores the pressing need to develop efficient remote sensing image compression networks. This demand has led to the emergence of more advanced networks, such as WebP [7] and BPG [8], which play a crucial role in the efficient storage and transmission of image data. However, these standards exhibit some notable limitations [9], [10]. First, due to the block-based hybrid coding approach, the encoding and decoding processes need to be processed block by block, which is prone to produce undesirable block effects or ringing artifacts in the decoded image. Second, these methods rely on complex module dependencies, which complicates the optimization of the overall algorithm. The resolution of remote sensing images continues to increase. The demand for diversified applications is growing. As a result, developing more advanced compression techniques has become increasingly important.

DOI: 10.13164/re.2025.0324

In the field of image compression, there has been a notable shift in focus towards learnable compression models [11–13]. Within the domain of deep learning-based image compression, prominent frameworks include Autoencoders (AE) [14] and Variational Autoencoders (VAE) [15]. These frameworks encompass two symmetrical data processing stages: compression and reconstruction. Since VAE has continuous mapping spatial capability compared to AE, it helps to reconstruct the image with a smooth transition. Therefore, VAE based framework has a more powerful image reconstruction capability. Therefore, Ballé et al. [11] develop a VAE-based image compression model that utilizes a hyperprior structure to capture spatial dependencies in the latent representation. In addition, the side information is used to estimate the variance of the parameter distributions. For more accurate compression modeling, Cheng et al. [10] proposed a Gaussian mixture model that uses a discretized Gaussian mixture likelihood to parameterize the distribution of potential codes, thus improving the accuracy of entropy model predictions. In order to improve the nonlocal modeling capability, Liu et al. [14] proposed a parallel structure of transformer and convolutional neural network (CNN). They cleverly fused the two methods of CNN and transformer, combining the local modeling ability of CNN and the nonlocal modeling ability of the transformer, in order to improve the overall performance of the image compression model.

In recent years, learnable compression models have also been introduced into remote sensing image compression due to their powerful feature extraction and representation capabilities. Although these techniques perform well in processing natural images, remote sensing images still face greater challenges in terms of compression effectiveness due to factors such as complex texture and spatial information [4]. Therefore, how to improve the reconstruction quality of remote sensing images by considering various factors comprehensively has become the focus of current research. Tang et al. [16] proposed an end-to-end image compression method combining graph attention and asymmetric CNN. The method overcomes the over-reliance of traditional CNNs in processing local features to a certain extent and promotes effective interaction between information. Although the CNN-based approach excels in extracting spatial information and local contextual features, it extracts latent features by applying convolutional filters in the local receptive domain. This approach leads the network to focus too much on the local details of the image, thus reducing the attention to the global visual features. To overcome this limitation, Zhang et al. [17] introduced a global anchored stripe self-attention mechanism. It captures global, local, and inter-channel information dependencies and enhances feature extraction during encoding and decoding with multi-scale modules. In addition, Pan et al. [18] developed a Coupled Compression Generation Network, which enhances information integrity and texture resolution through separate content and texture branches. In the generation stage, a Multi-Dimensional Residual Attention Module focuses on critical task information, while the texture branch employs a GAN-based training strategy. This strategy integrates the Enhanced Perception-Guided Refinement Stage and a Multi-Scale Fusion Discriminator to improve texture quality. Zhang et al. [19] proposed a low-complexity transformer-CNN hybrid model (LTCHM), which focuses on integrating non-local and channel information in remote sensing images by combining the dynamic attention model and the hyper-prior hybrid attention model.

Although existing remote sensing image compression algorithms have made significant progress in terms of rate-distortion performance, they still face several challenges when compressing high-resolution remote sensing images. First, the high resolution of remote sensing images results in low spatial continuity between neighboring pixels. It makes the compression efficiency suffer because the reduced continuity between pixels increases data redundancy. Second, remote sensing images contain multi-scale features such as vehicles, buildings, roads, and other landforms, which makes it more complicated to capture both small- and large-scale features simultaneously and efficiently. In addition, remotely sensed images usually contain a large amount of global information, such as terrain features and landform details, which are crucial for achieving efficient image compression and accurate image reconstruction. If the global information is not captured efficiently, the reconstructed image may lose critical information, resulting in blurring and detail loss.

To address these challenges, this paper proposes a novel remote sensing image compression model called FDEGCNet. It introduces three innovative approaches to overcome the limitations of existing methods. The specific contributions of this paper are summarized as follows:

- To overcome the static nature of traditional convolutional kernels and their limitations in feature learning, ODAM is proposed in this paper. By introducing Omni-Dimensional Dynamic Convolution (ODConv) [20] and deeply integrating it with the attention mechanism. As a result, ODAM makes the convolution kernel adaptive in multiple dimensions (including the number of convolution kernels, spatial dimensions, input channels, and output channels) according to different features of the input data. This mechanism enables the convolution strategy to be adaptively adjusted according to the specific features of the input data. Thus, it can capture the subtle texture and edge variations in remote sensing images more accurately.

- To address the challenge of multi-scale feature characterization in remote sensing images, this paper proposes the HEAM module. The module utilizes parallel convolutional branches and adaptive pooling methods to capture global spatial contextual information of an image in multiple spatial dimensions (height, width, and channel). HEAM can effectively enhance the fusion of local and global features by learning the feature weights across the space to improve the spatial feature representation. In addition, the HEAM module provides more accurate probabilistic guidance for the encoding and de-

coding process and realizes the effective coordination of global features.

- To effectively process local details and global structure information in remote sensing images, MCAM is proposed in this paper. This module can flexibly capture feature information at different scales by introducing a multi-scale convolution kernel, which enhances the model's ability to perceive multi-level features in images. Meanwhile, MCAM converts global features to the frequency domain for processing by transforming between the spatial domain and the frequency domain using the Fourier transform to fully utilize the global representation capability of the frequency domain. For local features, weighting operations are performed in the spatial domain to highlight the detailed information. This synergistic mechanism between spatial and frequency domains effectively combines the advantages of both, thus significantly improving the image compression performance.
- The experimental results demonstrate that the proposed network outperforms traditional image compression methods and advanced deep learning-based image compression methods on the DOTA, UC-Merced, and NWPU-RESISC45 datasets. The superior compression performance of the proposed network is evident in terms of Peak Signal-to-Noise Ratio (PSNR), Multi-Scale Structural Similarity (MS-SSIM), Learned Perceptual Image Patch Similarity (LPIPS), and Visual Information Fidelity in the Pixel Domain (VIFp).

The structure of this study is organized as follows: Section 2 provides a thorough overview of the FDEGC-Net compression framework, including its fundamental components—ODAM, HEAM, and MCAM. Section 3 discusses the experimental setup and the datasets utilized, followed by extensive experiments comparing and analyzing the proposed FDEGCNet with other compression methods. Section 4 concludes the study and discusses potential future research directions.

2. Proposed Method

2.1 Overall Framework

The proposed remote sensing image compression framework is illustrated in Fig. 1. The overall network architecture consists of a main encoder (g_a) and decoder (g_s), a hyperprior encoder (h_a) and decoder (h_s), a MCAM, and an entropy model. Q represents the quantizer, AE represents the arithmetic encoder, and AD represents the arithmetic decoder. The main encoder and decoder are constructed using the ODAM and residual blocks. The hyperprior encoder and decoder network include the HEAM and a downsampling module. Assuming an input image of dimension $X \in \mathbb{R}^{C\times H\times W}$, it is transformed into a latent representation y through the encoder, which consists of residual blocks and ODAM. The latent representation y is quantized using the quantization operation Q to obtain $\widehat{y}$, and the arithmetic encoder is then applied to $\widehat{y}$ to generate the compressed bitstream. During quantization, truncation errors (i.e., $y-Q(y)$) are introduced, leading to certain reconstruction distortions in the decoded image. To emulate the quantization process during training and circumvent non-differentiable operations, uniform noise $\mathcal{U}(-0.5, 0.5)$ is incorporated to approximate quantization [21]. In the prediction phase, the potential representation y is then discretized using a rounding function to obtain a discrete representation for actual image compression and reconstruction. In decompression, the reconstructed image is obtained with $\widehat{y}$ and decoder g_a network.

To encode $\widehat{y}$ with fewer bits, entropy models are commonly used to parameterize the distribution of $\widehat{y}$. After decoding $\widehat{z}$, this study utilizes the Gaussian Mixture Entropy Model proposed by Cheng et al. [10] to estimate $p_{(\widehat{y}|\widehat{z})}(\widehat{y}|\widehat{z})$, expressed as:

$$p_{(\widehat{y}|\widehat{z})}(\widehat{y}|\widehat{z}) \sim \sum_{k=1}^{K} w^{(k)} \mathcal{N}\left(\mu^{(k)}, \sigma^{2(k)}\right). \tag{1}$$

The entropy model is further expressed as:

$$\begin{aligned} p_{(\widehat{y}|\widehat{z})}(\widehat{y}|\widehat{z}) &= \prod_i p_{(\widehat{y}|\widehat{z})}(\widehat{y}_i|\widehat{z}), \\ p_{(\widehat{y}|\widehat{z})}(\widehat{y}|\widehat{z}) &= \left(\sum_{k=1}^{K} w^{(k)} \mathcal{N}\left(\mu_i^{(k)}, \sigma_i^{2(k)}\right)\right) \\ &\quad * \mathcal{U}(-0.5, 0.5)(\widehat{y}_i) \end{aligned} \tag{2}$$

where i represents the position of the feature map, k is the index of the mixture components, and each component is characterized by three parameters: weight ($w_i^{(k)}$), mean ($\mu_i^{(k)}$), and variance ($\sigma_i^{2(k)}$).

The core objective of an image compression network is to achieve an optimal balance on the rate-distortion curve. This balance is regulated by a Lagrange multiplier λ, which trades off between compression distortion and the desired bit rate. This relationship can be expressed as:

$$\begin{aligned} \mathcal{L} &= \mathcal{R}(\widehat{y}) + \mathcal{R}(\widehat{z}) + \lambda \cdot D(x, \widehat{x}) \\ &= \mathbb{E}\left[-\log_2\left(p_{\widehat{y}|\widehat{z}}(\widehat{y}|\widehat{z})\right)\right] + \mathbb{E}\left[-\log_2\left(p_{\widehat{z}|\varphi}(\widehat{z}|\varphi)\right)\right] \\ &\quad + \lambda \cdot D(x, \widehat{x}) \end{aligned} \tag{3}$$

where $\mathcal{R}(\widehat{y})$ and $\mathcal{R}(\widehat{z})$ represent the bit rates of $\widehat{y}$ and $\widehat{z}$, respectively. $D(x, \widehat{x})$ denotes the distortion between the original and reconstructed images, typically calculated using MSE or MS-SSIM. Since no prior information is available, a non-parametric fully factorized density model φ is used for entropy estimation:

$$p_{(\widehat{z}|\varphi)}(\widehat{z}|\varphi) = \prod_i \left(P_{z_i|\varphi}(\varphi) * \mathcal{U}(-0.5, 0.5)\right)(\widehat{z}_i) \tag{4}$$

where z_i represents the i-th element of z, and i denotes the position of each element.

