Multihead Global Attention and Spatial Spectral Information Fusion for Remote Sensing Image Compression

Cuiping Shi, *Member, IEEE*, Kaijie Shi, Fei Zhu, Zexin Zeng, and Liguo Wang, *Member, IEEE*

***Abstract*—In recent years, convolutional neural network (CNN) based methods have been widely used in remote sensing image compression tasks. However, CNN is commonly used to extract local information and does not fully utilize global contextual information. The transformer model can effectively extract the latent contextual information in remote sensing images due to its multihead self-attention mechanisms. Due to the multiscale local features and global low-frequency information of remote sensing images, existing deep-learning-based compression methods have not effectively combined CNN and transformer. In order to overcome the limitations of the above methods, a multihead global attention and spatial spectral information fusion network (MGSSNet) is proposed for remote sensing image compression. First, a spatial spectral information fusion attention module (SSIF-AM) is constructed to obtain multiscale local information. Second, a multihead global attention module (MHG-AM) is proposed to capture rich global context information. Third, a local global collaboration module is developed to explore the correlation between the multiscale local features obtained by SSIF-AM and the global visual features obtained by MHG-AM, and to efficiently model the intrinsic relationships between them to achieve effective feature fusion. Experimental results show that compared with advanced compression models, the proposed MGSSNet method achieves better compression performance. In addition, using reconstructed images obtained by different compression methods for classification tasks has proven that the proposed method can help achieve better classification performance, indicating that the proposed compression method can more fully preserve important information in the image.**

***Index Terms*—Attention network, compression, deep learning, remote sensing images, variational autoencoder (VAE).**

I. Introduction

REMOTE sensing images are digital representations of information about the Earth's surface obtained from satellites, or other sensors. Because of its ability to reflect a variety of material properties, remotely sensed imagery is widely used in many fields, including Earth science, environmental monitoring, urban planning, atmospheric science, and agriculture [1], [2], [3], [4]. However, the increase in spatial and spectral resolution means that more pixels and bits are needed to represent the data, which can lead to a sharp increase of the amount of data in remote sensing images, causing serious transmission or storage problems for remote sensing satellites and users. Therefore, the compression of remote sensing images is of great significance. Compared with natural images, remote sensing images have a larger imaging angle, which makes remote sensing images contain more ground object information and the content is more complex. Although there are many image compression technologies, most of them cannot effectively process both multiscale local information and global information in the image simultaneously, making it difficult for these methods to achieve high-quality compression of remote sensing images.

The huge amount of data in remote sensing images requires efficient compression methods for processing. There are three main categories of traditional image compression algorithms, including vector quantization based methods, predictive coding based methods, and transformation coding based methods. The theoretical basis of vector quantization coding is Shannon's rate distortion theory. In order to reduce the complexity of compression methods, Qian proposed a fast vector quantization compression method to achieve good compression performance. The vector quantization encoding method is to change the input vector into a codeword index that matches the input vector in the codebook for data transmission and storage; during decoding, just search the codebook [5]. Three-dimensional (3-D)-multiband linear predictor (MBLP) is a prediction-based technology that first removes spatial redundant information, then predicts the current frequency band, and finally uses an entropy decoder to encode the prediction residual [6]. 3-D-set partitioning in hierarchical trees (SPIHT) is a method for 3-D image compression transformation, which applies 3-D wavelet transform to the spatial and spectral domains [7]. Most of the traditional methods are based on transformation methods, which have high algorithm complexity and do not take into account the unique

Manuscript received 2 May 2024; revised 8 June 2024; accepted 18 June 2024. Date of publication 21 June 2024; date of current version 8 July 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 42271409 and Grant 62071084, and in part by the Qiqihar University Graduate Innovative Research Project under Grant QUZLTS_CX2023025. *(Corresponding author: Cuiping Shi.)*

Cuiping Shi is with the Department of Communication Engineering, Qiqihar University, Qiqihar 161000, China, and also with the College of Information Engineering, Huzhou University, Huzhou 313000, China (e-mail: shicuiping@qqhru.edu.cn).

Kaijie Shi, Fei Zhu, and Zexin Zeng are with the Department of Communication Engineering, Qiqihar University, Qiqihar 161000, China (e-mail: 2022910313@qqhru.edu.cn; 2022935750@qqhru.edu.cn; 2022910311@qqhru.edu.cn).

Liguo Wang is with the College of Information and Communication Engineering, Dalian Nationalities University, Dalian 116000, China (e-mail: wangliguo@hrbeu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3417690

characteristics of remote sensing images, which can lead to unsatisfactory remote sensing image compression performance.

In recent years, many researchers have focused on the increasingly development of deep-learning technologies. Among them, the most widely used is convolutional neural network (CNN) [8]. Compared with recurrent neural network [9] and generative adversarial network (GAN) [10], CNN processes information very similarly to the human visual system, which makes CNN plays a huge role in the field of computer vision. In contrast, GAN training is complex, and the training process is prone to oscillation and nonconvergence. The above methods cannot accurately model the relationship between multiscale local information and global visual features. If the above methods are directly used for compression, it is difficult to consider multiple feature information at the same time, and can only achieve suboptimal rate distortion performance.

At present, many learning-based image compression frameworks are constructed by combining neural networks and traditional compression methods. For these methods, the input image block is first downsampled multiple times through CNN to map the pixel data into a quantified representation. Then, some coding methods, such as arithmetic encoding, are used to further compress the data, resulting in less data. The goal of compression is to reduce the entropy of the entropy model between the sender and the receiver, so some compression models will add entropy models to the framework (such as single-core Gaussian model, mixed Gaussian model, Laplacian model, factorized entropy model, conditional entropy model, etc.) to introduce prior information to conduct more accurate modeling [11], [12], [13], [14]. Image compression consists of two main tasks, including efficient compression of images and high-quality reconstruction of images. These two parts are closely related, and in order to achieve high-performance compression, the model needs to fully consider the relationship between image encoding and decoding. Due to the reversal-like relationship between the two tasks, it is required that the encoding and decoding parts have a similar and symmetrical structure. Therefore, autoencoders (AE) and variational autoencoders (VAE), which have both symmetric structures and excellent image reconstruction capabilities, have attracted widespread interest. Some works have designed CNN-based image compression models that utilize AE to construct an end-to-end structure, mainly to learn a reversible mapping relationship that can convert pixels into quantifiable latent representations [15], [16], [17]. However, the mapping space of a fully automatic encoder is not continuous, and can only perform one-to-one mapping of inputs and outputs, which can result in its latent representation space being discontinuous and the reconstructed image pixels being excessively rough. In contrast, VAE has a continuous latent space, which helps to reconstruct images with smooth transitions. In addition, VAE can generate high-quality reconstructed images with more detailed information by learning the probability distribution of images [18], [19], [20]. If VAE is used for compression and reconstruction of remote sensing images, it is necessary to combine the imaging characteristics of the remote sensing images and design a reasonable compression method.

In the field of computer vision, CNNs are currently the most popular image learning techniques, mainly due to their powerful feature extraction capabilities [21], [22], [23]. Remote sensing images contain rich spatial information, and the use of CNNs can effectively remove spatial redundant data, thereby improving compression performance. Ma et al. [21] proposed the iWave framework for creating wavelet-like transforms suitable for image compression. It uses CNN for training, embedded in deep networks, and supports multiscale analysis. Tang et al. [23] proposed an end-to-end image compression method by combining graph attention and asymmetric CNN. To a certain extent, this method overcomes the excessive attention of traditional CNNs to local features, promotes information interaction, and takes into account the relationship and location information at the channel level. Although CNN-based methods have demonstrated excellent ability to extract spatial information and local contextual information, it cannot be ignored that these methods still have some limitations. On the one hand, CNN has a small receptive field. If you want to obtain long-distance contextual information, you need to use convolutions with large-size convolutional kernels for spatial information extraction. However, increasing the size of the convolution kernels can lead to a sharp increase in parameter amount and computational complexity. When there are long-distance features, such as rivers and bridges in the image, the model inevitably encounters the bottleneck of compression performance. On the other hand, CNNs utilize convolutional filters to extract potential features in the local receptive domain, which leads to the network paying too much attention to the local features of the image and reducing the attention to the global visual features.

In recent years, transformer-based image data processing methods have attracted much attention in the field of computer vision [24], [25], [26], [27]. Among them, vision transformer (ViT) has sparked a revolution in the field of image processing. It divides images into small pieces and then uses self-attention to capture the global latent representation, achieving powerful ability to capture long-distance contextual relationships. Li et al. [28] proposed a learning image compression network based on visual transformer by dividing the input image into blocks and using different types of transformer blocks in the encoder and decoder to achieve efficient image compression. At the same time, using a strategy based on residual coding, this method achieves a peak signal-to-noise ratio (PSNR) improvement of 0.75 dB at 0.15 bpp on the Kodak dataset, and has better performance on multiscale structural similarity index metric (MS-SSIM) at low bit rates. Although these methods can extract the global feature information of images, they do not make good use of local semantic features and feature information of different scales, which may result in the reconstructed image not showing good performance in terms of detailed features.

Although current-learning-based image compression methods have made great progress [29], [30], [31], [32], existing compression methods still have the following problems. The first problem is that the resolution of remote sensing images is generally meter-level, which leads to lower spatial continuity between adjacent pixels in remote sensing images compared with ordinary images. In addition, image compression methods typically rely on the continuity between adjacent pixels, and lower pixel continuity may lead to more data redundancy, resulting in a decrease in compression performance. The second problem is that, unlike ordinary images, remote sensing images often contain features of different scales of land cover, including small-scale features, such as vehicles and buildings, as well as large-scale geographic features that run through the entire image, such as rivers, roads, and oceans. Therefore, how to effectively capture both small- and large-scale features simultaneously is a worthwhile research question. At present, in order to solve this problem, researchers have proposed several methods, including multiscale analysis and multimodal data fusion [33], [34], aimed at effectively associating small-scale and large-scale features. However, these methods do not fully consider global information. The third problem is that remote sensing images not only contain local information at various scales, but also global information, such as terrain and landforms. If global information is ignored, the model may lose its understanding of important context, such as the distribution of ground objects and geomorphic features, resulting in artifacts in the reconstructed image. Currently, researchers often use large-sized convolution or pyramid networks to capture global features, but these methods do not fully consider multiscale local information [35]. Finally, the current remote sensing image compression frameworks often use fully factorized entropy models, which will lead to some statistical redundancy information being retained in the distribution of latent representations, resulting in poor compression performance.