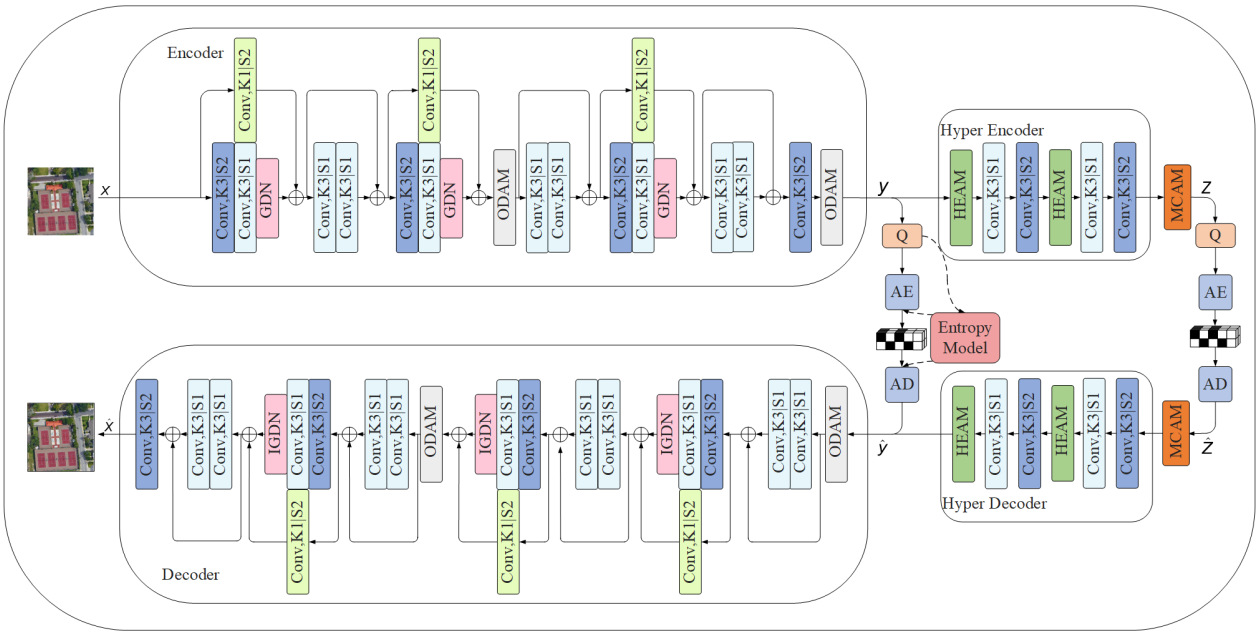

Fig. 1. The overall compression framework. The Entropy Model is the Gaussian Mixture Entropy Model. ODAM denotes the Omni-Dimensional Attention Model. HEAM denotes the Hyperprior Efficient Attention Model. MCAM denotes the Multi-Kernel Attention Model. K denotes the kernel size, where $K1$ denotes a kernel size of 1 and $K3$ denotes a kernel size of 3. S denotes the stride size, where $S1$ denotes a stride of 1 and $S2$ denotes a stride of 2.

2.2 Omni-Dimensional Attention Model

To enable the main encoder and decoder to dynamically focus on the critical regions of an image, this study proposes the ODAM, whose core component is the ODConv [20]. The structure of ODConv is shown in Fig. 2(a). Assuming the input feature map is $X \in \mathbb{R}^{C\times H\times W}$, the model first applies adaptive average pooling to the input to reduce its dimensions and extract global feature information across channels. Subsequently, the input is passed through a 1×1 convolution layer to generate a more accurate channel representation, effectively compressing the information and reducing computational complexity. Next, BatchNorm normalization and ReLU activation are performed to achieve dimensionality reduction and nonlinear mapping, which in turn generates four different attention weights. Specifically:

- Convolutional Kernel Attention Scalar (CKAS): The input features are adaptively average pooled and 1×1 convolved to generate the attention weights of the convolution kernels. This weight is used to dynamically adjust the weights of the convolution kernel combinations, allowing the model to flexibly select different convolution kernels and enhance the diversity of feature extraction.
- Output Channel Attention Scalar (OCAS): The number of channels is mapped to the number of output channels after the input features have been adaptively average pooled and 1×1 convolved. The attention weights of the filters are restricted to the range $[0, 1]$ by a Sigmoid activation function, which dynamically adjusts the contribution of each convolution kernel channel to enhance the attention to the important filters.
- Input Channel Attention Scalar (ICAS): The input features are subjected to adaptive mean pooling to compress the spatial dimension (H, W) to $(1, 1)$ to obtain global information at the channel level. This global information is dimensionalized and activated by a 1×1 convolutional layer, and the number of channels is recovered by another 1×1 convolutional layer. Finally, the channel weights are mapped to the $[0, 1]$ interval by a Sigmoid activation function in order to multiply them channel-by-channel with the input feature map to enhance or suppress the importance of different channels.
- Spatial Dimension Attention Scalar (SDAS): The input features are first processed through adaptive average pooling and a 1×1 convolution to generate spatial attention weights, which are applied to each spatial location of the input features. This mechanism allows the model to weigh important regions in the spatial dimensions, enabling it to focus on more meaningful spatial positions.

In the model, the ICAS is first multiplied channel-by-channel with the input feature map to enhance or suppress the importance of different channel features. Then, the SDAS and CKAS are applied to the spatial dimension of the convolution kernel and the number of convolution kernel groups, respectively, to generate the dynamic convolution kernel. Finally, the dynamic convolution kernel is applied to the input feature maps and the convolution operation is performed to obtain the output feature maps, which are then multiplied with the OCAS to further weight the output channels. The operation of ODConv can be defined as follows:

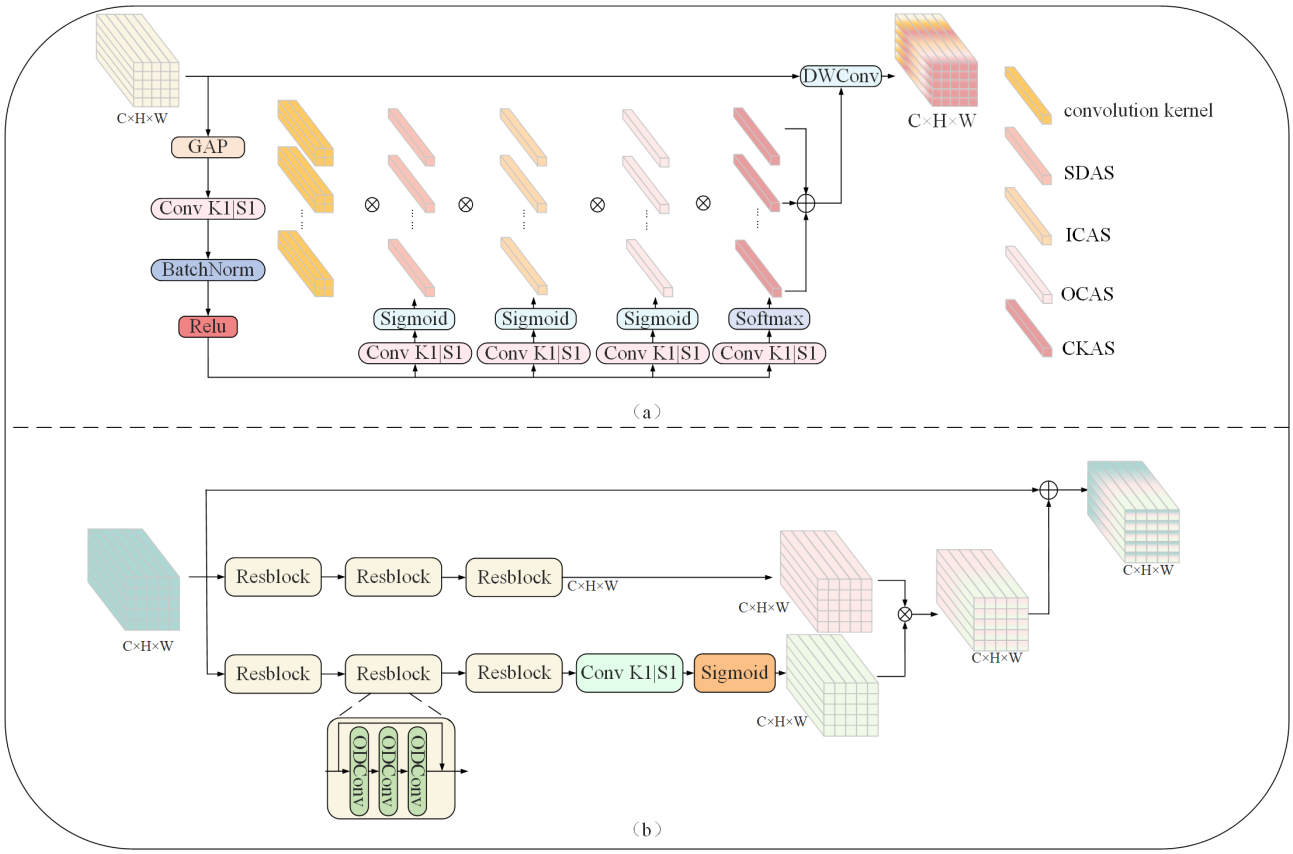

Fig. 2. (a) Omni-Dimensional Dynamic Convolution (ODConv): C, K, and S represent the number of channels, the number of convolutional kernels, and the stride size, respectively. $C = 192$ when $\lambda = \{128, 256, 512\}$ and $C = 256$ when $\lambda = \{1024, 2048, 4096\}$. $K1|S1$ indicates that the kernel size is 1 and the stride is 1. GAP refers to Global Average Pooling. (b) Omni-Dimensional Dynamic Attention Model (ODAM): C, K, and S represent the number of channels, the number of convolutional kernels, and the stride size, respectively. $K1|S1$ indicates that the kernel size is 1 and the stride is 1.

$$y = \left(\alpha_{w1} \odot \alpha_{f1} \odot \alpha_{c1} \odot \alpha_{s1} \odot W_1 + \cdots + \alpha_{wn} \odot \alpha_{fn} \odot \alpha_{cn} \odot \alpha_{sn} \odot W_n\right) * x \tag{5}$$

where $\odot$ denotes the multiplication operation and * denotes the convolution operation. W_i represents the convolutional kernel, while α_{wi}, α_{fi}, α_{ci}, and α_{si} represent the four scalar attention weights of the convolutional kernel: CKAS, OCAS, ICAS, and SDAS, respectively. x is the input feature, and y is the output feature.

Attention mechanisms have been extensively applied in the domain of image compression [10, 13, 17, 22] and have been demonstrated to enhance compression performance and improve rate-distortion efficiency. Building upon this foundation, we propose an ODAM based on ODConv. As shown in Fig. 2(b), ODConv is used to construct residual blocks, which process input features during the convolution operation and add the processed results to the original input through residual connections. This mechanism effectively prevents information loss and gradient vanishing problems. Unlike traditional static convolutions, ODConv generates four types of attention: input channel attention, output channel attention, spatial attention, and convolutional kernel attention. These dynamically generated scalars are used to adjust the weights of the convolution kernels, allowing the convolution operation to adaptively select different convolution kernels based on the characteristics of the input data, thus flexibly enhancing the focus on important features and effectively suppressing unimportant features.

2.3 Hyperprior Efficient Attention Model

Convolutional Neural Networks have demonstrated proficiency in capturing local features due to their localized receptive fields. However, their limitations become apparent when addressing tasks that require global information modeling [23]. To enhance global feature modeling and effectively suppress information redundancy, this paper proposes a novel HEAM, which is deeply integrated with the hyperprior encoder and decoder framework to improve the accuracy of feature extraction and representation. As shown in Fig. 3, assuming the input feature map is $X \in \mathbb{R}^{C\times H\times W}$, the features X are divided into G sub-feature groups through a grouping operation, enabling more efficient feature processing and global information modeling. This process can be expressed as follows:

$$group_x = \text{reshape}\left(X\left(B \cdot G, C/G, H, W\right)\right). \tag{6}$$

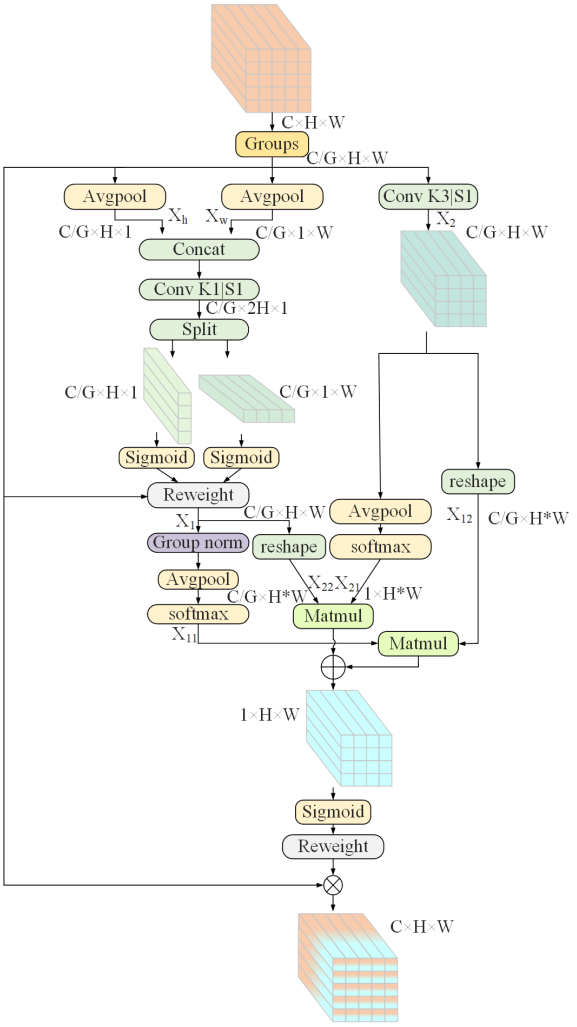

Fig. 3. HEAM Network Structure. Groups denote the grouping operation. C, K, and S represent the number of channels, the number of convolutional kernels, and the stride size, respectively. $C = 192$ when $\lambda = \{128, 256, 512\}$ and $C = 256$ when $\lambda = \{1024, 2048, 4096\}$. $K3|S1$ indicates that the kernel size is 3 and the stride is 1. Reweight refers to the dynamic spatial reweighting within each group.

Next, through three independent sub-networks, the feature map is first subjected to pooling operations along the height and width dimensions to calculate the average values for each row (x_h) and each column (x_w), respectively. Then, the cross-channel information interaction between two parallel paths in a 1×1 branch is realized by aggregating the two-channel attentions through a simple multiplication operation, thus fusing the feature information in the height and width directions. Finally, the concatenated features are split back into height features x_h and width features x_w. The specific process can be expressed by the following formulas:

$$\begin{aligned} x_h &= \text{pool}_h\left(group_x\right), \\ x_w &= \text{pool}_w\left(group_x\right), \\ hw &= \text{conv}_{1\times1}\left(\text{cat}\left([x_h, x_w]\right)\right), \\ x_h, x_w &= \text{split}\left(hw, [H, W]\right). \end{aligned} \tag{7}$$

The activation results of x_h and x_w are multiplied and then group normalized to obtain the normalized feature x_1. In 3×3 branching, 3×3 convolution is applied to capture the local cross-channel interactions, extend the feature space, and further extract the local features to obtain an enhanced feature representation.

$$\begin{aligned} x_{11} &= \text{reshape}(\text{softmax}\left(Gap\left(x_1\right)\right), \\ x_{12} &= \text{reshape}\left(x_2\right), \\ x_{21} &= \text{reshape}(\text{softmax}\left(Gap\left(x2\right)\right), \\ x_{22} &= \text{reshape}\left(x_1\right), \\ weights &= \left(\text{matmul}\left(x_{11}, x_{12}\right) + \text{matmul}\left(x_{21}, x_{22}\right)\right), \\ output &= \text{reshape}\left(group_x \cdot \text{sigmoid}\left(weights\right)\right) \end{aligned} \tag{8}$$

where x_{11} represents the dependency and weights between the computed channels, x_{12} represents the number of channels and spatial dimensions of the feature map after convolution. x_{21} represents the Softmax weights of the feature map obtained after convolution. x_{22} represents the number of channels and spatial dimensions of the processed feature map. The weights are the final weights. The output represents the output of the network.

2.4 Multi-Kernel Convolutional Attention Model

Cui et al. [24] proposed an Omni-Kernel Model (OKM) that effectively captures the multi-scale receptive fields required for image restoration through dual-domain processing and large kernel-scale depthwise convolution modulation. On this basis, we design a MCAM, which combines multi-kernel convolution with a Dual-Domain Attention Module (DDAM), to optimize the modulation effect of the global representation, so as to achieve an efficient fusion of global and local information. To reduce the computational cost of convolution operations, the MCAM is deployed between the hyperprior encoder and decoder, where the feature map has the smallest dimensions within the compression network. The architecture of MCAM is shown in Fig. 4(a). Given the input feature $X \in \mathbb{R}^{C\times H\times W}$, it first undergoes channel transformation via a 1×1 convolution and is then fed into the global branch, mixed-kernel branch, and local branch.

In the global branch, the features are fed into the DDAM, whose structure is shown in Fig. 4(b). The global features X_{Global} are first subjected to global average pooling and 1×1 convolution to generate channel attention to extract global information and guide frequency domain weighting. Then the global features X_{Global} are switched to the frequency domain by Fourier transform to obtain the frequency domain features. The frequency domain features are weighted using channel attention to enhance important frequency domain components and suppress redundant features. Then, return to the spatial domain by inverse Fourier transform to extract the spatial context information and get the spatial attention weights X_{Global1}. This process can be expressed as:

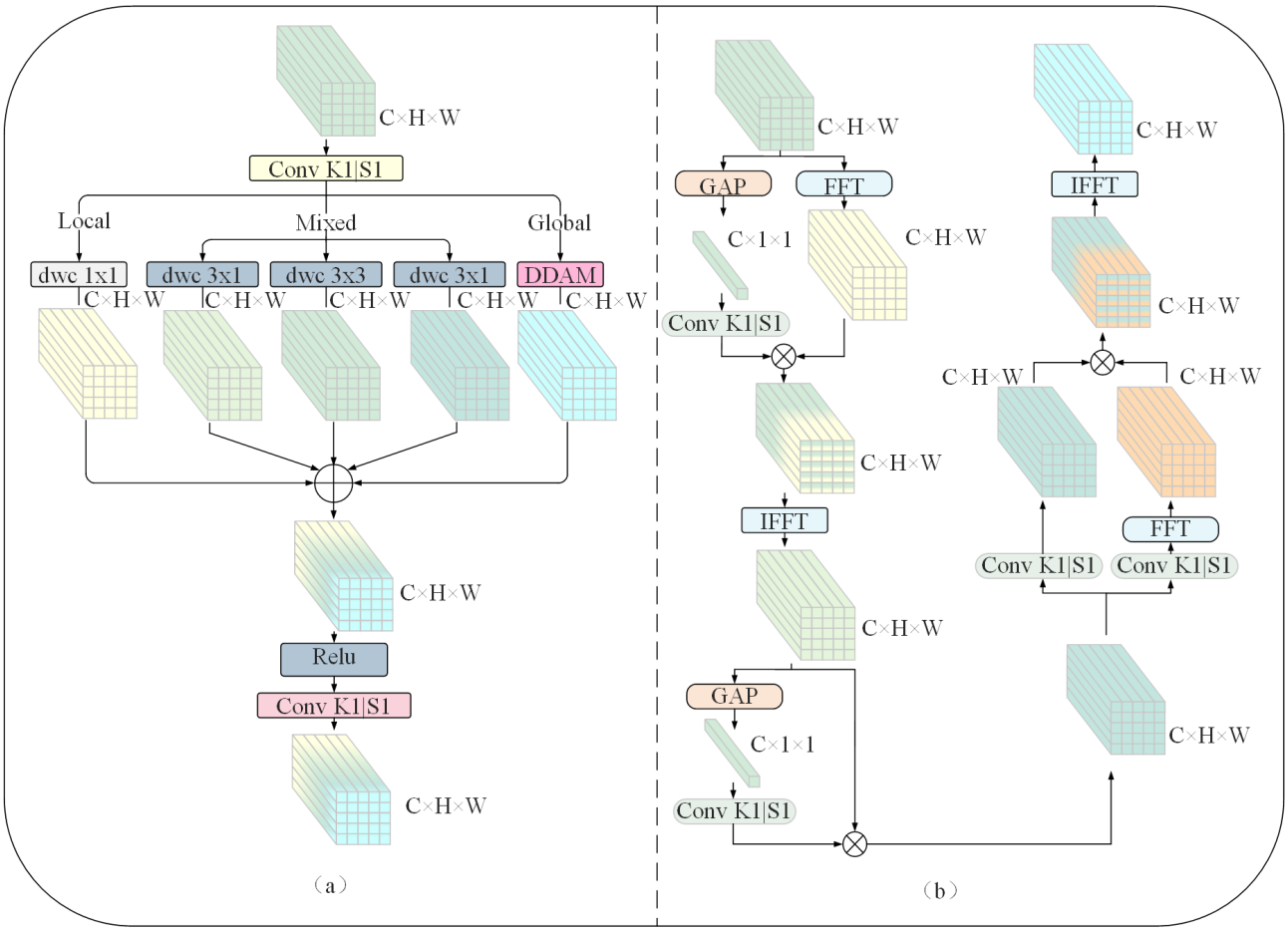

Fig. 4. (a) Multi-Kernel Convolutional Attention Model (MCAM): DWC refers to Depth-wise Convolution. C, K, and S represent the number of channels, convolutional kernels, and stride size, respectively. $C = 192$ when $\lambda = \{128, 256, 512\}$ and $C = 256$ when $\lambda = \{1024, 2048, 4096\}$. $K1|S1$ indicates a kernel size of 1 and a stride of 1. (b) Dual-Domain Attention Module (DDAM): FFT and IFFT denote Fast Fourier Transform and its Inverse Fourier Transform, respectively. GAP represents Global Average Pooling.

$$X_{\text{Global1}} = \mathcal{F}^{-1}\left(\mathcal{F}\left(X_{\text{Global}}\right) \otimes \text{conv}_{1\times1}\left(\text{GAP}\left(X_{\text{Global}}\right)\right)\right) \tag{9}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the Fast Fourier Transform and its Inverse Fourier Transform. GAP denotes Global Average Pooling. $\otimes$ denotes element-by-element multiplication.