In this article, a remote sensing image compression network based on multihead global attention and spatial spectral information fusion is designed. In order to solve the problem of poor interpixel continuity of remote sensing images, the proposed multihead global attention and spatial spectral information fusion network (MGSSNet) selects VAE as the basic framework and uses its continuous latent space to generate smooth transition images, so as to reconstruct high-quality images with more detailed features. In order to capture both the small- and large-scale features in the remote sensing images, a local attention module [spatial spectral information fusion attention module (SSIF-AM)] is constructed by introducing bar convolution and pooling layers of different scales into the local attention block. In addition, in order to increase the receptive field without increasing the additional computational burden, the bar convolution is utilized in SSIF-AM instead of the ordinary convolution. At the same time, this article constructs a multihead global attention module (MHG-AM) by using the multihead self-attention (MHSA) mechanism to capture global context information. In general, CNN acts on the local acceptance domain in the form of convolutional filters, which makes the features extracted by CNN contain more high-frequency information, such as edges, contours, and texture information [36]. In contrast, self-attention is generally used to capture global information, and therefore it is considered a low-pass filter [37]. Under the guidance of self-attention mechanism, the network can adaptively focus on different regions of the image. This mechanism is beneficial for capturing the relationships between a wide range of features of an image, making the image more visually coordinated. Considering the complementarity of CNN and self-attention in feature extraction, the integration of these two modules is helpful to fully extract the local features of multiscale and the global features of the whole image. Therefore, this article designs a local global collaboration module (LGCM) to explore the correlation between the multiscale local features obtained by SSIF-AM and the global visual features obtained by MHG-AM, so as to efficiently model the intrinsic relationship between them and achieve effective integration. In addition, general encoding methods directly assume the probability model of image blocks or even the entire image, while this article adopts a layered prior method. The core idea of this method is to learn the probability distribution model of each quantized representation by capturing edge information to generate a more accurate entropy encoding model. Through this layered prior method, MGSSNet can better adapt to the features of different images, improving the efficiency and quality of image compression. In order to verify the effectiveness of the proposed architecture in image compression, a rate distortion optimization strategy is adopted in this article, which can be represented as

$$\begin{aligned}\text{Minloss} &= R + \lambda \cdot D \\ &= \underbrace{\mathbb{E}_{X\sim p_X}[-\log_2 p_{\hat{y}}(Q(M_a(X)))]}_{\text{rate}} \\ &\quad + \underbrace{\mathbb{E}_{X\sim p_X}[-\log_2 p_{\hat{z}}(\hat{z})]}_{\text{rate}} \\ &\quad + \lambda \cdot \underbrace{\mathbb{E}_{X\sim p_X}[d(X,\hat{X})]}_{\text{distortion}}\end{aligned} \tag{1}$$

where p_X is the unknown distribution of the image, Q is the quantification, M_a is the main encoder, $\hat{y} = Q(y)$ is the quantified latent representation, $p_{\hat{y}}$ is the entropy model that can be learned, X is the chunks of remote sensing image data passed into the network, and $\hat{X}$ is the reconstructed image. rate is cross-entropy between the latent marginal distribution and the learned entropy model. $d(X,\hat{X})$ is the loss between the original image and the reconstructed image and $d()$ denotes mean square error (MSE). To put it simply, R represents the entropy rate, D represents the distortion between the original image and the reconstructed image, and different bitrates can be controlled by adjusting the penalty coefficient λ.

In this article, a large number of experiments were carried out on the remote sensing image dataset San Francisco [38] and Northwestern Polytechnical University (NWPU)-RESISC45 [39]. Experimental results show that compared with some image learning compression methods, i.e., Minnen et al. [14], Minnen et al. [14] (mean), Balle et al. [40] (hyperprior), Balle et al. [40] (factorized-relu), Cheng et al. [11], and some traditional image compression methods, including JPEG2000 [41], WebP [42], and BPG [43], the compression performance of the proposed network can provide better performance in PSNR and MS-SSIM. In addition, a variety of ablation experiments were conducted to verify the effectiveness of the proposed SSIF-AM, MHG-AM, and LGCM, respectively.

The main contributions of this article are summarized as follows.

1) In this article, we propose an SSIF-AM that can effectively capture both small- and large-scale features of images.

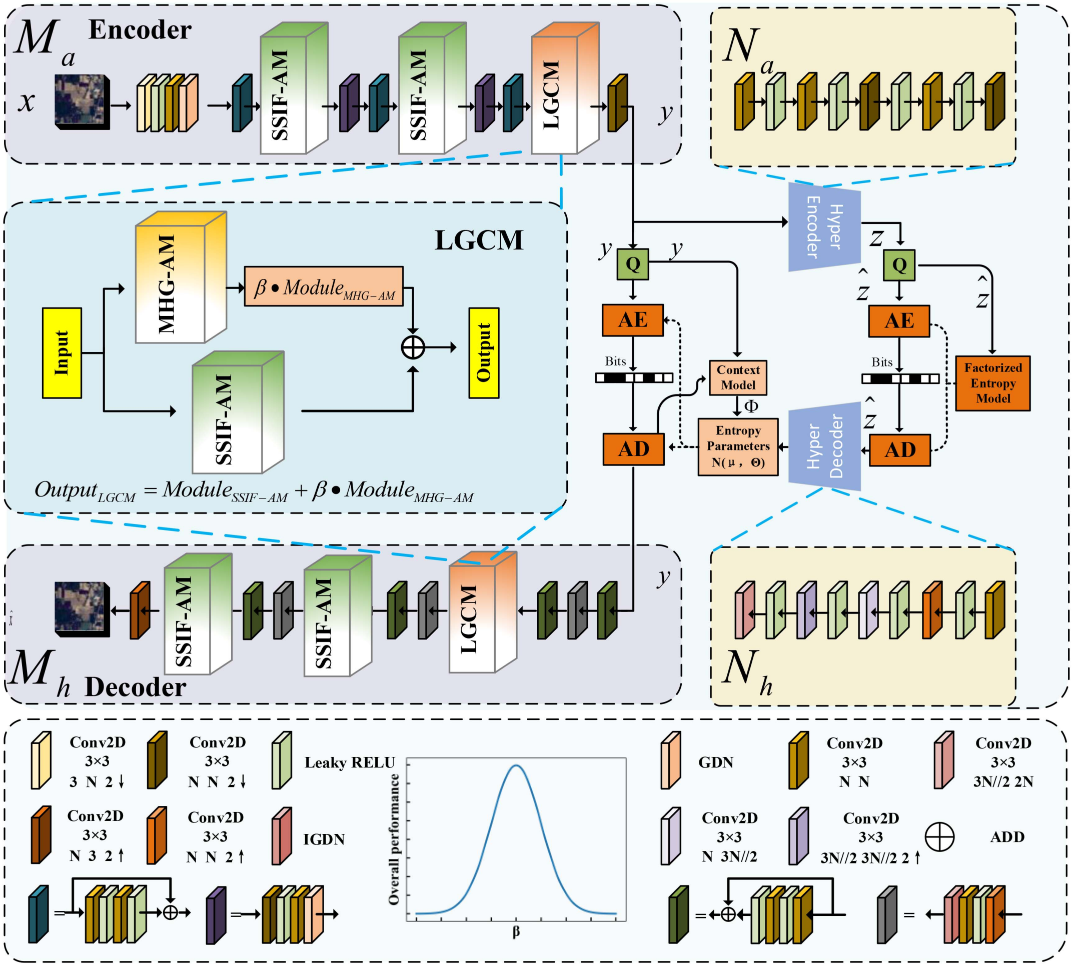


Fig. 1. Overall structure diagram of proposed MGSSNet (in the overall performance curve for β, β represents a harmony hyperparameter, and "overall performance" refers to various performance metrics, such as PSNR and MS-SSIM).

Through the bar convolution of different scales, SSIF-AM enables the model to extract potential representation features at different levels while increasing the receptive field.

2) In order to better understand the contextual information of images, an MHG-AM was constructed. The module realizes the accurate capture of global context information through an MHSA mechanism, which enables the compression model to better preserve the overall features of the image, reduce the probability of artifacts, and thus improve the quality of reconstructed images.
3) In order to effectively fuse high-frequency local information and low-frequency global information, a harmonic hyperparameter β was introduced, and an LGCM was constructed. This article effectively explores the correlation between multiscale local information and global visual features, and efficiently models the intrinsic relationship between them, achieving effective feature fusion.
4) In this article, the designed SSIF-AM, MHG-AM, LGCM, context model, and factorized entropy model are effectively embedded into the VAE framework, and a MGSSNet is proposed for remote sensing image compression. The MGSSNet can effectively extract the abundant high-frequency and low-frequency information in the image, and realize the efficient compression of remote sensing images.

In this article, the good compression performance of MGSSNet was verified by sufficient experiments on San Francisco and NWPU-RESISC45. The rest of this article is organized as follows. In Section II, SSIF-AM, MHG-AM, LGCM, and the components of MGSSNet are introduced in detail. In Section III, the experimental results and analysis are provided. Section IV makes a summary of this article. Finally, Section V concludes this article.

II. Methodology

In this section, we will introduce the proposed SSIF-AM, MHG-AM, LGCM, and MGSSNet in detail.

A. The Overall Framework of the Proposed MGSSNet

In this article, VAE is utilized as the basic framework to design the MGSSNet for remote sensing image compression. The overall structure diagram is shown in Fig. 1. The structure is divided into two parts. For the image compression part, the encoder maps the remote sensing image data block as a latent representation feature by combining SSIF-AM, MHG-AM, LGCM, and convolutional blocks. Then, the quantization, arithmetic coding, text context module, entropy parameter, and factorized entropy model are utilized to further remove the statistical redundancy, and the minimum bit stream after data processing of the model is obtained. For the image decoding part, the reconstructed image with high quality is obtained through arithmetic decoding, text context module, entropy parameter module, factorized entropy model, and decoder.

Specifically, according to the spatial information distribution characteristics of remote sensing images, this article proposes SSIF-AM based on multiscale bar convolution, MHG-AM based on MHSA, and LGCM with efficient fusion of multiscale local information and global visual features. SSIF-AM uses bar convolution of different scales, point convolution, maximum pooling layer, and average pooling layer to extract multiscale information and spatial spectral fusion information from images. Moreover, MHG-AM retains the equally important global context information, which is based on MHSA and embeds local information into the global information by introducing convolutional branches, thereby obtaining latent representation features that are more conducive to image reconstruction. In addition, the proposed LGCM explores the correlation between multiscale local information and global visual features, and efficiently models the intrinsic relationship between them to achieve effective fusion. Then, the context model for capturing contextual information, as well as mean and scale parameters for generating hyperprior information, are added to the network model. Further, unlike traditional entropy coding that directly assumes image models, this article designs a factorized entropy model to learn each quantized representation by obtaining side information, making the entropy model more accurate. Finally, GDN is also adopted in this model, which is a nonlinear activation function. Compared to other normalization functions, GDN is more suitable for image compression.

In Fig. 1, M_a and M_h are the main encoder and main decoder, respectively, which are used to learn the latent representation features of the remote sensing image. N_a and N_h are the hyper encoder and hyper decoder, respectively. In this article, the hyperprior network is adopted to learn the entropy model on which entropy coding depends. It is also used to generate the parameters of the entropy model. Q represents the quantizer, and AE and AD represent arithmetic encoding and arithmetic decoding, respectively. The one between AE and AD is the smallest form (bit stream) in which the data exist in this model. The context model is an autoregressive model for capturing context information, and entropy parameters are suitable for generating mean and scale parameters conditional on hyperprior information. The factorized entropy model is a model that can capture more edge information [13], [14], [40].

Here, after each downsampling and upsampling, LeakyReLU, convolution with stride 1, and GDN will be sequentially added. In this way, the data are normalized and gradient explosions are prevented. After each GDN, a buffer module is added, which is mainly composed of two convolutions with stride 1 and two LeakyReLU. In order to speed up the training of the network, the buffer module adds some residual connections. The buffer module is mainly designed to further stabilize the data after downsampling.

First, the remote sensing image X is input to the main encoder M_a, and after passing through multiple residual blocks, SSIF-AM, MHG-AM, and LGCM, the latent representation features containing multiscale local information and global information is obtained, i.e.,

$$\begin{aligned} Y &= M_a(X;\Upsilon_a) \\ &= W_{d5} * (W_{\mathrm{SSIF-AM3}}(W_{d4} * (W_{\mathrm{LGCM}} \\ &\quad (W_{d3} * (W_{\mathrm{SSIF-AM2}}(W_{d2} * (W_{\mathrm{SSIF-AM1}}(W_{d1} * X)))))))). \end{aligned} \tag{2}$$

Here, Υ_a represents the parameters of each part, $*$ represents the relevant convolution operations, $W_{\mathrm{SSIF-AM}_i} (i = 1, 2, 3)$ represents the SSIF-AM, W_{LGCM} represents the weight parameters of LGCM, and $W_{di} (i = 1, 2, 3, 4 \cdots)$ represents the various residual blocks, convolutions, and GDN operations in Fig. 1. Since the entropy model corresponds to the prior information of the compact latent representation, the edge information can be regarded as the prior information of the entropy model parameters. Finally, the obtained Y is sent to hyper encoder N_a to capture edge information, i.e.,

$$\begin{aligned} Z &= N_a(Y;\Omega_a) \\ &= W_{d5} * W_{\mathrm{LeakyRELU4}}(W_{d4} * W_{\mathrm{LeakyRELU3}} \\ &\quad (W_{d3} * W_{\mathrm{LeakyRELU2}}(W_{d2} * W_{\mathrm{LeakyRELU1}}(W_{d1} * y)))). \end{aligned} \tag{3}$$

Here, Ω_a represents the parameters of the hyper encoder, $W_{di}(i = 1, 2, 3, 4, 5)$represents the convolutional layer, $W_{\mathrm{LeakyRELU}i}(i = 1, 2, 3, 4)$ represents the Leaky RELU layer. Then, Z is quantized, and the quantized Z is further compressed by an arithmetic encoder to obtain the bitstream. Then, a factorized entropy model and an arithmetic decoder are used to obtain $\hat{Z}$, which is then sent to a hyper decoder to generate entropy parameters, i.e.,

$$\begin{aligned} \theta_{N_h} &= N_h(\widehat{Z};\Omega_h) \\ &= W_{d5} * W_{\mathrm{LeakyRELU4}}(W_{d4} * W_{\mathrm{LeakyRELU3}} \\ &\quad (W_{d3} * (W_{\mathrm{LeakyRELU2}}(W_{d2} * (W_{\mathrm{LeakyRELU1}}(W_{d1} * \widehat{Z})))))). \end{aligned} \tag{4}$$

Here, Ω_h represents the hyper decoder and $\hat{Z}$ parameters, $W_{di}(i = 1, 2, 3, 4, 5)$ represents the convolutional layer, and $W_{\mathrm{LeakyRELU}i}(i = 1, 2, 3, 4)$ represents the Leaky RELU layer.