Next, the spatial attention is generated using a 1×1 convolution on X_{Global1} to localize important regions in space. The results of the frequency domain enhancement are weighted point by point according to the spatial attention X_{Global1} weights to generate new features X_{Global2}. The spatial features X_1 and the frequency domain features X_2 are then extracted and fused through element-wise multiplication, resulting in features with enhanced global perception capabilities. This process can be expressed using the following formula:

$$\begin{aligned} X_{\text{Global2}} &= X_{\text{Global1}} \otimes \text{conv}_{1\times1}\left(\text{GAP}\left(X_{\text{Global1}}\right)\right), \\ X_1 &= \text{conv}_{1\times1}\left(X_{\text{Global2}}\right), \\ X_2 &= \mathcal{F}\left(\text{conv}_{1\times1}\left(X_{\text{Global2}}\right)\right), \\ X_{\text{out}} &= \mathcal{F}^{-1}\left(X_1 \otimes X_2\right) \end{aligned} \tag{10}$$

where X_{out} denotes the final output of the DDAM model.

In the mixed kernel branch, 1×3 convolution, 3×1 convolution, and 3×3 convolution are used to capture local features in the vertical direction, local features in the horizontal direction, and feature information in a larger range of receptive fields, respectively, so as to make up for the limitations of a single receptive field and to fully integrate information at different scales.

In the local branch, a 1×1 convolution is applied to maintain consistency between the input and output channels, facilitating feature integration. Finally, the convolution results from all branches are summed with the input features and the spatial enhancement results to produce the final output features.

3. Experiments

In this section, the proposed FDEGCNet method was extensively evaluated on multiple standard remote sensing datasets. These datasets include DOTA [25], UC-Merced [26], and NWPU-RESISC45 [27], all of which

contain rich geospatial information, providing an effective benchmark for assessing the performance of the FDEGCNet method. This section compares the results of FDEGCNet with several state-of-the-art traditional image compression methods and deep learning-based image compression methods. The conventional image compression methods include JPEG2000 [6], WebP [7], JPEG XL [28] and AVIF [29], while the deep learning-based methods include those proposed by Cheng et al. (2020) [10], Zou et al. (2022) [22], Jiang et al. (2023) [30], Liu et al. (2023) [31], Liu et al. (2024) [32], and Zhang et al. (2024) [19]. The final experimental results show that the proposed FDEGCNet achieves significant performance improvements in all four metrics (PSNR, MS-SSIM, LPIPS, and VIFP) when compared with existing deep learning-based and traditional methods. These results highlight the superiority of FDEGCNet in remote sensing image compression tasks.

3.1 Datasets

In the experiments, three remote sensing image datasets were used to train and evaluate the proposed model: DOTA [25], UC-Merced [26], and NWPU-RESISC45 [27]. The DOTA dataset contains 2806 images with pixel resolutions ranging from 800×800 to 4000×4000. It includes objects of various scales, orientations, and shapes, ensuring diversity and adaptability in the experiments. The UC-Merced dataset comprises 2100 images across 21 categories of remote sensing scenes, with each image having a resolution of 256×256. The NWPU-RESISC45 dataset consists of 45 different categories of remote sensing images, including airports, deserts, sports fields, forests, and harbors, with 700 images in each category. From each category, 70 images were randomly selected to form the dataset, with each image having a resolution of 256×256. During the experiments, each dataset was randomly divided into a ratio of 8:1:1 for training, validation, and testing. The images were cropped into 256×256 blocks and then processed for subsequent tasks.

3.2 Training Details

To ensure the fairness of the experiments, all models were implemented using Python and the PyTorch framework and developed with the publicly available CompressAI [33] image compression library. All training processes were conducted on an NVIDIA GeForce RTX 3090 GPU. The framework version used was PyTorch 2.1.1, and the CUDA version was 11.8. The compression level of the model was controlled by adjusting the Lagrange multiplier λ, with its values set to $\{128, 256, 512, 1024, 2048, 4096\}$. Six models were trained to correspond to these λ values. Low-bitrate models corresponded to λ values of 128, 256, and 512, with the number of channels set to 192, while high-bitrate models corresponded to λ values of 1024, 2048, and 4096, with the number of channels set to 256. During training, the Adam optimizer [34] was used with an initial learning rate of 1×10^{-4}. After 100k iterations, the learning rate was reduced to 1×10^{-5} and maintained until the end of training. The batch size was set to 8.

3.3 Traditional Codecs

We use the official FFmpeg library obtained from the official website https://ffmpeg.org/ to compress and decompress JPEG2000, WebP, and AVIF. The compression quality is set to {60, 50, 40, 30, 20, 10} for JPEG2000, {1, 10, 20, 30, 40, 50} for WebP, and {60, 55, 50, 45, 35, 30} for AVIF.

For JPEG XL, we use the libjxl library obtained from [28] with the default configuration. The compression quality for JPEG XL is set to {5, 15, 25, 35, 45, 55}.

3.4 Evaluation Strategies

To evaluate the rate-distortion performance of the designed compression model, four commonly used metrics were adopted: PSNR, MS-SSIM, LPIPS, and VIFp. These metrics provide a comprehensive assessment of image reconstruction distortion.

- PSNR: PSNR is based on Mean Squared Error (MSE) to compare the difference between the original image and the compressed reconstructed image, reflecting the degree of distortion. A higher PSNR value indicates that the reconstructed image is closer to the original image, signifying better quality. PSNR is expressed as:

$$\begin{aligned} PSNR\left(X, \widehat{X}\right) &= \frac{1}{C}\sum_{i=1}^{C} 10\log_{10}\left(\frac{\max^2\left(X^i\right)}{MSE_i}\right), \\ MSE\left(X, \widehat{X}\right) &= (1/H \times W \times C)\left\|X - \widehat{X}\right\|^2 \end{aligned} \tag{11}$$

  where $\max^2\left(X^i\right)$ represents the square of the maximum pixel value in the i-th band, and C represents the number of bands.

- MS-SSIM: MS-SSIM is a metric used to measure the similarity of images across multiple scales. By taking the weighted average of image details at different resolutions, it evaluates the differences between the original image and the reconstructed image. The value ranges from [0, 1], where a value closer to 1 indicates higher similarity between the two images [35]. MS-SSIM is expressed as:

$$\begin{aligned} MS\text{-}SSIM &= -10\log_{10}\left(1 - D_{MS\text{-}SSIM}\right), \\ D_{MS\text{-}SSIM} &= 1 - \prod_{m=1}^{M}\left(\frac{2\mu_X\mu_{\widehat{X}} + C_1}{\mu_X^2 + \mu_{\widehat{X}}^2 + C_1}\right)^{\beta_m} \\ &\quad \left(\frac{2\sigma_{X\widehat{X}} + C_2}{\sigma_X^2 + \sigma_{\widehat{X}}^2 + C_2}\right)^{\gamma_m} \end{aligned} \tag{12}$$

  where $D_{MS\text{-}SSIM}$ is a normalized value with a range of 0−1. M represents different scales, μ_X and $\mu_{\widehat{X}}$ represent the mean of the original image and the reconstructed image, σ_X and $\sigma_{\widehat{X}}$ represent the standard deviation between the original image and the reconstructed image, $\sigma_{X\widehat{X}}$ represents the covariance between the original image and the reconstructed image, β_m and γ_m define the

relative weighting of the two components, and C_1 and C_2 are constants introduced to avoid division by 0. Under the same bit rate condition, the larger the value of MS-SSIM, the better the quality of the reconstructed image and the higher the similarity between the reconstructed image and the original image.

- LPIPS: LPIPS is a metric used to measure the perceptual differences between the original image and the reconstructed image. Unlike PSNR and MS-SSIM, LPIPS does not rely solely on pixel-wise comparisons. Instead, it quantifies the perceptual similarity by calculating the distance between the feature maps of the two images. The value ranges from [0, 1], where a lower score indicates that the original and reconstructed images are perceptually very similar, while a higher score indicates significant differences. The formula is expressed as:

$$LPIPS\left(X, \widehat{X}\right) = \sum_{l=1}^{L} \| w^l \odot \left(F^l\left(X\right) - F^l\left(\widehat{X}\right)\right) \|_2 \quad (13)$$

where l represents the network layer, F^l denotes the feature map at the l layer, w^l is the weight of the l layer, and $\odot$ indicates element-wise multiplication.

- VIFP: VIFP is a metric used to measure the degree of visual information retained in the reconstructed image relative to the original image. It calculates the similarity of visual information between the input image and the reconstructed image using a channel model. The value of VIFP ranges from 0 to 1, where a value closer to 1 indicates a higher retention of visual information in the reconstructed image compared to the original image.

$$\begin{aligned} VIFP &= \prod_{i=1}^{N} \frac{I_i^{\mathrm{r}}}{I_i^{\mathrm{o}}}, \\ I_i &= -\sum_{x} p\left(x\right) \log\left(p\left(x\right)\right) \end{aligned} \quad (14)$$

where I_i^{o} is the amount of local information in block i of the original image, I_i^{r} is the amount of local information in block i of the reconstructed image, and N is the total number of image blocks.

3.5 Rate-Distortion Performance

We selected ten state-of-the-art image compression methods for comparison with FDEGCNet, including four traditional image compression methods and six deep learning-based methods. Figures 5–7 show the rate-distortion performance curves obtained with different compression methods on the three datasets [25–27], respectively. It can be observed that, regardless of high or low bitrates, deep learning-based image compression methods significantly outperform traditional methods across all evaluation metrics. Among the four traditional image compression methods, AVIF has better rate-distortion performance than JPEG XL [28], WebP, and JPEG2000. Its performance advantage is mainly based on the efficient compression capability of advanced AV1 codec technology, which effectively removes the redundant information in the image to achieve higher compression efficiency.

For deep learning-based image compression methods, the method [32] obtains a better rate-distortion performance by adaptively convolving different regions according to the mask. However, on the DOTA dataset, the method [32] rate-distortion performance performs poorly, indicating poor robustness of the compression model. The method [10] also achieves good performance due to its more reasonable residual convolution structure and excellent entropy model. In comparison, method [19]is more advantageous mainly because it combines dynamic convolution and Hyper-Prior hybrid attention model. The other comparison methods [22, 30, 31] based on deep learning have average performance, mainly because they lack a strong attention mechanism and excellent rate-distortion optimization strategies. However, the proposed FDEGCNet achieves the best performance of this paper's method compared to method [19], and its advantage is especially evident in high bit rate conditions. This is mainly attributed to the fact that FDEGCNet improves the dynamic convolution by applying the attention mechanism to the four dimensions of the convolution kernel, which can effectively enhance the image feature extraction ability of the main encoder and decoder processes. It also effectively handles multi-scale information through the hyperprior model, and combines the Fourier transform and inverse Fourier transform to simultaneously process the high-frequency local features and low-frequency global information of the image. It provides more accurate and comprehensive prior information for the main encoder and decoder processes, and better realizes the global synergy between global and local information.

To intuitively demonstrate the improvement achieved by different compression methods, this study adopts PSNR-BPP curves to calculate BD-Rate [36] and BD-PSNR as quantitative evaluation metrics. Using JPEG2000 as the baseline anchor point (with a BD-Rate of 0%), we compare the BD-Rate and BD-PSNR results of our proposed method and other methods [7, 10, 19, 22, 29–32] on three datasets. As shown in Tab. 1, compared to JPEG2000, FDEGCNet achieves BD-Rate improvements of 81.325%, 77.354%, and 44.173% on the DOTA, UC-Merced, and NWPU-RESISC45 datasets, respectively, with corresponding BD-PSNR gains of 5.64 dB, 5.745 dB, and 2.061 dB. Compared to the next-best method, FDEGCNet achieves bit rate savings of 3.152%, 3.193%, and 2.652%, with BD-PSNR improvements of 0.273 dB, 0.422 dB, and 0.186 dB, respectively.