Before image reconstruction, this article studies the problem of minimizing rate distortion loss. In this article, each latent representation variable $\widehat{y_i}$ is modeled into a model that conforms to the Gaussian distribution, which ensures that the distributions of the main encoder and the main decoder can be well matched during training under the premise of adding additive uniform noise. In addition, mean parameters and scale parameters are added to the model to improve the model's ability to reconstruct images. The predicted Gaussian model parameters are as follows:

$$p_{\widehat{y}}(\widehat{y}|\widehat{z}, \theta_{N_h}, \theta_{\rm cm}, \theta_{\rm eq}) = \prod \left(N(\mu_i, {\sigma_i}^2) * U\left(-\frac{1}{2}, \frac{1}{2}\right)\right)(\widehat{y}_i). \tag{5}$$

Here, θ_{N_h} is the parameter of the hyper decoder, $\theta_{\rm cm}$ is the parameter of the text context model, $\theta_{\rm eq}$ is the parameter of the entropy parameter network, $N(\mu_i, {\sigma_i}^2)$ is the Gaussian distribution, μ_i is the mean, ${\sigma_i}^2$ is the variance, and $U(-\frac{1}{2}, \frac{1}{2})$ is the uniformly distributed noise.

Finally, this article reconstructs the remote sensing image $\widehat{X}$ by using the main decoder M_h, i.e.,

$$\begin{aligned} \widehat{X} &= M_h(\widehat{Y}; \Upsilon_h) \\ &= W_{d5} * W_{\rm SSIM-AM}(W_{d4} * W_{\rm SSIF-AM2} \\ &\quad (W_{d3} * W_{\rm LGCM}(W_{d2} * W_{\rm SSIF-AM1}(W_{d1} * \widehat{Y})))). \end{aligned} \tag{6}$$

Here, Υ_h is the parameter of the main decoder, $\widehat{Y}$ is Y after decoding by the arithmetic decoder, $*$ is the relevant convolution operation, $W_{{\rm SSIF-AM}_i} (i = 1, 2, 3)$ is the weight parameter of SSIF-AM, $W_{\rm LGCM}$ is the weight parameter of LGCM, and $W_{di} (i = 1, 2, 3, 4 \cdots)$ is the various residual block, convolution, and GDN operations in Fig. 1.

It should be emphasized that the SSIF-AM, MHG-AM, and LGCM proposed in this article are embedded between the main encoder and the main decoder. On the one hand, SSIF-AM can adaptively extract microlatent representations and large-scale features at different scales under the framework of VAE. On the other hand, MHG-AM is embedded in the middle of the main encoder and main decoder, allowing the network to capture the appropriate global features. MHG-AM is not embedded in the front of the network, which may cause the model to obtain too much global information, so that the reconstructed image would lack some detailed features. At the same time, this article did not embed MHG-AM at the tail of the main encoder because after the feature map at the tail of the main encoder was downsampled four times, the feature map became very small in the spatial dimension, which results in only a small amount of global information available for learning on the feature map, so using SSIF-AM for global feature extraction is not very meaningful. Furthermore, LGCM can effectively integrate local information and global visual features. Finally, in order to keep the hyper encoder and hyper decoder simple and capture sparse edge information, the model uses Leaky RELU, which is more efficient than RELU.

B. SSIF-AM

Remote sensing images have large spatial data redundancy, and SSIF-AM helps the model to identify feature information at different scales, so that the reconstructed images retain more detailed features. In general, the model uses a large convolutional kernel to capture large-scale feature information. However, as the size of the convolutional kernel increases, the computational complexity increases dramatically. In contrast, bar convolution has the advantage of acquiring information over long distances with fewer parameters. SSIF-AM uses bar convolution instead of ordinary convolution to increase the receptive field at the same complexity. The process of SSIF-AM can be represented as

$$\begin{aligned} I_{\rm SSIF-AM} = I \otimes W_{\rm conv1\times1}(&W_{\rm conv5\times1} * W_{\rm conv1\times5} * W_{\rm conv5\times5}(I) \\ &+ W_{\rm conv7\times1} * W_{\rm conv1\times7} * W_{\rm conv5\times5}(I) \\ &+ W_{\rm conv9\times1} * W_{\rm conv1\times9} * W_{\rm conv5\times5}(I) \\ &+ W_{\rm AvgPool2D} * W_{\rm conv5\times5}(I) \\ &+ W_{\rm MaxPool2D} * W_{\rm conv5\times5}(I) \\ &+ W_{\rm conv5\times5}(I)). \end{aligned} \tag{7}$$

Here, I represents the input of SSIF-AM, W represents the weight parameters of each convolutional layer and pooling layer, $I_{\rm SSIF-AM}$ represents the output of SSIF-AM, and $\otimes$ represents Hadamard Product.

The block diagram of SSIF-AM is shown in Fig. 2. First, the 2-D convolution of size 5 × 5 is used to extract the local spatial information of the image. Then, three bar convolutions of different scales are utilized to process the feature map to extract local information at different scales. In addition, AvgPooling and MaxPooling are adopted to extract the average and maximum information of the feature map, respectively. Then, the feature map obtained by bar convolution and the latent representation obtained by two pooling are fused by adding operations to complete the efficient extraction of local multiscale spatial information. In addition, remote sensing images also contain certain spectral information. Therefore, after extracting local spatial information, a layer of point convolution is added to fuse spectral information. Finally, in order to speed up the training of the model and prevent overfitting, some residuals are introduced into the SSIF-AM module is embedded.

C. MHG-AM

In recent years, ViT technologies have been widely used in the field of computer vision. Due to the ability of ViT to capture long-distance dependencies between long-distance features, the model with ViT has demonstrated excellent performance. This ability to capture global features mainly comes from its core component, i.e., MHSA. This ability to capture global low-frequency information makes transformer often understood as a low-pass filter; CNN, on the other hand, is generally utilized to capture local information, so it is often considered to be a kind of high-pass filter. Considering the different but complementary nature of transformer and CNN, this article proposes an MHG-AM that embeds the core component of transformer (MHSA) and CNN together. This integration method is beneficial for the full extraction of multilevel information.

The structure diagram of MHG-AM is shown in Fig. 3. The global semantic features of MHG-AM are obtained through modeling, which includes MHSA, a Softmax layer, a Hadamard product, a dropout layer, two matrix multiplication, two point additions, two convolutional layers, and several linear layers. MHSA generally maps data into different projection spaces, and then uses tensor multiplication of information from different

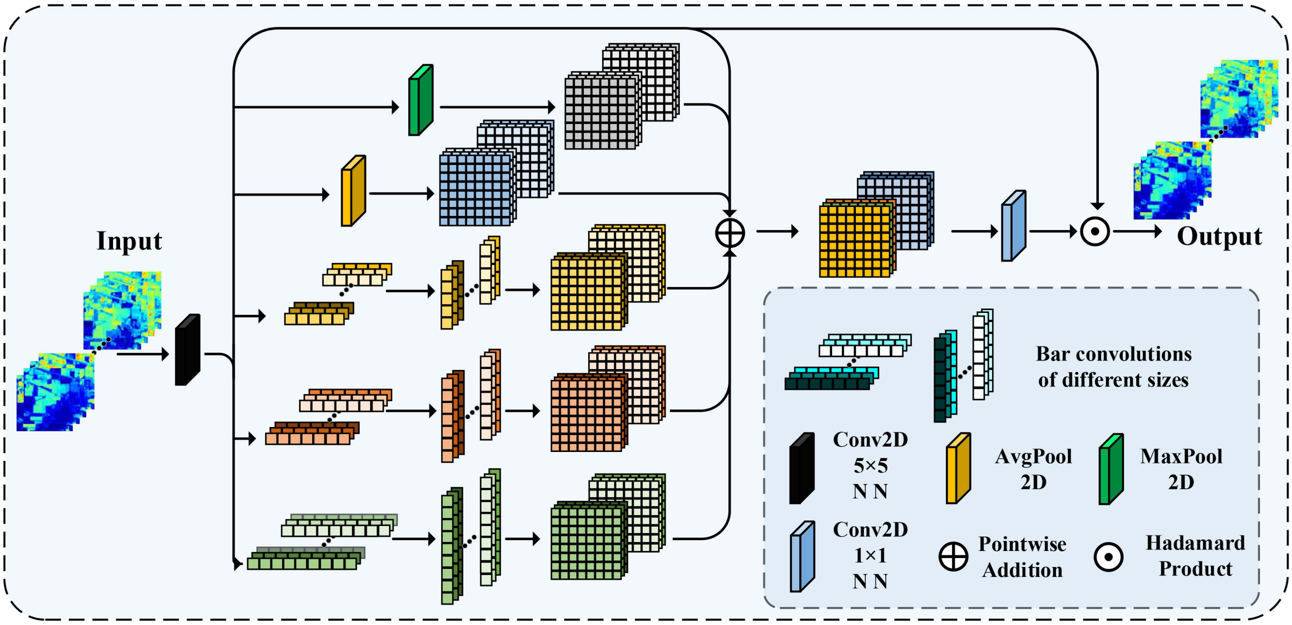


Fig. 2. Schematic diagram of SSIF-AM.

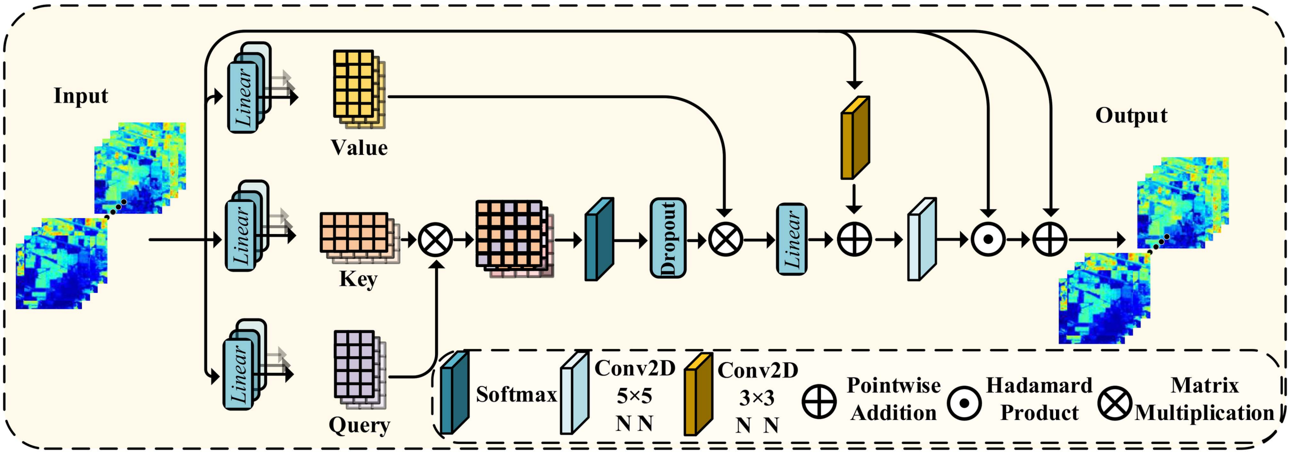


Fig. 3. Schematic diagram of MHG-AM structure.

projection spaces to obtain context features with global relationships. The process of MHG-AM can be represented as

$$\text{Output} = \text{Input} \odot (W_{\text{Conv}5\times5}(W_{\text{Linear}}(V \otimes (W_{\text{Dropout}} \\ (W_{\text{Softmax}}(Q \otimes K)))) + W_{\text{Conv}3\times3}(\text{Input}))) + \text{Input}. \quad (8)$$

Here, Input represents the input, $W_{\text{LeakyRELU}}$ represents the LeakyRELU layer, W_{Conv} represents the convolutional layer, W_{Dropout} represents the random deactivation layer, $\otimes$ represents matrix multiplication, $+$ represents addition, Q represents Query, K represents Key, V represents Value, and Output represents output. The process of the proposed MHG-AM is shown in Algorithm 1.

D. LGCM

The high-frequency information of the image usually includes edges, textures, small objects, and some changes in the ground objects. In the process of image compression, more detailed features should be retained, which is conducive to the enhancement of image edge information and the recovery of enhanced texture. In addition, low-frequency information also plays a crucial role in the reconstruction process of the image, including balancing the overall brightness and color, reducing discontinuities between pixels, maintaining the uniformity of the image, and helping to restore the overall structure. Therefore, when designing the network, corresponding subnetworks are designed for the extraction of low-frequency and high-frequency information of images, including SSIF-AM for capturing high-frequency local information and MHG-AM for capturing global information.

Algorithm 1: The Feature Extraction Process of the Proposed MHG-AM.

```
Input: Remote sensing data $X \in \mathbb{R}^{b\times c\times h\times w}$
for $i = 1$ to $T$ do
    Perform Flatten, Reshape, and Linear, the result denoted as $x \in \mathbb{R}^{b\times n\times 3c}$.
    Perform Split, $x$ denoted as $Q \in \mathbb{R}^{b\times n\times c}$, $K \in \mathbb{R}^{b\times n\times c}$ and $V \in \mathbb{R}^{b\times n\times c}$.
    Perform $Conv_{3\times 3}$, $x \in \mathbb{R}^{b\times c\times h\times w}$ denoted as $x_1 \in \mathbb{R}^{b\times c\times h\times w}$.
    Perform Reshape and Transpose, $Q$, $K$, $V$ denoted as $Q \in \mathbb{R}^{b\times head\times n\times headd}$, $K \in \mathbb{R}^{b\times head\times n\times headd}$,
    $V \in \mathbb{R}^{b\times head\times n\times headd}$, respectively.
    Perform matrix multiplication of the transpose of $Q$ and $K$, softmax, dropout, the result denoted as
    $attn \in \mathbb{R}^{b\times head\times n\times n}$.
    Perform $attn$ and $V$ matrix multiplications, the result denoted as $attn_1 \in \mathbb{R}^{b\times head\times n\times headd}$.
    Perform Transpose, Flatten, Linear, the result denoted as $attn_2 \in \mathbb{R}^{b\times n\times c}$.
    Perform Transpose, Reshape, the result denoted as $attn_3 \in \mathbb{R}^{b\times c\times h\times w}$.
    Add $attn_3$ and $x_1$ to get $attn_4 \in \mathbb{R}^{b\times c\times h\times w}$.
    Perform $Conv_{5\times 5}$ on $attn_4$ to get $attn_5$, then perform the Hadamard product of $attn_5$ and $x_1$ to get
    $attn_6 \in \mathbb{R}^{b\times c\times h\times w}$.
    Perform a residual operation and add $x$ and $attn_6$ to get output $attn_7 \in \mathbb{R}^{b\times c\times h\times w}$.
end for
Output: Feature map $\tilde{X} \in \mathbb{R}^{b\times c\times h\times w}$ after feature extraction.
```

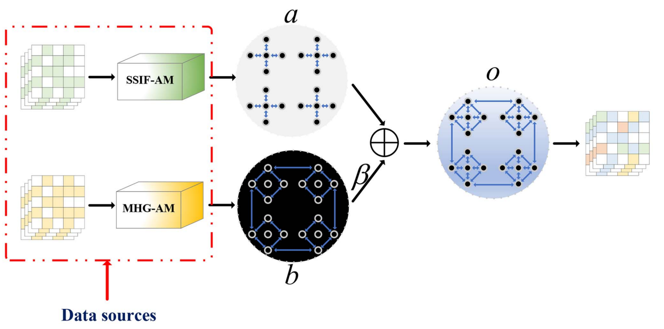


Fig. 4. Proposed LGCM ($\beta$is a harmony hyperparameter and $\oplus$ is pointwise addition).

These two modules are utilized to extract both high-frequency and low-frequency information to achieve high-quality image reconstruction. In the experiment, SSIF-AM showed a higher improvement in the quality of reconstructed images compared to MHG-AM in evaluation metrics, such as PSNR and MS-SSIM. Therefore, in this article, the multiscale local features obtained by SSIF-AM are taken as the main body, and the global features obtained by MHG-AM are used as auxiliary information, and the LGCM can be constructed to effectively integrate local information and global information.

The LGCM structure is shown in Fig. 4. a represents the high-frequency local feature information processed by the subnetwork SSIF-AM, which mainly extracts the relationship between pixels and surrounding pixels; b represents the low-frequency global information processed by the subnetwork MHG-AM, which mainly represents the relationship between global pixels; β is the harmony hyperparameter; and o represents the high-quality feature information after the effective fusion of high-frequency global information and low-frequency global information, which obtains the harmonious relationship between a single pixel and the surrounding pixels, as well as the global pixels. In this article, SSIF-AM was used as the main body, and the local feature map of the input image was first extracted, then the global information extracted by MHG-AM was multiplied by the harmonic hyperparameter, and finally the feature maps of the two branches were added point by point to obtain the final feature map that effectively fused the local information and the global information. In this article, the harmony between the two types of information can be adjusted by adjusting the value of the hyperparameter β. The process of LGCM is

$$\mathrm{Output}_{\mathrm{LGCM}} = \mathrm{Module}_{\mathrm{SSIF-AM}} + \beta \bullet \mathrm{Module}_{\mathrm{MHG-AM}}. \quad (9)$$

Here, $\mathrm{Module}_{\mathrm{SSIF-AM}}$ represents the output of the SSIF-AM module, β represents the harmonic hyperparameter, $\mathrm{Module}_{\mathrm{MHG-AM}}$ represents the output of MHG-AM, and $\mathrm{Output}_{\mathrm{LGCM}}$ represents the output of LGCM.

III. Experimental Results and Analysis

Sufficient experiments were carried out on the dataset San Francisco [38] and the dataset NWPU-RESISC45 [39], and the two datasets contain rich ground feature information, which could effectively evaluate the effectiveness of the proposed MGSSNet method. In this article, MGSSNet is compared with commonly used image compression methods JPEG2000 [41], WebP [42], BPG [43], and several latest methods based on deep learning, including Minnen et al. [14], Minnen et al. [14] (mean), Balle et al. [40] (hyperprior), Balle et al. [40] (factorized-ReLU), and Cheng et al. [11]. Experimental results show that the proposed MGSSNet method can provide excellent compression performance in PSNR, MS-SSIM, and other commonly used evaluation indicators. This section includes experimental setup, experimental results on PSNR and MS-SSIM, visualization experiments of reconstructed images, ablation experiments, and classification performance of reconstructed images.

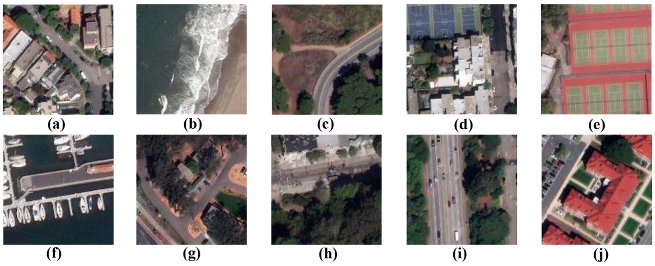


Fig. 5. Some images from San Francisco dataset. (a) Residential buildings. (b) Coastline. (c) Highway. (d) Basketball court. (e) Badminton court. (f) Port. (g) Junction. (h) Forest. (i) Vehicle. (j) School.

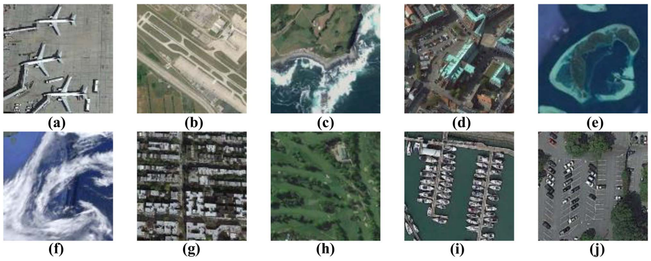


Fig. 6. Some images from NWPU-RESISC45 dataset. (a) Airports. (b) Roads. (c) Coastlines. (d) Factories. (e) Islands. (f) Airflow. (g) Cities. (h) Forests. (i) Ports. (j) Parking lots.

A. Introduction to Remote Sensing Image Datasets

1) San Francisco Dataset: San Francisco is a dataset of remotely sensed images from [38]. San Francisco is a remote sensing image with a resolution of 17 408 × 17 408. It contains information on various categories of features, such as cities, highways, rivers, ports, oceans, etc. The bit depth of the image is 24, so it contains a wealth of feature information. In this article, San Francisco was cropped to 256 × 256 size, and finally obtained 3000 valid images (the parts that contain black bars or only a single pixel are removed). The dataset is then divided into a training set, a validation set, and a test set at an 8:1:1 ratio. Fig. 5 shows some samples of this dataset.

2) NWPU-RESISC45 Dataset: NWPU-RESISC45 is a widely used remote sensing image dataset for remote sensing image classification tasks. The dataset was provided by NWPU in China. The dataset contains a total of 45 different remote sensing image scene categories, and each category contains 700 images. Each image has a resolution of 256 × 256 pixels in RGB color format. The dataset contains a variety of geographical environments and scenes, covering different categories of images, such as cities, farmlands, rivers, forests, grasslands, and airports. A total of 140 images from each category were selected to form a dataset of 6300 remotely sensed images. After that, it is divided into training set, validation set, and test set according to the ratio of 8:1:1. Fig. 6 shows some samples of this dataset.

B. Evaluation Indicators

In terms of image quality evaluation, this article adopts two common indicators, the PSNR and the MS-SSIM, to comprehensively evaluate the quality of the reconstructed images.

1) PSNR: It compares the reconstructed image to the original image from the point of view of the MSE. The higher the PSNR value, the higher the fidelity of the reconstructed image. The PSNR can be expressed as

$$\mathrm{PSNR}(X,\widehat{X}) = \frac{1}{C}\sum_{i=1}^{C} 10\log_{10}\left(\frac{\max^2(X^i)}{\mathrm{MSE}_i}\right). \quad (10)$$

Here, $\mathrm{MSE}(X,\widehat{X}) = (1/H \times W \times C)\|X-\widehat{X}\|_F^2$, $\max^2(X^{(i)})$ represents the square of the largest pixel in the ith band, and C represents the number of bands.