To provide a more comprehensive evaluation of the proposed compression method, cross-dataset testing was conducted. The model was trained using the DOTA dataset and tested on the UC-Merced and NWPU-RESISC45 datasets. The compression performance results are shown in Figs. 8 and 9. The proposed model consistently achieves superior performance across all four metrics—PSNR, MS-SSIM, LPIPS, and VIFP—which effectively demonstrates its strong generalization capability.

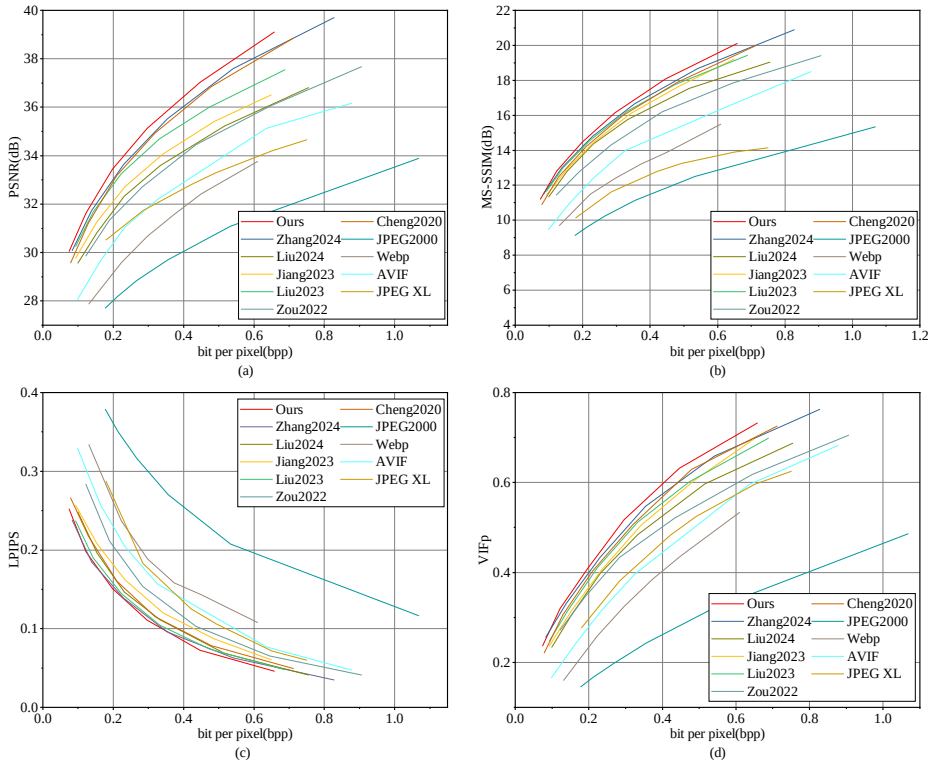

Fig. 5. Rate-distortion performance of different methods trained on the DOTA dataset: (a) Comparison of PSNR results; (b) Comparison of MS-SSIM results; (c) Comparison of LPIPS results; and (d) Comparison of VIFP results.

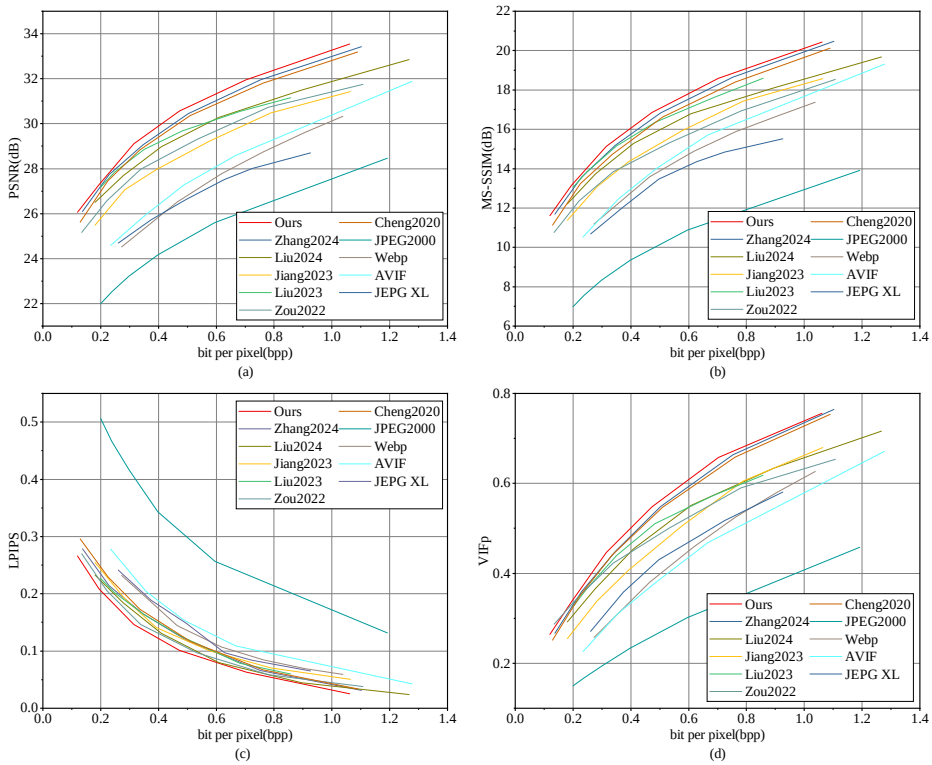

Fig. 6. Rate-distortion performance of different methods trained on the UC-Merced dataset: (a) Comparison of PSNR results; (b) Comparison of MS-SSIM results; (c) Comparison of LPIPS results; and (d) Comparison of VIFP results.

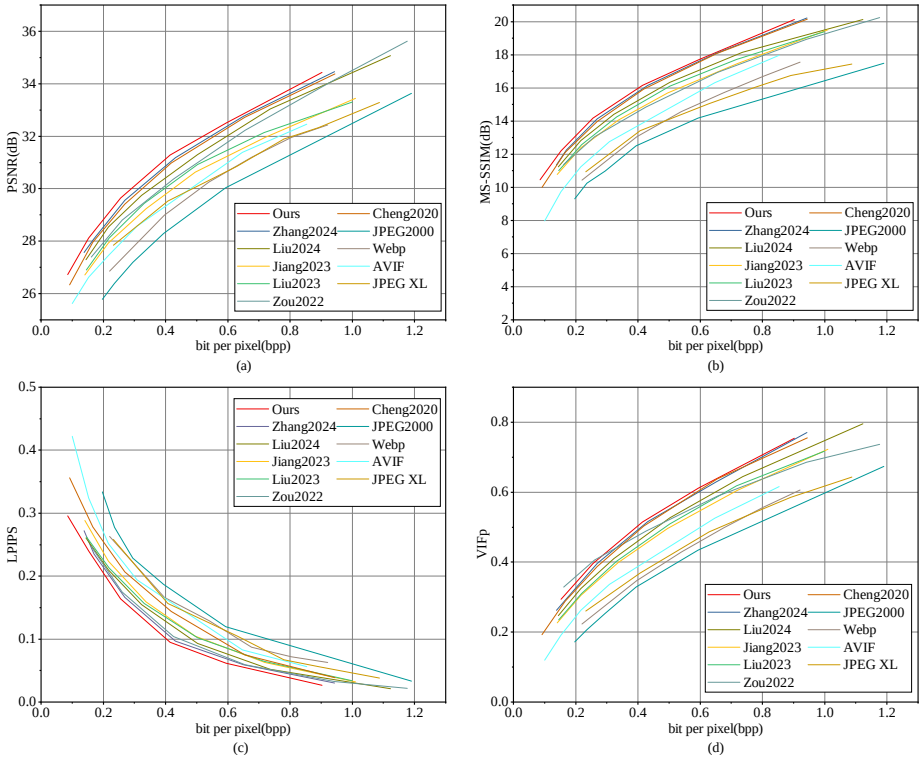

Fig. 7. Rate-distortion performance of different methods trained on the NWPU-RESISC45 dataset: (a) Comparison of PSNR results; (b) Comparison of MS-SSIM results; (c) Comparison of LPIPS results; and (d) Comparison of VIFP results.

Model	DOTA		UC-Merced		NWPU-RESISC45	
	BD-rate	BD-PSNR	BD-rate	BD-PSNR	BD-rate	BD-PSNR
WebP [7]	–32.843%	1.371 dB	–39.253%	2.062 dB	–9.158%	0.467 dB
JPEG XL [28]	–47.033%	2.439 dB	–33.937%	1.631 dB	–8.113%	0.425 dB
AVIF [29]	–55.261%	2.923 dB	–45.991%	2.554 dB	–22.469%	1.035 dB
Cheng2020 [10]	–76.892%	5.196 dB	–74.161%	5.268 dB	–40.333%	2.232 dB
Zou2022 [22]	–67.438%	3.639 dB	–66.604%	4.048 dB	–32.329%	1.777 dB
Jiang2023 [30]	–74.542%	4.320 dB	–61.870%	3.733 dB	–29.251%	1.383 dB
Liu2023 [31]	–78.173%	4.999 dB	–73.464%	4.664 dB	–33.029%	1.623 dB
Liu2024 [32]	–70.789%	3.912 dB	–69.543%	4.462 dB	–34.830%	1.991 dB
Zhang2024 [19]	–77.239%	5.367 dB	–73.841%	5.323 dB	–41.521%	2.415 dB
Ours	**–81.325%**	**5.640 dB**	**–77.354%**	**5.745 dB**	**–44.173%**	**2.601 dB**

Tab. 1. Comparison of BD-rate and BD-PSNR for different methods across DOTA, UC-Merced, and NWPU-RESISC45 datasets.

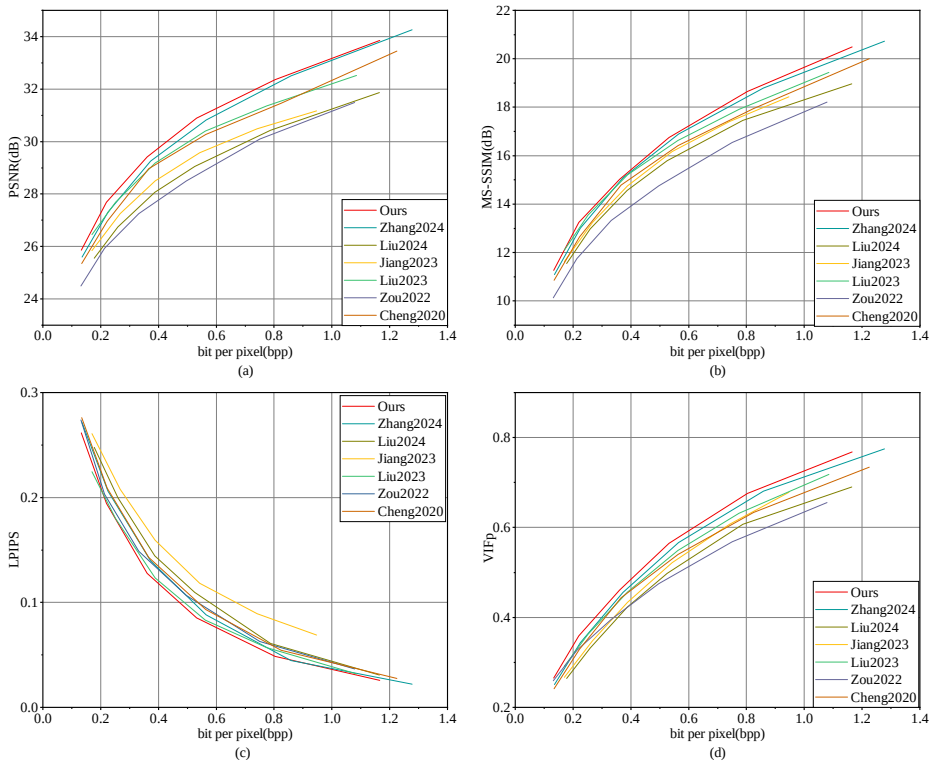

Fig. 8. Rate-distortion performance of different methods trained on the DOTA dataset and tested on the UC-Merced dataset: (a) Comparison of PSNR results; (b) Comparison of MS-SSIM results; (c) Comparison of LPIPS results; and (d) Comparison of VIFP results.

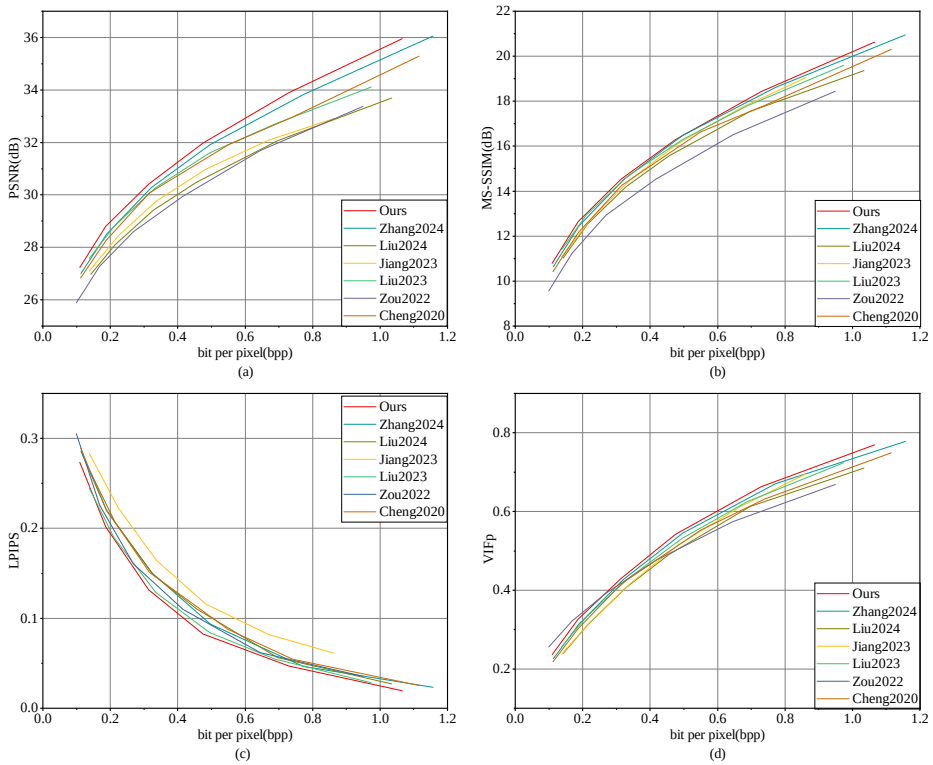

Fig. 9. Rate-distortion performance of different methods trained on the DOTA dataset and tested on the NWPU-RESISC45 dataset: (a) Comparison of PSNR results; (b) Comparison of MS-SSIM results; (c) Comparison of LPIPS results; and (d) Comparison of VIFP results.

3.6 Visualization

To verify the visual quality of reconstructed images using FDEGCNet, one image was selected from each of the DOTA, UC-Merced, and NWPU-RESISC45 datasets. These images were reconstructed using different image compression methods, and the local regions of the reconstructed images were magnified for comparison. Figures 10–12 show the reconstruction results of FDEGCNet and ten other methods on the three datasets, respectively. Figure 10 presents an example of a resort area. By magnifying the regions of the sports field, buildings, and trees in the upper right corner, the reconstruction effects of different image compression methods are compared. Among traditional compression methods, AVIF exhibits significantly better rate-distortion performance than JPEG2000, JPEG XL, and WebP. As shown in Fig. 10, in the AVIF reconstructed image, the white lines on the sports field are clearer, and the outlines and textures of the buildings are more detailed, with natural and smooth color transitions. In contrast, WebP is slightly inferior in detail performance, especially in the area of white lines with high contrast, which shows slight distortion. Images compressed with JPEG XL have color gradients in large areas and slight blurring of the image at the edges. As for JPEG2000, the detail of the image is more blurred, especially the outline of the building is not as clear as WebP and JPEG XL, and the loss of detail in some areas is more obvious.

The excellent compression effect of AVIF is due to the advanced AV1 encoding algorithm, which is able to effectively retain the image details during the compression process while maintaining a high compression ratio. The AV1 algorithm performs particularly well in the retention of high-frequency details, sharp edges, and complex textures, especially in high-contrast areas, and is able to retain details and reduce distortion much better than the traditional JPEG2000, JPEG XL, and WebP methods. Compared to traditional methods, deep learning-based image compression methods achieve significant improvements in both rate-distortion performance and visual quality. However, as shown in Fig. 10, methods [10, 19, 22, 30–32] exhibit some blurriness in the reconstructed sports field region. Compared with method [19], which delivers the best reconstruction among these methods, the proposed method has less blurring, especially in complex areas (e.g., building roofs and ballparks), where the details remain clearer and the sharpness of the image is maintained better. This shows that the proposed method is effective in achieving dynamic enhancement of global detail features and synergistically capturing global information, which results in better visual quality of reconstructed images.

In the UC-Merced and NWPU-RESISC45 datasets, a remote sensing image of a transportation hub was selected for visualization experiments. The local regions of the images, such as highway white lines and building rooftops, were magnified to provide an intuitive comparison of reconstruction quality. Figures 11 and 12 show the reconstruction results of the original image, the proposed method, and ten other comparison methods. The experimental results demonstrate that the proposed method effectively captures the multi-scale detail features and global structural information of the images. Compared with other methods, the reconstructed images produced by the proposed method exhibit superior overall visual quality, with key texture details better preserved. The results show minimal artifacts and noise interference, presenting a more natural and clearer visual effect. As a result, the proposed method outperforms existing approaches, achieving the best performance in terms of image quality, with a noticeable improvement in both subjective visual perception and objective evaluation metrics.

3.7 Ablation Study

To demonstrate the effectiveness of each component, several ablation experiments were conducted. Figure 13 presents the rate-distortion performance results of the ablation experiments on the DOTA dataset. The configurations for the ablation experiments are as follows: 1) baseline: The baseline network serves as the foundational model for the ablation study. 2) baseline + ODAM: ODAM is integrated into the primary encoder-decoder of the baseline model. 3) baseline + MCAM: MCAM is added to the baseline model after the hyperprior encoder and before the hyperprior decoder. 4) baseline + HEAM: HEAM is incorporated into the hyperprior encoder-decoder module of the baseline model. 5) baseline + ODAM + MCAM + HEAM: This represents the integration of ODAM, MCAM, and HEAM into the baseline model. As shown in Fig. 13, the baseline model exhibits the lowest rate-distortion performance. The performance of baseline + ODAM outperforms the baseline at the same bit rate. This indicates that the method enhances the image features by dynamically adjusting the feature map weights using full-dimensional dynamic attention, which improves the rate-distortion performance. Similarly, baseline + MCAM performs significantly better than the baseline. This indicates that multi-scale extraction of local information and the use of frequency domain information to capture the global feature distribution plays an important role. A further comparison between baseline + HEAM and the baseline highlights the critical role of preserving key information across channels for reconstructing remote sensing images. At the same bit rate, the proposed method achieves the best performance in terms of PSNR, MS-SSIM, LPIPS, and VIFP, with its advantages being even more pronounced at higher bit rates. This shows that the method in this paper can efficiently integrate ODAM, MCAM, and HEAM, and can realize information collaboration for high-quality image reconstruction.

3.8 Convergence Analysis

To evaluate the efficiency of our method, we compared the PSNR of all methods with respect to the Epochs, and the results are shown in Fig. 14. The graph shows that our method exhibits the best performance throughout the training process, with a stable PSNR value of around 35.1 dB, which is significantly better than the other methods.

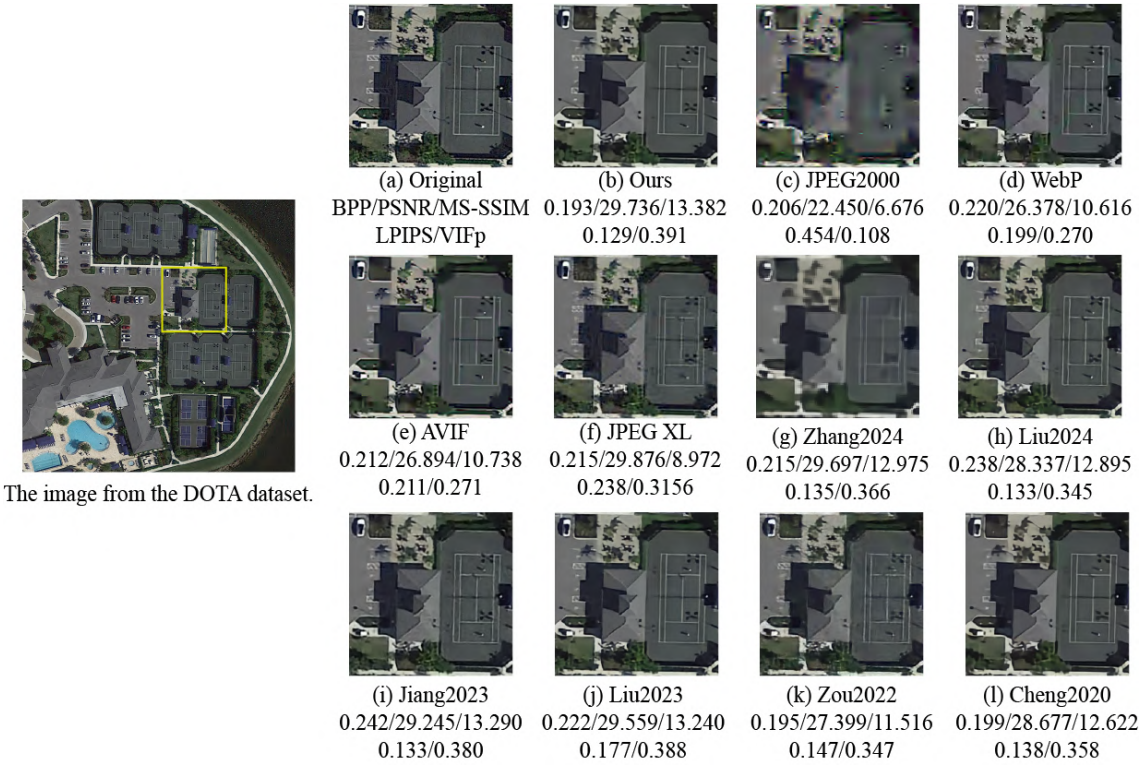

Fig. 10. Visual comparison of reconstructed images by different methods on the DOTA dataset.