2) MS-SSIM: It is a multiscale structural similarity metric used to measure the differences between the original and reconstructed images, including image details merged at different resolutions [44]. Its value ranges from 0 to 1, with higher values indicating higher similarity, i.e., higher quality of the reconstructed image. In order to better compare the difference between MS-SSIM values, this article converts it into a decibel value, which can be expressed as

$$\mathrm{MS-SSIM} = -10\log_{10}(1 - D_{\mathrm{MS-SSIM}}) \quad (11)$$

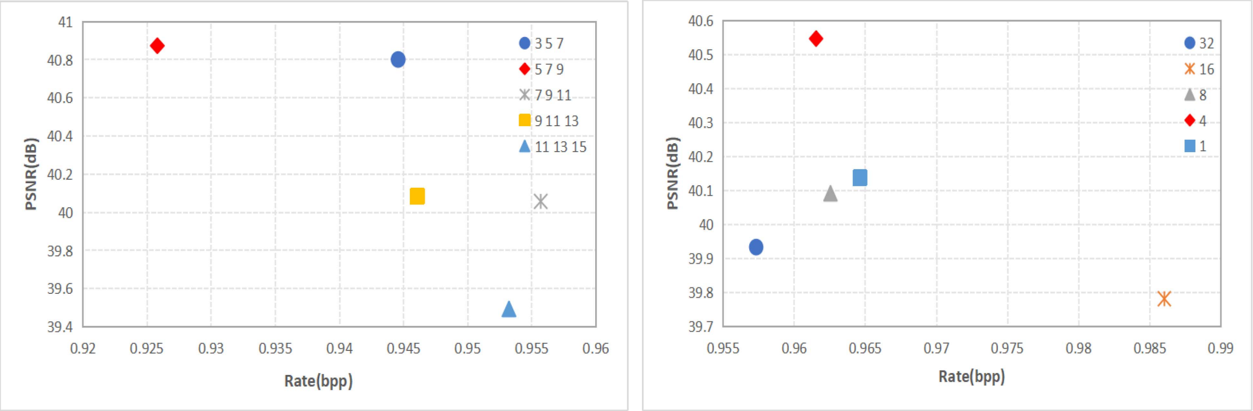
$$D_{\mathrm{MS-SSIM}} = 1 - \prod_{m=1}^{M}\left(\frac{2\mu_X\mu_{\widehat{X}} + C_1}{{\mu_X}^2 + {\mu_{\widehat{X}}}^2 + C_1}\right)^{\beta_m} \left(\frac{2\sigma_{X\widehat{X}} + C_2}{{\sigma_X}^2 + {\sigma_{\widehat{X}}}^2 + C_2}\right)^{\gamma_m}. \quad (12)$$

Here, $D_{\mathrm{MS-SSIM}}$ 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 represent the relative importance between the two terms, and C_1 and C_2 are constant terms to prevent the divisor from being 0.

Under the same bit rate conditions, the larger the values of PSNR and MS-SSIM, the better quality of the reconstructed image and the higher the similarity between the reconstructed image and the original image.

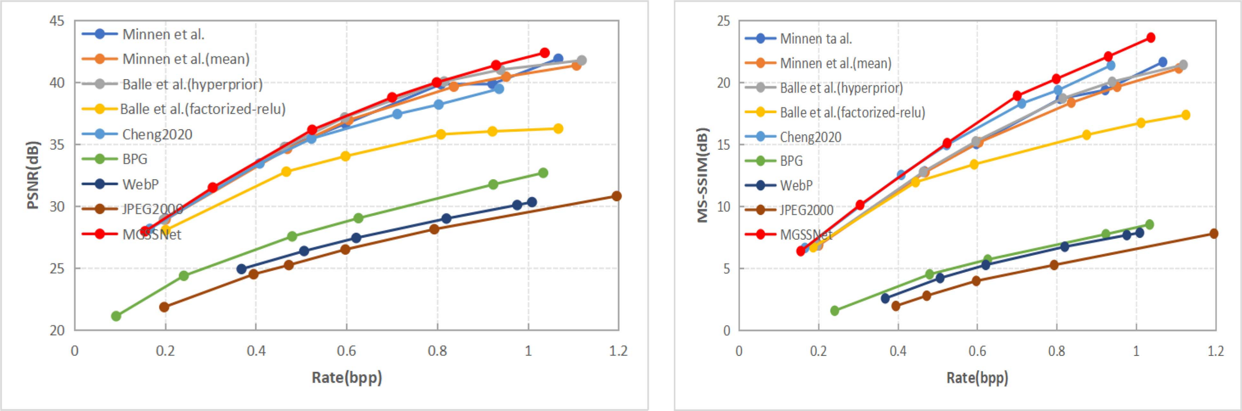
C. Experimental Environment and Parameter Settings

In this article, the PyTorch framework is used to implement all the compression methods based on the deep learning. All models were trained on an NVIDIA GeForce RTX 3090 using the Adam optimizer. In addition, all codecs are made on the same CPU (i9-9900K CPU@3.60 GHz). Two optimizers are used in this network, one in the main network and the other in the hyper codec. For the main optimizer, this article initializes the learning rate to 10^{-4}. In addition, the learning rate of the optimizer in the hyper encoding and decoding is set to 10^{-3}. The batch size is set to 8 during training. Each model was trained 300 times until convergence. For the sake of fairness in the experimental comparison, all experiments in this article were performed under the same experimental conditions. In the experiment, three traditional image compression methods, including JPEG2000, WebP, and BPG, and five latest image



(a) (b)

Fig. 7. Effects of different parameters on SSIF-AM and MHG-AM. (a) Influence of different sizes of convolution kernels adopted in SSIF-AM on PSNR. (b) Effect of the number of heads on the PSNR in MHG-AM.



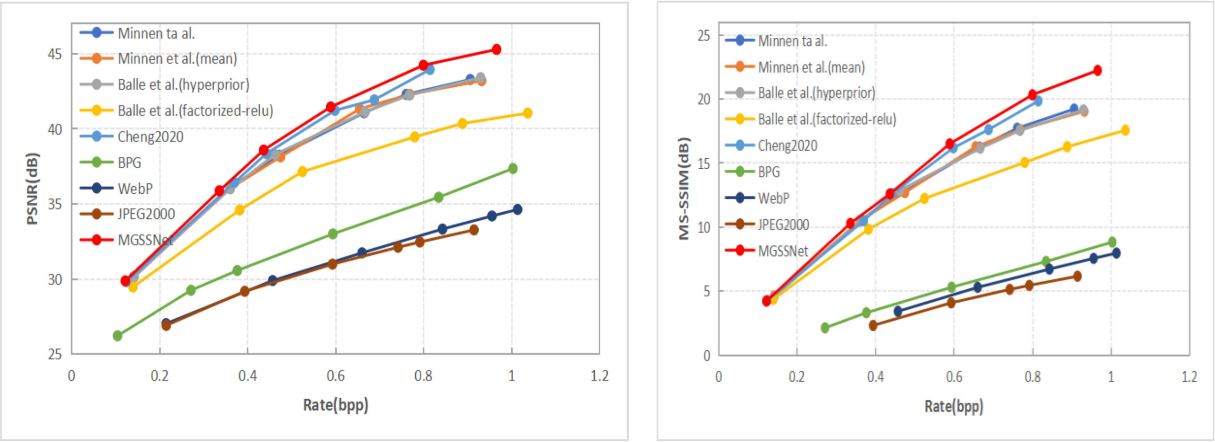
(a) (b)

Fig. 8. Compression performance of different methods on San Francisco. (a) Comparison of experimental results on PSNR. (b) Comparison of experimental results on MS-SSIM.

compression methods based on deep learning, including Minnen et al. [14], Minnen et al. [14] (mean), Balle et al. [40] (hyperprior), Balle et al. [40] (factorized-relu), Cheng et al. [11], etc., were selected for comparison. The ffmpeg version used by JPEG2000 is 6.0 and the version used by BPG is 0.9.8. The penalty coefficient λ used in this article is [0.660, 0.508, 0.211, 0.072, 0.033, 0.013, 0.007]. The sizes of the three pairs of strip convolution kernels in SSIF-AM are $k = 5$, $k = 7$, and $k = 9$, respectively. The number of self-attention heads in MHG-AM is 4. The harmony hyperparameter β in LGCM is set to 0.1. Fig. 7 shows the results of experiments conducted on the dataset San Francisco. Fig. 7(a) represents the effect of three convolutional kernels of different sizes on PSNR in SSIF-AM, and Fig. 7(b) shows the effect of the number of heads on PSNR in MHG-AM. The experimental results show that the parameters selected in this article are optimal.

D. Experimental Results on PSNR and MS-SSIM

In this article, the rate distortion performance of all methods is evaluated by PSNR and MS-SSIM. Fig. 8 shows the rate distortion performance curves of PSNR and MS-SSIM obtained experimentally on the dataset San Francisco. A total of eight state-of-the-art image compression methods are selected for comparison, including three traditional compression methods and five image compression methods based on deep learning. Among them, Minnen et al. [14], Minnen et al. [40] (mean), Balle et al. [40] (hyperprior), Balle et al. [40] (factorized-relu), and Cheng et al. [11] are all methods based on VAEs. In experiments, whether at high bit rates or low bit rates, the PSNR and MS-SSIM performance based on deep-learning methods are better than those based on traditional image coding and decoding methods. In addition, the MGSSNet method proposed in this article achieves the highest PSNR and MS-SSIM performance among all image compression methods based on deep learning. As can be seen in Fig. 8, the model based on hyperprior has an advantage over the method that is not based on hyperprior information, and this advantage becomes more pronounced in the case of high bit rates. However, the proposed model achieves better rate distortion performance than the model based on hyperprior, which is mainly due to the accurate local information and relatively complete global information captured by MGSSNet,



(a) (b)

Fig. 9. Compression performance of different methods on NWPU-RESISC45. (a) Comparison of experimental results on PSNR. (b) Comparison of experimental results on MS-SSIM.

as well as the effective fusion of local information and global information.

For the traditional image compression methods, it can be found that the rate distortion performance of BPG is significantly better than that of WebP and JPEG2000. This is due to BPG's support for multichannel encoding, which allows for independent processing of different color channels, resulting in increased control over color and detail, helping to achieve higher quality reconstructed images.

In order to fully verify the compression performance of the proposed MGSSNet method, some same experiments are carried out on the dataset NWPU-RESISC45. Fig. 9 shows the PSNR and MS-SSIM results obtained by different methods on the dataset NWPU-RESISC45. As can be seen from Fig. 9, the proposed MGSSNet achieves the best compression performance. Especially in the case of high bit rate, the advantages of the proposed MGSSNet are more obvious compared with other compression methods, which fully proves the effectiveness of the proposed method.

E. Visualization Comparison of Reconstructed Images

In order to verify the visual quality of the reconstructed images obtained by the proposed MGSSNet, this article compares the visual quality of the reconstructed images obtained by different methods. We selected an image from the dataset San Francisco that included trees and zebra crossings, and used different compression methods to reconstruct the image at around 0.3 bpp. We zoomed in on the trees in the upper left corner and the zebra crossing in the lower right corner to compare the quality of the reconstructed images of different methods. Fig. 10 gives the reconstructed image obtained by the proposed MGSSNet and eight comparison methods. For the traditional image compression method, it can be found that although the compression performance of the JPEG2000 is lower than that of BPG and WebP in PSNR and MS-SSIM, the reconstructed trees can retain more texture information, which is due to the JPEG2000 using wavelet transform and supporting multiresolution encoding. However, the trees reconstructed using BPG and WebP have become very blurry and have lost all the details. On the other hand, for the reconstructed images obtained by BPG and WebP, the edges of the zebra crossing are slightly blurred, but they are much clearer than the zebra crossing reconstructed by JPEG2000. For image compression methods based on deep learning, they all achieve better visual effects than those of traditional encoding methods and retain more texture features. As can be seen from Fig. 10, some artifacts will appear in the reconstructed trees obtained by Minnen et al. [14] (mean), Balle et al. [40], Balle et al. [40] (factorized-relu), Cheng et al. [11], and other methods. Finally, comparing Minnen et al., which has the best visual effect in the comparison method, with the proposed MGSSNet, it can be seen that the reconstructed trees obtained by MGSSNet are significantly better than those of Minnen et al. in the left branch details and the texture features in the middle of the tree. This fully illustrates the effectiveness of the proposed MGSSNet in capturing multiscale detailed features and global information.