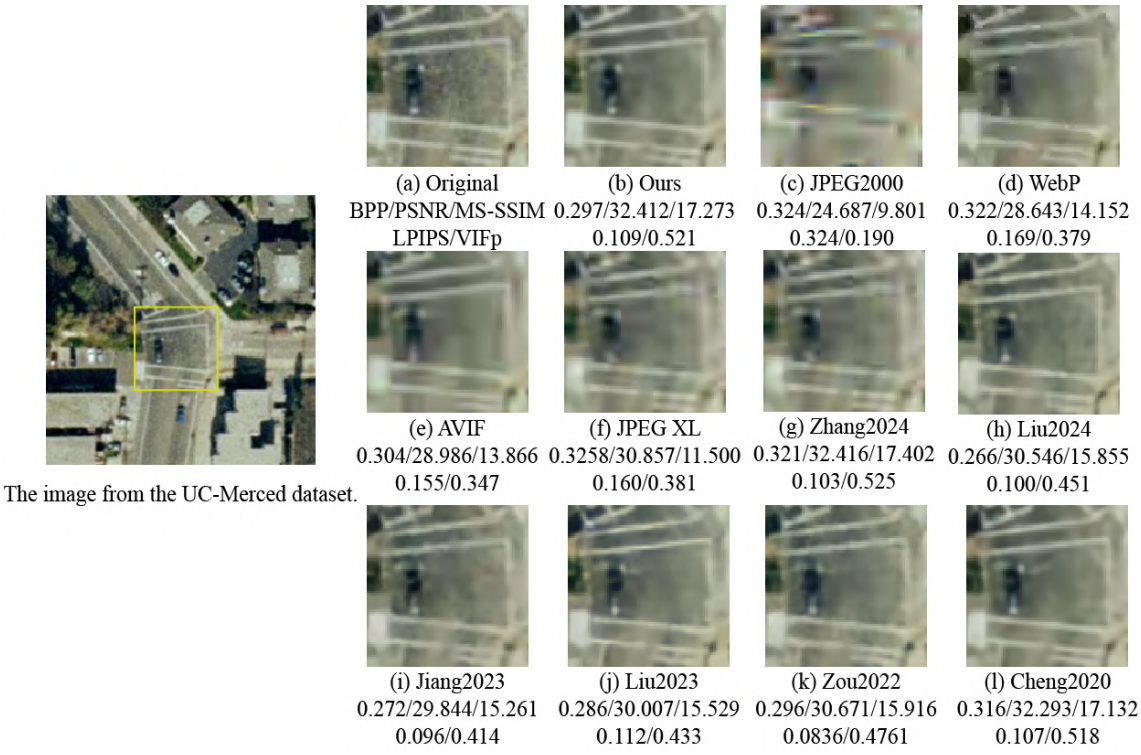

Fig. 11. Visual comparison of reconstructed images by different methods on the UC-Merced dataset.

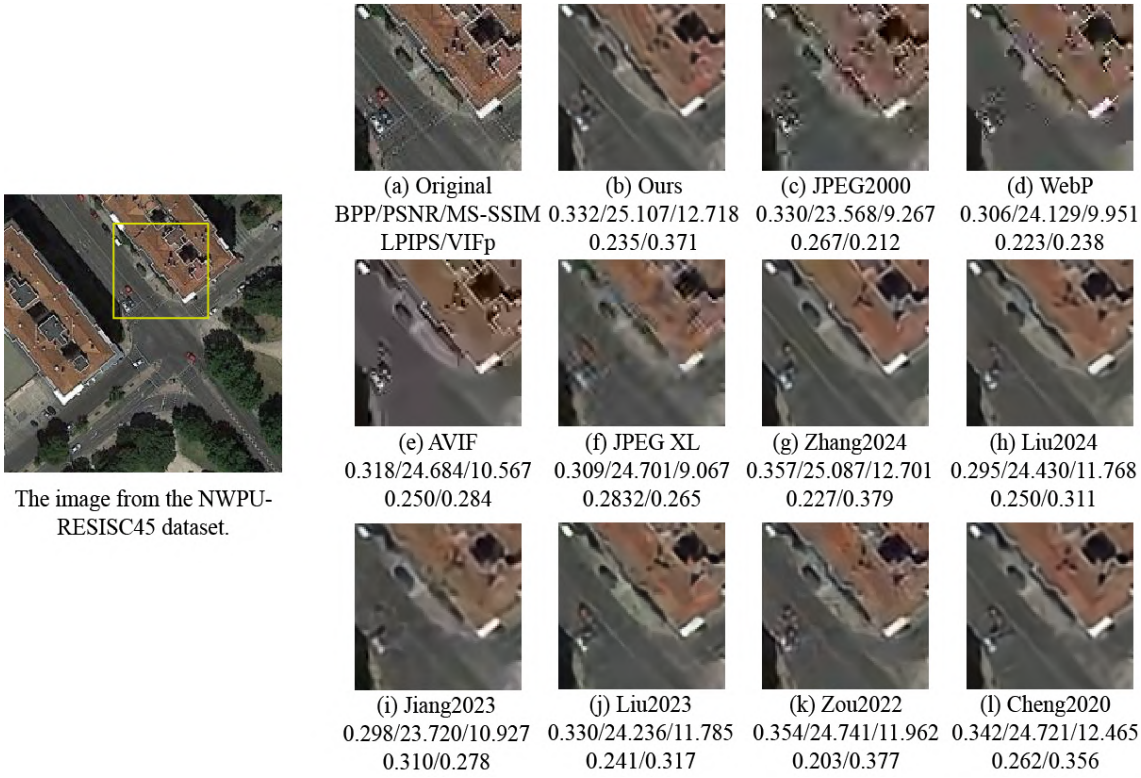

Fig. 12. Visual comparison of reconstructed images by different methods on the NWPU-RESISC45 dataset.

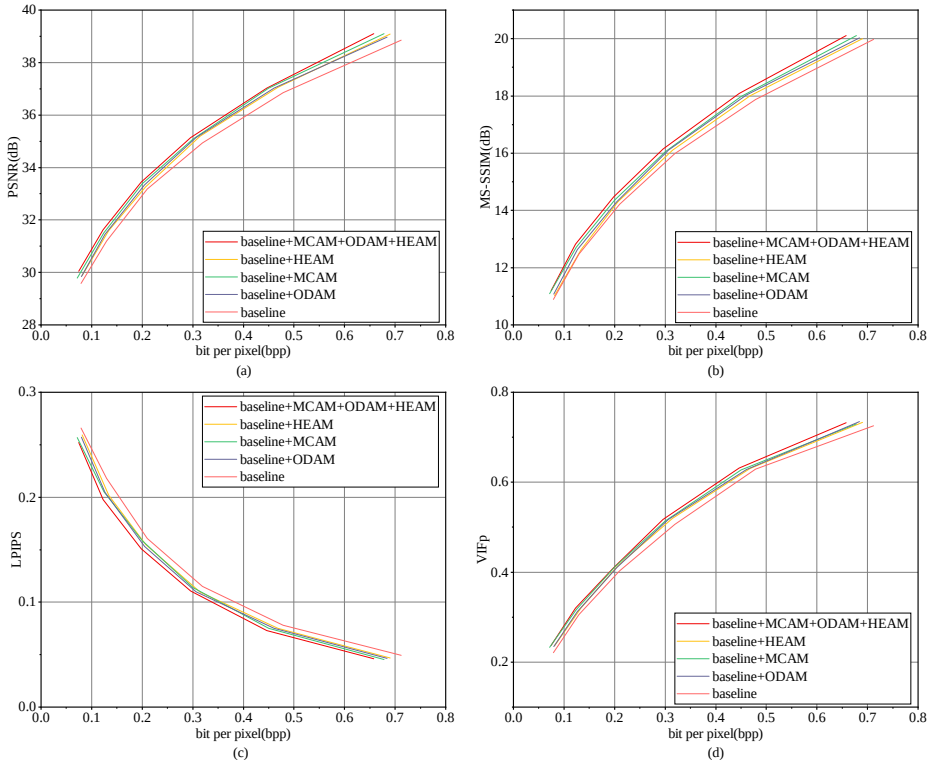

Fig. 13. Ablation results of different methods on the DOTA dataset. (a) Comparison of PSNR results, (b) Comparison of MS-SSIM results, (c) Comparison of LPIPS results, and (d) Comparison of VIFP results.

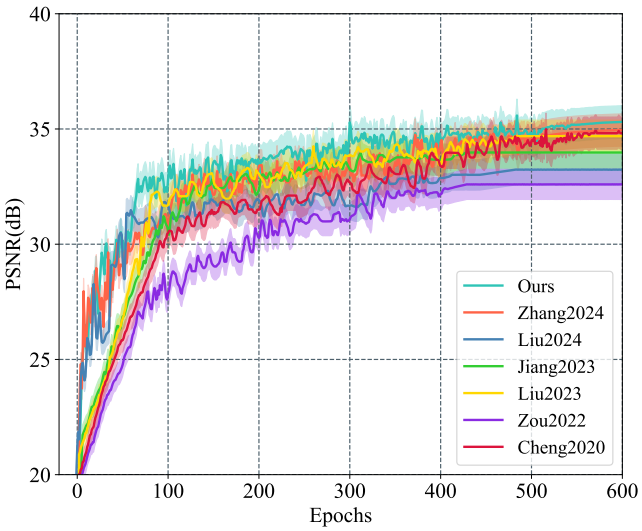

Fig. 14. Relationship between PSNR and Epochs.

Method	Params (M)	FLOPs (G)	Times (s)	
			Encoding	Decoding
Cheng2020 [10]	8.78	27.41	0.1656	0.1502
Zou2022 [22]	75.00	48.70	0.2835	0.2722
Jiang2023 [30]	116.48	82.36	0.3291	0.2843
Liu2023 [31]	75.90	116.78	0.1431	0.1522
Liu2024 [32]	91.17	190.55	0.0910	0.1010
Zhang2024 [19]	12.34	24.91	0.124	0.127
Ours	**17.07**	**54.96**	**0.1991**	**0.1897**

Tab. 2. Complexity comparison of various compression methods.

It validates the effectiveness of ODAM, HEAM, and MCAM modules. In addition, our method demonstrates faster convergence, especially in the early stage of training (0–100 epochs), where the PSNR value quickly improves from 20 dB to more than 32 dB. After training to 500 epochs, its PSNR value stabilizes and fluctuates less, indicating that the method has good stability during training.

Among other methods [10, 19, 22, 30–32], the method [19] shows better reconstruction quality and convergence ability, with a stable PSNR value of around 35 dB, second only to our method. Its excellent performance may be due to its low-complexity transformer-CNN architecture. The design of the architecture may make it easier to converge quickly and maintain high stability in the early stages of training.

In contrast, the method [22] has the worst performance, with a stable PSNR value of only around 32.6 dB. Too short a training period (e.g., only 50 epochs) leads to underfitting, e.g., at this point the PSNR value of method [10] is only 25.5 dB, which is much lower than its final performance (34.8 dB). Conversely, too long a training period (e.g., 600 epochs) may trigger overfitting, especially in methods such as method [22], which manifests itself as a small decrease in the PSNR value. Therefore, a reasonable choice of training period will ensure its efficient convergence capability.

3.9 Complexity Analysis

To compare the computational complexity and Encoding-Decoding Times of different deep learning-based compression methods, the models were evaluated on the test set of the DOTA dataset using metrics such as Parameters, FLOPs, Encoding Times, and Decoding Times. Since hardware conditions and input image size can affect the Encoding-Decoding Times, this study averages the time for all inputs in Tab. 2 with a fixed image size of $3 \times 256 \times 256$. The complexity analysis results are shown in Tab. 2. Compared with other methods, the number of parameters of the proposed method is second only to Cheng2020 and Zhang2024. This means that the model has fewer parameters and takes up less storage space, which make it ideal for resource-constrained environments. In terms of FLOPs, the number of parameters increases by 27.55G, 6.26G, and 4.73G compared with Cheng2020, Zou2022, and Zhang2024, respectively, but decreases significantly by 27.4G, 61.82G, and 135.59G compared with Jiang2023, Liu2023, and Liu2024, respectively. This indicates that the proposed method has moderate computational complexity, which helps to improve the efficiency of the compression process and reduce the consumption of computational resources. In terms of Encoding and Decoding Times, the performance of the proposed method is in the middle of all compared methods. This is because the arrangement of weight calculation may affect the speed of Encoding and Decoding, but it is still within reasonable limits. These experiments demonstrate that the proposed method achieves superior rate-distortion performance with relatively low computational complexity.

4. Conclusion

This paper proposes a novel remote sensing image compression network called the FDEGCNet. The ODAM dynamically adjusts attention weights to direct the network's focus to key image regions. This design improves rate-distortion performance. The HEAM enables multi-scale contextual information extraction. It enhances spatial dependencies in latent representations and improves prior modeling capabilities. Finally, the MCAM is proposed, which utilizes strip depthwise convolutions and standard depthwise convolutions to capture local information while employing dual-domain attention mechanisms to modulate global representations. This enables the HEAM to capture local features and global contextual features. It provides more precise probability distributions to the primary channel and achieves global coordination. Through comparison, our proposed method achieves the best performance across the DOTA, UC-Merced, and NWPU-RESISC45 datasets, compared to Cheng2020, Zou2022, Jiang2023, Liu2023, Liu2024, and Zhang2024. Specifically, on the DOTA dataset at 0.65 bpp, the PSNR of our method improves by 1.83%, 8.21%, 6.88%, 4.67%, 7.97%, and 1.60%, respectively. The results show that the proposed method is highly generalizable and indicate that the introduced modules can be easily integrated into

other network models. In future work, the core modules of FDEGCNet can also be integrated into other application scenarios such as urban building detection. For example, in urban building detection, effectively reducing the amount of data while providing high-quality compressed images will provide clearer input images to the detection model, which in turn will improve the detection accuracy of urban building targets. In addition, we will explore how to realize the adjustment of compression strategies according to different scenes and resolutions. This can ensure the image quality while saving bandwidth and storage space to a greater extent, so as to adapt to diversified application requirements.

References

[1] HUANG, B., ZHAO, B., SONG, Y. Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery. *Remote Sensing of Environment*, 2018, vol. 214, p. 73–86. DOI: 10.1016/j.rse.2018.04.050

[2] LI, J., HOU, X. Object-fidelity remote sensing image compression with content-weighted bitrate allocation and patch-based local attention. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, vol. 62, p. 1–14. DOI: 10.1109/TGRS.2024.3395708

[3] WANG, D., CAO, W., ZHANG, F., et al. A review of deep learning in multiscale agricultural sensing. *Remote Sensing*, 2022, vol. 14, no. 3, p. 1–27. DOI: 10.3390/rs14030559

[4] LU, X., WANG, B., ZHENG, X., et al. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 2017, vol. 56, no. 4, p. 2183–2195. DOI: 10.1109/TGRS.2017.2776321

[5] BASCONES, D., GONZALEZ, C., MOZOS, D. Hyperspectral image compression using vector quantization, PCA and JPEG2000. *Remote Sensing*, 2018, vol. 10, no. 6, p. 1–13. DOI: 10.3390/rs10060907

[6] JPEG COMMITTEE *JPEG2000 Official Software OpenJPEG.* [Online] Cited 2024-11-01. Available at: https://jpeg.org/jpeg2000/software.html

[7] GOOGLE. *WebP Image Format.* [Online] Cited 2024-11-23. Available at: https://developers.google.com/speed/webp/

[8] BELLARD, F. *BPG Image Format*. [Online] Cited 2024-11-25. Available at: http://bellard.org/bpg/

[9] PAN, T., ZHANG, L., SONG, Y., et al. Hybrid attention compression network with light graph attention module for remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 2023, vol. 20, p. 1–5. DOI: 10.1109/LGRS.2023.3275948

[10] CHENG, Z., SUN, H., TAKEUCHI, M., et al. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle (USA), 2020, p. 7939–7948. DOI: 10.1109/CVPR42600.2020.00796

[11] BALLE, J., MINNEN, D., SINGH, S., et al. Variational image compression with a scale hyperprior. *arXiv Preprint*, 2018, p. 1–23. DOI: 10.48550/arXiv.1802.01436

[12] HE, D., YANG, Z., PENG, W., et al. Elic: efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans (USA), 2022, p. 5718–5727. DOI: 10.48550/arXiv.2203.10886

[13] CHEN, T., LIU, H., MA, Z., et al. End-to-end learnt image compression via non-local attention optimization and improved context modeling. *IEEE Transactions on Image Processing*, 2021, vol. 30, p. 3179–3191. DOI: 10.1109/TIP.2021.3058615

[14] LIU, J., YUAN, F., XUE, C., et al. An efficient and robust underwater image compression scheme based on autoencoder. *IEEE Journal of Oceanic Engineering*, 2023, vol. 48, no. 3, p. 925–945. DOI: 10.1109/JOE.2023.3249243

[15] XU, Q., XIANG, Y., DI, Z., et al. Synthetic aperture radar image compression based on a variational autoencoder. *IEEE Geoscience and Remote Sensing Letters*, 2021, vol. 19, p. 1–5. DOI: 10.1109/LGRS.2021.3097154

[16] TANG, Z., WANG, H., YI, X., et al. Joint graph attention and asymmetric convolutional neural network for deep image compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, vol. 33, no. 1, p. 421–433. DOI: 10.1109/TCSVT.2022.3199472

[17] ZHANG, L., HU, X., PAN, T., et al. Global priors with anchored-stripe attention and multiscale convolution for remote sensing images compression. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023, vol. 17, p. 138–149. DOI: 10.1109/JSTARS.2023.3326957

[18] PAN, T., ZHANG, L., QU, L., et al. A coupled compression generation network for remote-sensing images at extremely low bitrates. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, vol. 61, p. 1–14. DOI: 10.1109/TGRS.2023.3270271

[19] ZHANG, L., WANG, X., LIU, J., et al. A low-complexity transformer-CNN hybrid model combining dynamic attention for remote sensing image compression. *Radioengineering*, 2024, vol. 33, no. 4, p. 642–659. DOI: 10.13164/re.2024.0642

[20] LI, C., ZHOU, A., YAO, A. Omni-dimensional dynamic convolution. In *Proceedings of the International Conference on Learning Representations (ICLR)*. Virtual Event, 2022, p. 1–20. DOI: 10.48550/arXiv.2209.07947

[21] BALLE, J., LAPARRA, V., SIMONCELLI, E. P. End-to-end optimized image compression. *arXiv Preprint*, 2016, p. 1–14. DOI: 10.48550/arXiv.1611.01074

[22] ZOU, R., SONG, C., ZHANG, Z. The devil is in the details: Window-based attention for image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans (US), 2022, p. 17492–17501. DOI: 10.48550/arXiv.2203.08450

[23] KIM, J.-H., HEO, B., LEE, J.-S. Joint global and local hierarchical priors for learned image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans (US), 2022, p. 5992–6001. DOI: 10.48550/arXiv.2112.04487

[24] CUI, Y., REN, W., KNOLL, A. Omni-kernel network for image restoration. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, no. 2, p. 1426–1434. DOI: 10.1609/aaai.v38i2.27907

[25] DING, J., XUE, N., XIA, G.-S., et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, vol. 44, no. 11, p. 7778–7796. DOI: 10.1109/TPAMI.2021.3117983

[26] YANG, Y., NEWSAM, S. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*. San Jose (USA), 2010, p. 270–279. DOI: 10.1145/1869790.1869829

[27] CHENG, G., HAN, J., LU, X. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 2017, vol. 105, no. 10, p. 1865–1883. DOI: 10.1109/JPROC.2017.2675998

[28] JPEG. *JPEG XL Reference Software*. [Online] Cited 2024-11-25. Available at: https://gitlab.com/wg1/jpeg-xl/

[29] AOM WORKING GROUP. *AV1 image file format (AVIF)*. [Online] Cited 2024-12-10. Available at: https://aomediacodec.github.io/av1-avif/

[30] JIANG, W., YANG, J., ZHAI, Y., et al. MLIC: Multi-reference entropy model for learned image compression. In *Proceedings of the 31st ACM International Conference on Multimedia*. Ottawa (Canada), 2023, p. 7618–7627. DOI: 10.1145/3581783.3611694

[31] LIU, J., SUN, H., KATTO, J. Learned image compression with mixed transformer-cnn architectures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vancouver (Canada), 2023, p. 14388–14397. DOI: 10.1109/CVPR52729.2023.01383

[32] LIU, Y., YANG, W., BAI, H., et al. Region-adaptive transform with segmentation prior for image compression. *arXiv Preprint*, 2024, p. 1–19. DOI: 10.48550/arXiv.2403.00628

[33] BEGAINT, J., RACAPE, F., FELTMAN, S., et al. CompressAI: A PyTorch library and evaluation platform for end-to-end compression research. *arXiv Preprint*, 2020, p. 1–19. DOI: 10.48550/arXiv.2011.03029

[34] KINGMA, D. P., BA, J. Adam: A method for stochastic optimization. *arXiv Preprint*, 2014, p. 1–15. DOI: 10.48550/arXiv.1412.6980

[35] WANG, Z., SIMONCELLI, E. P., BOVIK, A. C. Multiscale structural similarity for image quality assessment. In *Proceedings of the Thirty-Seventh Asilomar Conference on Signals, Systems & Computers*. Pacific Grove (USA), 2003, p. 1398–1402. DOI: 10.1109/ACSSC.2003.1292216

[36] BJONTEGAARD, G. *Calculation of Average PSNR Differences Between RD-Curves*. ITU SG16 Doc. VCEG-M33, 2001. [Online]. Available at: https://www.itu.int/wftp3/av-arch/video-site/0104_Aus/VCEG-M33.doc

About the Authors . . .

Qizhi FANG received a B.E. degree from Shenyang Aerospace University, Shenyang, China, in 2002 and received a M.E. degree from Northeastern University, Shenyang, China, in 2008. He is currently an Associate Professor at the College of Electronic Information Engineering, Shenyang Aerospace University and Liaoning General Aviation Academy, Shenyang, China. His current research interests include image compression, radar-based human activity classification, and SAR-based ship detection.

Shibo GU received a B.E. degree from Shenyang Aerospace University, Shenyang, China, in 2023. He is currently a postgraduate student at the College of Electronic Information Engineering, Shenyang Aerospace University, Shenyang, China. His current research interests include deep learning and image compression.

Jingang WANG received a B.E. degree from Shenyang Aerospace University, Shenyang, China, in 2023. He is currently a postgraduate student at the College of Electronic Information Engineering, Shenyang Aerospace University, Shenyang, China. His current research interests include deep learning and multispectral image compression.

Lili ZHANG received her B.E., M.E., and Ph.D. degrees from Jilin University, Changchun, China, in 2002, 2005, and 2012. She is currently an Associate Professor at the College of Electronic Information Engineering, Shenyang Aerospace University, Shenyang, China. Her current research interests include image compression, radar-based human activity classification, and SAR-based ship detection.