In order to verify the robustness of the proposed model, a remote sensing image containing a church was selected from the NWPU-RESISC45 dataset to carry out the visualization experiment of the reconstructed image. At around 0.25 bpp, different compression methods are used for image compression and reconstruction. In the experimental results, two local areas were selected for magnification, including the roof in the upper left and the clock tower in the upper right, for visual comparison of the reconstructed images. Fig. 11 gives the reconstructed image obtained by the proposed MGSSNet and eight comparison methods. Because the MGSSNet network can fully extract the multiscale detail features and global information of the image, more texture features are retained in the reconstructed image, and the overall quality of the reconstructed image is also improved, with almost no artifacts and noise. Therefore, the proposed MGSSNet achieves the best visual quality of the reconstructed image.

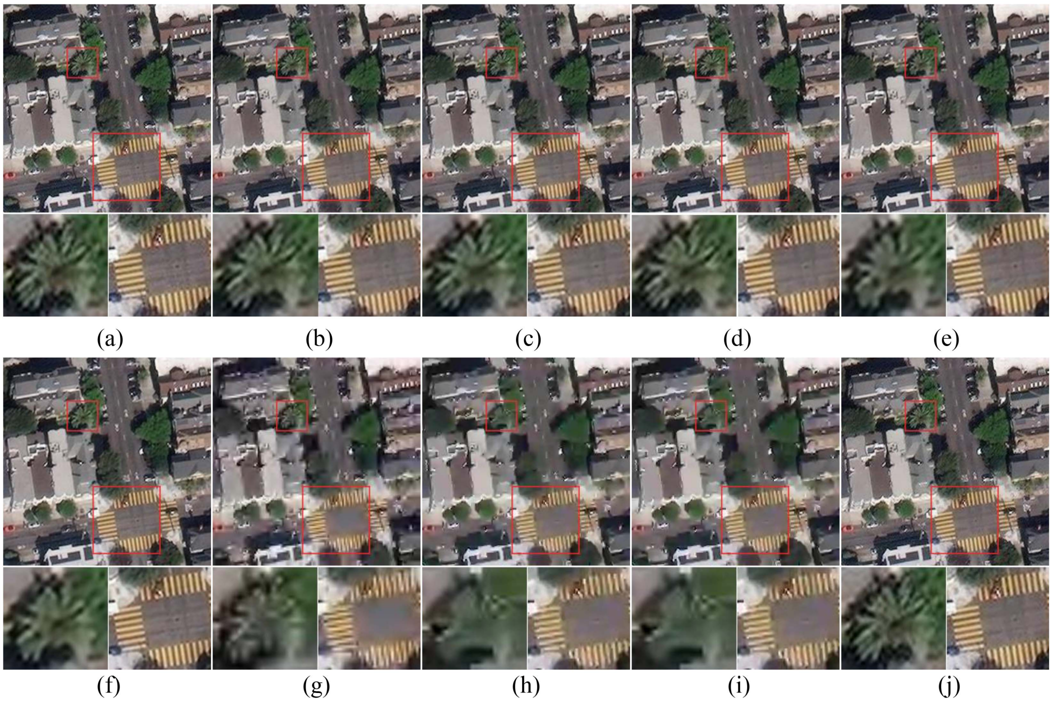


Fig. 10. Visualization comparison of reconstructed images obtained by different methods on the San Francisco dataset. (a) Original. (b) Minnen et al. (bpp: 0.2976; PSNR: 31.02; MS-SSIM: 9.58). (c) Minnen et al. (mean) (bpp: 0.2971; PSNR: 30.91; MS-SSIM: 9.54). (d) Balle et al. (hyperprior) (bpp: 0.2944; PSNR: 30.91; MS-SSIM: 9.62). (e) Balle et al. (factorized-ReLU) (bpp: 0.2938; PSNR: 29.348; MS-SSIM: 8.96). (f) Cheng et al. (bpp: 0.3054; PSNR: 29.58; MS-SSIM: 9.50). (g) JPEG2000 (bpp: 0.3157; PSNR: 23.81; MS-SSIM: 1.34). (h) Webp (bpp: 0.3681; PSNR: 24.98; MS-SSIM: 2.17). (i) BPG (bpp: 0.3192; PSNR: 25.51; MS-SSIM: 2.27). (j) MGSSNet (bpp: 0.3043; PSNR: 31.50; MS-SSIM: 9.956).

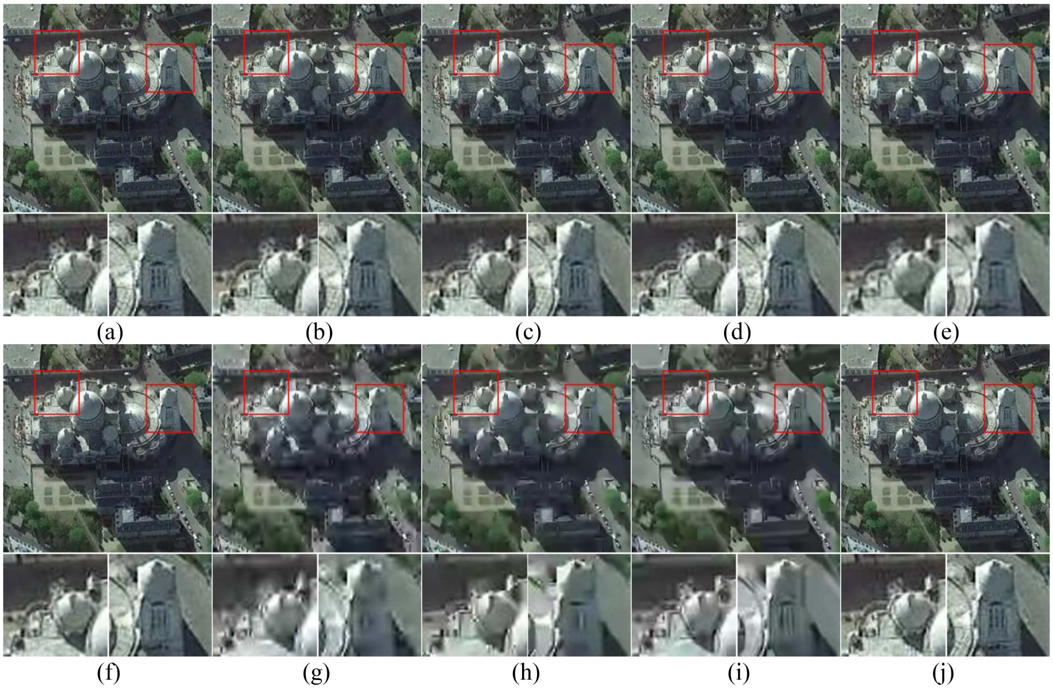
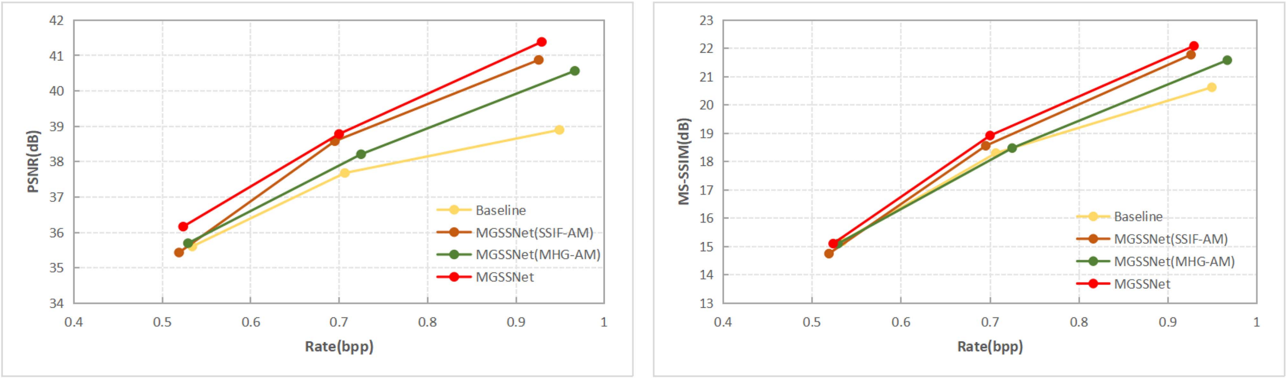
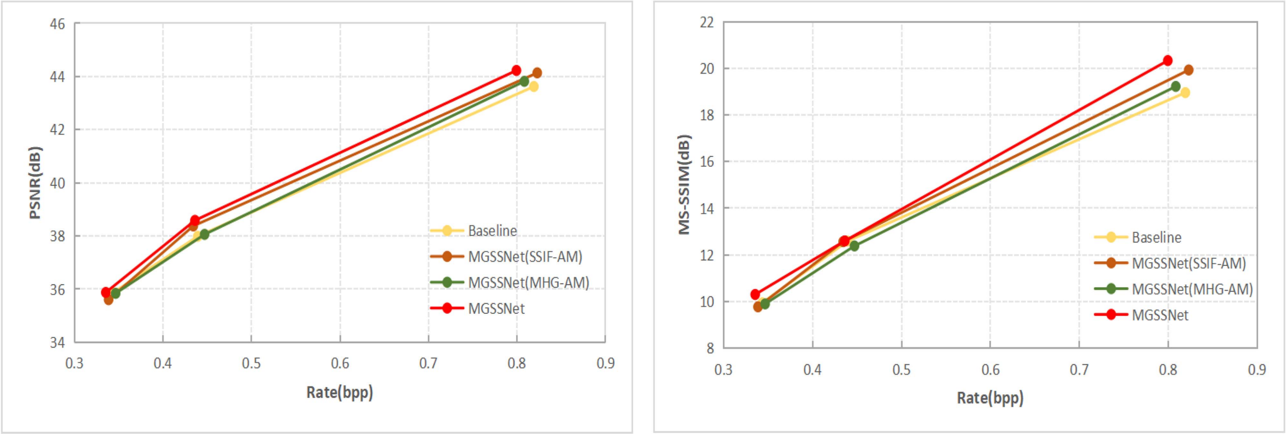


Fig. 11. Visualization comparison of reconstructed images obtained by different methods on the dataset NWPU-RESISC45. (a) Original. (b) Minnen et al. (bpp: 0.2499; PSNR: 32.96; MS-SSIM: 7.69). (c) Minnen et al. (mean) (bpp: 0.2469; PSNR: 32.90; MS-SSIM: 7.64). (d) Balle et al. (hyperprior) (bpp: 0.2473; PSNR: 33.00; MS-SSIM: 7.61). (e) Balle et al. (factorized-ReLU) (bpp: 0.2404; PSNR: 31.47; MS-SSIM: 7.47). (f) Cheng et al. (bpp: 0.2600; PSNR: 32.67; MS-SSIM: 8.38). (g) JPEG2000 (bpp: 0.2591; PSNR: 23.43; MS-SSIM: 1.16). (h) Webp (bpp: 0.3193; PSNR: 24.62; MS-SSIM:1.36). (i) BPG (bpp: 0.2770; PSNR: 25.09; MS-SSIM: 1.57). (j) MGSSNet (bpp: 0.2565; PSNR: 32.86; MS-SSIM: 8.41).



(a) (b)

Fig. 12. Comparison of ablation experimental results of different methods on the San Francisco dataset. (a) PSNR. (b) MS-SSIM.



(a) (b)

Fig. 13. Comparison of ablation experimental results of different methods on the NWPU-RESISC45 dataset. (a) PSNR. (b) MS-SSIM.

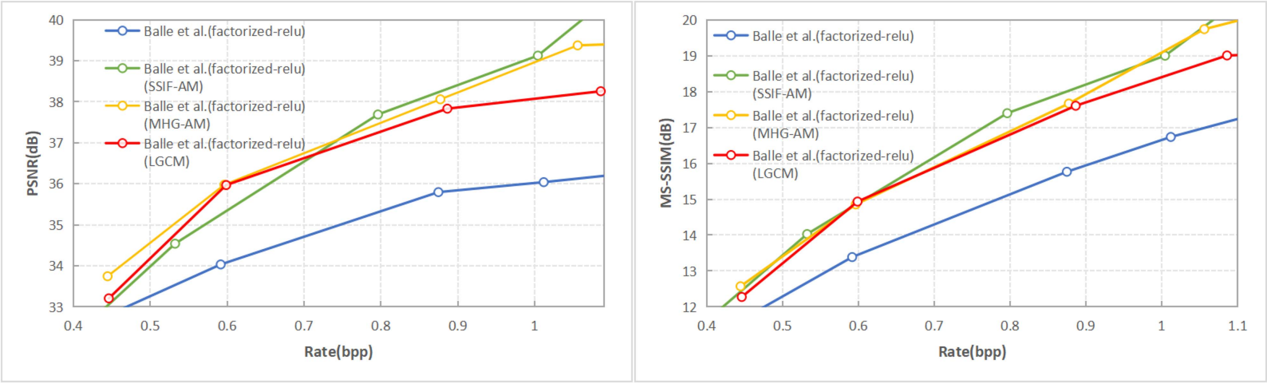
F. Ablation Experiments

In order to verify the effectiveness of each module, some ablation experiments were performed in this article. Among them: 1) baseline: baseline model; 2) MGSSNet (SSIF-AM): SSIF-AM is added after the first three downsampling of the main encoder and the first three upsampling of the main decoder, respectively; 3) MGSSNet (MHG-AM): MHG-AM is added after the third downsampling of the main encoder and the first upsampling of the main decoder; 4) MGSSNet: SSIF-AM is added after the first and second downsampling of the main encoder and the second and third upsampling of the main decoder, and LGCM is added after the third downsampling of the main encoder and the first upsampling of the main decoder. Fig. 12 shows the results of the ablation experiment on the dataset San Francisco, and Fig. 13 shows the results of the ablation experiment on the dataset NWPU-RESISC45. As can be seen in Figs. 12 and 13, baseline has the lowest performance in both PSNR and MS-SSIM. In most cases, the performance of MGSSNet (MHG-AM) is better than that of Baseline, which also verifies the importance of global information for compressing the network. In addition, the performance of MGSSNet (SSIF-AM) on PSNR and MS-SSIM is significantly higher than that of baseline, which indicates that multiscale local information plays a significant role in improving the overall quality of the reconstructed image and improving the structural similarity. Further comparison between MGSSNet (SSIF-AM) and MGSSNet (MHG-AM) shows that multiscale local information is more important than global information in this network, which can preserve the edge and texture information of the image to a greater extent. At the same bit rate, the proposed MGSSNet has the highest PSNR and MS-SSIM, and the performance advantage of the proposed method is further increased at the high bit rate. This shows that MGSSNet not only effectively extracts multiscale local information and global information, but also effectively fuses the two.

IV. Discussion

A. Generalization Analysis of Modules

To verify the generalization ability of the proposed SSIF-AM, MHG-AM, and LGCM modules, some generalization experiments were conducted. In this article, a publicly image compression method [Balle et al. (factorized-ReLU)] was selected as the baseline network. Subsequently, each module was embedded



(a)

(b)

Fig. 14. Generalization results of different modules on the San Francisco dataset. (a) PSNR. (b) MS-SSIM.

into the baseline network to verify the generalization ability of the module itself. The dataset used is San Francisco. Fig. 14 shows the rate distortion performance curve of the network after each module is added. It can be seen that in Fig. 14, the rate-distortion performance of the baseline network is the lowest. After adding the proposed modules, i.e., Balle et al. (factorized-ReLU) (SSIF-AM), Balle et al. [40] (factorized-ReLU) (MHG-AM), and Balle et al. [40] (factorized-ReLU) (LGCM), the significant improvements in PSNR and MS-SSIM have been achieved. This fully proves that both multiscale local features and long-distance global features play an effective role in improving the compression performance of remote sensing images. It also confirms the generalization of each module. It is worth mentioning that the improvement in rate-distortion performance brought by the SSIF-AM module on the baseline network Balle et al. [40] (factorized-ReLU) is remarkable. At 1.1 bpp, compared with Balle et al. (factorized-ReLU), the PSNR of Balle et al. [40] (factorized-ReLU) (MHG-AM) and Balle et al. [40] (factorized-ReLU) (LGCM) achieve improvements of 8.8% and 7.0%, respectively. Especially, Balle et al. (factorized-ReLU) (SSIF-AM) reached a remarkable 11.4% improvement. The reason is that Balle et al. (factorized-ReLU) itself cannot effectively fuse multiscale local features and long-range global features. However, SSIF-AM, MHG-AM, and LGCM just compensate for this deficiency, resulting in significant performance improvements.

B. The Impact of Reconstructed Images on Some Applications

In order to further verify the effectiveness of the proposed MGSSNet method, the reconstructed images obtained by the proposed MGSSNet method and Minnen et al. [14], Minnen et al. [14] (mean), Balle et al. [40] (hyperprior), Balle et al. [40] (factorized-ReLU), and Cheng et al. [11] are compared in remote sensing scene image classification. The quality of these reconstructed images is evaluated by classification performance. The classification method used in this article is an efficient multiscale transformer and cross-level attention learning for remote sensing scene classification [45], and the dataset used for training is NWPU-RESISC45, and the training ratio is 10%.

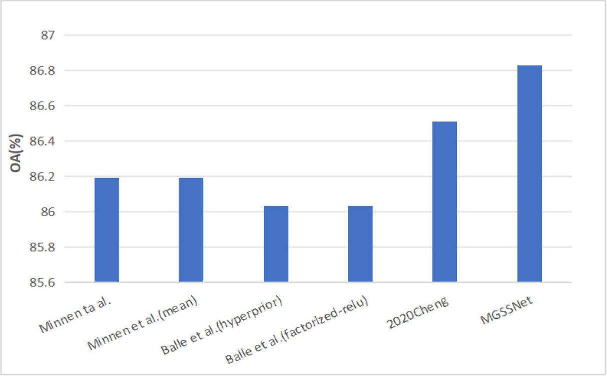
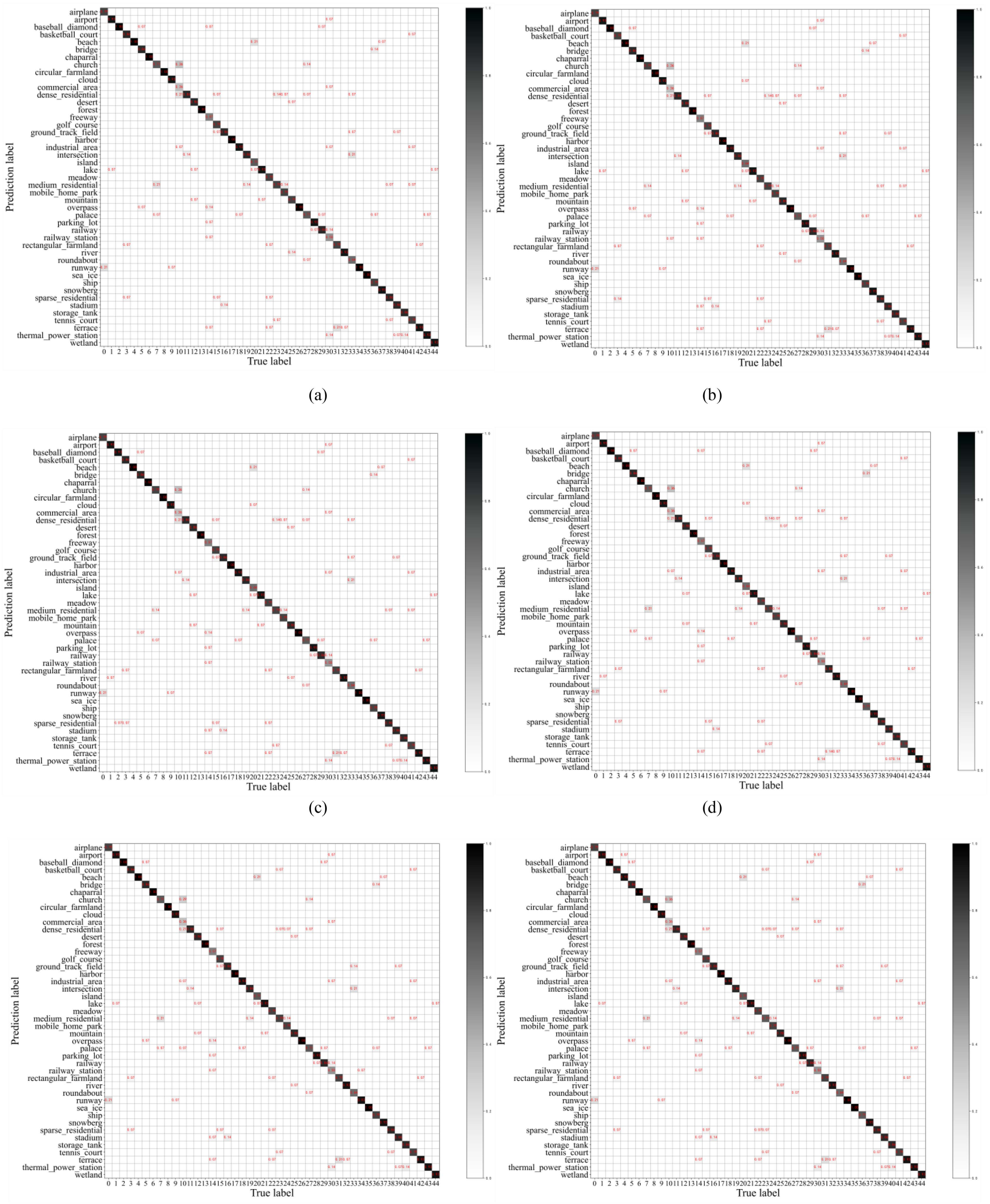


Fig. 15. OA value of the reconstructed image obtained by different compression methods in remote sensing scene classification (the dataset used is NWPU-RESISC45).

The images used for compression and the images used for remote sensing scene classification training are not crossed, and the reconstructed images are only used for the test of classification performance, not for the training of classification networks. To ensure the fairness of the experiment, the bitrate of all compression methods is set at around 0.81 bpp. It can be seen that in Fig. 15, the proposed MGSSNet achieves the highest OA, which is more than 0.73%, 0.73%, 0.91%, 0.91%, and 0.37% than that of Minnen et al. [14], Minnen et al. [14] (mean), Balle et al. [40] (hyperprior), Balle et al. [40] (factorized-ReLU), and Cheng et al. [11], respectively.

Fig. 16 shows the confusion matrix of the reconstructed images of Minnen et al. [14], Minnen et al. [14] (mean), Balle et al. [40] (hyperprior), Balle et al. [40] (factorized-ReLU), Cheng et al. [11], and the proposed MGSSNet method for remote sensing image classification. It can be seen that for the classes of "river," "baseball_diamond," and "cloud," the classification effect of the proposed method is better than that of other methods. This is due to the fact that these types of scenes contain more global visual features, and the proposed MGSSNet network can enhance the capture of global information. For remote sensing scene classification, it is important to effectively obtain local features with complex spatial information and geometric structure, and



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 16. Confusion matrix of reconstructed images obtained by different compression methods for remote sensing scene classification. (a), (b), (c), (d), (e), and (f) correspond to the confusion matrices of Minnen et al. [14], Minnen et al. [14] (mean), Balle et al. [40] (hyperprior), Balle et al. [40] (factorized-ReLU), Cheng et al. [11], and MGSSNet, respectively.

to efficiently retain global visual features. The proposed MGSSNet uses SSIF-AM to obtain multiscale local features, uses MHG-AM to obtain global visual features, and effectively fuses global visual features and multilevel local information through the designed LGCM, ultimately generating efficient discriminative features. This is why the proposed MGSSNet method can provide the best classification performance for remote sensing scene image.

V. Conclusion

In this article, we propose an MGSSNet network based on multihead global attention and spatial spectral information fusion for the compression of remote sensing images. Specifically, this article designs an SSIF-AM for capturing both microfeatures and large-scale features; in addition, this article constructs an MHG-AM for capturing global visual features. Finally, this article proposes an LGCM to coordinate the local features of SSIF-AM and the global context features of MHG-AM, so as to help MGSSNet achieve high-quality image compression and reconstruction work, and achieve very superior rate distortion performance. Specifically, at 1.1 bpp, MGSSNet achieves PSNR improvements of 2.9%, 4.2%, 3.2%, 18.9%, and 4.4% compared to that of Minnen et al. [14], Minnen et al. [14] (mean), Balle et al. [40] (hyperprior), Balle et al. [40] (factorized-ReLU), and Cheng et al. [11], respectively. The quantitative and visual analysis results on different remote sensing datasets show that the introduced SSIF-AM, MHG-AM, and LGCM can adaptively retain multiscale detail features and global context relationships, thereby improving the compression performance. It is worth noting that, the proposed SSIF-AM, MHG-AM, and LGCM can be inserted into any neural-network-based image compression method to improve their performance. This indicates that the proposed method has strong generalization and can be easily integrated into other compression networks. In future article, we will try to perform the hierarchical compression processing for images. By optimizing multilevel features, the information gap between the latent representation and the specific task is narrowed, so as to further improve the compression performance.

References

[1] Q. Yuan et al., “Deep learning in environmental remote sensing: Achievements and challenges,” *Remote Sens. Environ.*, vol. 241, 2020, Art. no. 111716.

[2] B. Huang, B. Zhao, and Y. Song, “Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery,” *Remote Sens. Environ.*, vol. 214, pp. 73–86, 2018.

[3] Q. Vanhellemont and K. Ruddick, “Advantages of high quality SWIR bands for ocean colour processing: Examples from Landsat-8,” *Remote Sens. Environ.*, vol. 161, pp. 89–106, 2015.

[4] D. Wang et al., “A review of deep learning in multiscale agricultural sensing,” *Remote Sens.*, vol. 14, no. 3, 2022, Art. no. 559.

[5] S. E. Qian, “Hyperspectral data compression using a fast vector quantization algorithm,” *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1791–1798, Aug. 2004.

[6] R. Pizzolante and B. Carpentieri, “Multiband and lossless compression of hyperspectral images,” *Algorithms*, vol. 9, no. 1, 2016, Art. no. 16.

[7] L. Thornton et al., “Unequally protected SPIHT video codec for low bit rate transmission over highly error-prone mobile channels,” *Signal Process.: Image Commun.*, vol. 17, no. 4, pp. 327–335, 2002.

[8] 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.

[9] S. H. Kim, J. H. Park, and J. H. Ko, “Target-dependent scalable image compression using a reconfigurable recurrent neural network,” *IEEE Access*, vol. 9, pp. 119418–119429, 2021.

[10] D. Liu et al., “View synthesis-based light field image compression using a generative adversarial network,” *Inf. Sci.*, vol. 545, pp. 118–131, 2021.

[11] 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.*, 2020, pp. 7939–7948.

[12] D. Liu, X. Sun, F. Wu, and Y.-Q. Zhang, “Edge-oriented uniform intra prediction,” *IEEE Trans. Image Process.*, vol. 17, no. 10, pp. 1827–1836, Oct. 2008.

[13] F. Kong et al., “Multi-scale spatial-spectral attention network for multispectral image compression based on variational autoencoder,” *Signal Process.*, vol. 198, 2022, Art. no. 108589.

[14] D. Minnen, J. Ballé, and G. D. Toderici, “Joint autoregressive and hierarchical priors for learned image compression,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 10794–10803.

[15] R. La Grassa et al., “Hyperspectral data compression using fully convolutional autoencoder,” *Remote Sens.*, vol. 14, no. 10, 2022, Art. no. 2472.

[16] J. Liu, F. Yuan, C. Xue, Z. Jia, and E. Cheng, “An efficient and robust underwater image compression scheme based on autoencoder,” *IEEE J. Ocean. Eng.*, vol. 48, no. 3, pp. 925–945, Jul. 2023.

[17] A. A. Jeny, M. B. Islam, M. S. Junayed, and D. Das, “Improving image compression with adjacent attention and refinement block,” *IEEE Access*, vol. 11, pp. 17613–17625, 2023.

[18] V. Alves de Oliveira et al., “Reduced-complexity end-to-end variational autoencoder for on board satellite image compression,” *Remote Sens.*, vol. 13, no. 3, 2021, Art. no. 447.

[19] Q. Xu et al., “Synthetic aperture radar image compression based on a variational autoencoder,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4015905.

[20] S. Xiang and Q. Liang, “Remote sensing image compression with long-range convolution and improved non-local attention model,” *Signal Process.*, vol. 209, 2023, Art. no. 109005.

[21] H. Ma, D. Liu, R. Xiong, and F. Wu, “iWave: CNN-based wavelet-like transform for image compression,” *IEEE Trans. Multimedia*, vol. 22, no. 7, pp. 1667–1679, Jul. 2020.

[22] Z. Jin, M. Z. Iqbal, W. Zou, X. Li, and E. Steinbach, “Dual-stream multipath recursive residual network for JPEG image compression artifacts reduction,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 467–479, Feb. 2021.

[23] Z. Tang, H. Wang, X. Yi, Y. Zhang, S. Kwong, and C.-C. J. Kuo, “Joint graph attention and asymmetric convolutional neural network for deep image compression,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 1, pp. 421–433, Jan. 2023.

[24] W. Wang et al., “Pvt v2: Improved baselines with pyramid vision transformer,” *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–424, 2022.

[25] X. Meng, N. Wang, F. Shao, and S. Li, “Vision transformer for pansharpening,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5409011.

[26] O. Dalmaz, M. Yurt, and T. Çukur, “ResViT: Residual vision transformers for multimodal medical image synthesis,” *IEEE Trans. Med. Imag.*, vol. 41, no. 10, pp. 2598–2614, Oct. 2022.

[27] D. Wang et al., “Advancing plain vision transformer toward remote sensing foundation model,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5607315.

[28] B. Li, J. Liang, and J. Han, “Variable-rate deep image compression with vision transformers,” *IEEE Access*, vol. 10, pp. 50323–50334, 2022.

[29] S. Xiang, Q. Liang, and L. Fang, “Discrete wavelet transform-based Gaussian mixture model for remote sensing image compression,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3000112.

[30] X. Ma, “High-resolution image compression algorithms in remote sensing imaging,” *Displays*, vol. 79, 2023, Art. no. 102462.

[31] P. Han, B. Zhao, and X. Li, “Edge-guided remote sensing image compression,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5524515.

[32] S. Xiang and Q. Liang, “Remote sensing image compression with long-range convolution and improved non-local attention model,” *Signal Process.*, vol. 209, 2023, Art. no. 109005.

[33] F. Kong et al., “End-to-end multispectral image compression framework based on adaptive multiscale feature extraction,” *J. Electron. Imag.*, vol. 30, no. 1, 2021, Art. no. 013010.

[34] F. Kong et al., "Multi-scale spatial-spectral attention network for multispectral image compression based on variational autoencoder," *Signal Process.*, vol. 198, 2022, Art. no. 108589.

[35] S. Pan et al., "Content-based hyperspectral image compression using a multi-depth weighted map with dynamic receptive field convolution," *Int. J. Interactive Multimedia Artif. Intell.*, vol. 7, pp. 85–92, 2022.

[36] S. Tao et al., "SAR image despeckling using a CNN guided by high-frequency information," *J. Electromagn. Waves Appl.*, vol. 37, no. 3, pp. 441–451, 2023.

[37] N. Park and S. Kim, "How do vision transformers work?," 2022, *arXiv:2202.06709*.

[38] 2016. [Online]. Available: https://resources.maxar.com/product-samples/analysis-ready-data-san-francisco-california

[39] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

[40] J. Ballé et al., "Variational image compression with a scale hyperprior," 2018, *arXiv:1802.01436*.

[41] JPEG2000 official software OpenJPEG, 2015. [Online]. Available: https://jpeg.org/jpeg2000/software.html

[42] M. Maldonado, "WebP J: A new web oriented image format," Universitat Oberta de Catalunya, 2010.

[43] F. Bellard, "BPG image format," 2018. [Online]. Available: http://bellard.org/bpg/

[44] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, 2003, pp. 1398–1402.

[45] X. Tang, M. Li, J. Ma, X. Zhang, F. Liu, and L. Jiao, "EMTCAL: Efficient multiscale transformer and cross-level attention learning for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5626915.

Cuiping Shi (Member, IEEE) received the M.S. degree in signal and information processing from the Yangzhou University, Yangzhou, China, in 2007, and the Ph.D. degree in information and communication engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2016.

From 2017 to 2020, she held Postdoctoral Research with the College of Information and Communications Engineering, Harbin Engineering University. Since 2024, she has been with the college of Information Engineering, Huzhou University, Huzhou, China. She is currently a Professor with the Department of Communication Engineering, Qiqihar University, Qiqihar, China. She has authored/coauthored two academic books about remote sensing image processing and more than 80 papers in journals and conference proceedings. Her research interests include remote sensing image processing, pattern recognition, and machine learning.

Dr. Shi was the recipient of the nomination award of Excellent Doctoral Dissertation of HIT, in 2016.

Kaijie Shi received the bachelor's degree in electronic information science and technology from the Heilongjiang University of Science and Technology, Harbin, China, in 2021. He is currently pursuing a Master's degree in communication and information systems at Qiqihar University, Qiqihar, China.

His research interests include remote sensing image compression and machine learning.

Fei Zhu received the bachelor's degree in internet of things engineering from Luoyang Institute of Science and Technology, Luoyang, China, in 2021. He is currently pursuing a Master's degree in electronic information at Qiqihar University, Qiqihar, China.

His research interests include hyperspectral image processing and machine learning.

Zexin Zeng received the bachelor's degree in electronic information engineering from the Guangzhou Maritime University, Guangdong, China, in 2021. He is currently pursuing a Master's degree in communication and information systems at Qiqihar University, Qiqihar, China.

His research interests include hyperspectral image processing and machine learning.

Liguo Wang (Member, IEEE) received the M.S. and Ph.D. degrees in signal and information processing from the Harbin Institute of Technology, Harbin, China, in 2002 and 2005, respectively.

From 2006 to 2008, he held a Postdoctoral Research position with the College of Information and Communications Engineering, Harbin Engineering University, where he is currently a Professor. Since 2020, he has been with the College of Information and Communication Engineering, Dalian Nationalities University, Dalian, China. He has authored/coauthored two books about hyperspectral image processing and more than 130 papers in journals and conference proceedings. His research interests include remote sensing image processing and machine learning.

SCI 收录报告

经查 Web of Science-Core Collection ，石翠萍提供的如下文章已经被 SCI-Expanded（科学引文索引 ）收录，其收录记录简要信息摘选如下：

标题: Multihead Global Attention and Spatial Spectral Information Fusion for Remote Sensing Image Compression

作者: Shi, CP (Shi, Cuiping); Shi, KJ (Shi, Kaijie); Zhu, F (Zhu, Fei); Zeng, ZX (Zeng, Zexin); Wang, LG (Wang, Liguo)

来源出版物: IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING 卷: 17 页: 999-1015 DOI: 10.1109/JSTARS.2024.3417690 Published Date: 2024

Web of Science 核心合集中的 "被引频次": 2

被引频次合计: 2

入藏号: WOS:001270275700003

语言: English

文献类型: Article

地址: [Shi, Cuiping; Shi, Kaijie; Zhu, Fei; Zeng, Zexin] Qiqihar Univ, Dept Commun Engn, Qiqihar 161000, Peoples R China.

[Shi, Cuiping] Huzhou Univ, Coll Informat Engn, Huzhou 313000, Peoples R China.

[Wang, Liguo] Dalian Nationalities Univ, Coll Informat & Commun Engn, Dalian 116000, Peoples R China.

通讯作者地址: Shi, CP (通讯作者)，Qiqihar Univ, Dept Commun Engn, Qiqihar 161000, Peoples R China.

电子邮件地址: shicuiping@qqhru.edu.cn; 2022910313@qqhru.edu.cn; 2022935750@qqhru.edu.cn; 2022910311@qqhru.edu.cn; wangliguo@hrbeu.edu.cn

Affiliations: Qiqihar University; Huzhou University; Dalian Minzu University

IDS 号: YR7N2

ISSN: 1939-1404

eISSN: 2151-1535

来源出版物页码计数: 17

特此证明

湖州师范学院图书馆
查证人 钱绍
2024年11月12日