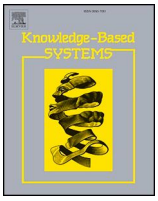
Knowledge-Based Systems 310 (2025) 112996

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Dynamic feature enhancement network guided by multi-dimensional collaborative edge information for remote sensing image compression

Cuiping Shi [a], Kaijie Shi [b,*], Zexin Zeng [b], Fei Zhu [b]

[a] College of Information Engineering, Huzhou University, Huzhou 313000, China
[b] College of Communication and Electronic Engineering, Qiqihar University, Qiqihar 161000, China

ARTICLE INFO

Keywords:
Remote sensing image compression
Structural features
Latent spatial representation capability
Deep learning
Dynamic feature enhancement

ABSTRACT

At present, there are some common problems in lossy compression methods for remote sensing images, such as block effect and blur effect, which are particularly evident at high compression ratios. Although some models have been developed that apply prior knowledge of local smoothing to probabilistic models to address these issues, it can result in significant loss of structural features. In this paper, a dynamic feature enhancement network guided by multi-dimensional collaborative edge information for remote sensing image compression (DMENet) is proposed, which can achieve high fidelity remote sensing image compression while preserving more structural features. Firstly, a multi-dimensional feature extraction module guided by edge information (MDEI) is carefully designed to extract structural features and edge features from images. These features are aligned structurally through loss to achieve high-quality restoration of structural features Secondly, a slice dynamic pyramid module (SDPM) is constructed to achieve dynamic extraction of irregular shaped features and multi-scale features. Furthermore, a latent representation space enhancement module (LSM) is proposed to address the issue of deep level feature loss in probabilistic models due to low information capacity. Finally, a high-quality remote sensing image compression is performed through the entire network under the guidance of a novel rate distortion optimization strategy (a constraint that focuses more on structural features). The experimental results show that compared with some advanced compression models, DMENet can compress remote sensing images more effectively.

1. Introduction

Remote sensing images can reflect various land features, such as terrain, temperature, and crop categories. Therefore, remote sensing images have been widely used in many fields such as environmental monitoring, geological science, military reconnaissance, etc. [1,2,3]. However, with the development of sensor technology, the resolution of remote sensing images continues to improve [4,5]. In addition, billions of remote sensing images are captured and transmitted every day [6,7]. Based on the above reasons, there is an urgent need for high fidelity remote sensing image compression methods with higher compression efficiency.

At present, traditional image compression methods have achieved some results [8,9]. The classic JPEG [10] and JPEG 2000 [11] mainly consist of three parts: image transformation, quantization, and entropy encoding. Firstly, transform and de quantify the image; Next, retain important information through quantification; Finally, use entropy encoding to compress the correlation coefficients of the solution. In addition, BPG [12,13] and WebP [14] with superior performance have also emerged. Some scholars have conducted targeted research and improvements on remote sensing images due to their high information entropy, rich texture, and various scale features. For example, Báscones et al. proposed a method that combines principal component analysis and JPEG2000 to compress hyperspectral image data, achieving dimensionality reduction and preserving the main spectral information [15]. Li et al. used MDSI as a quality evaluation index to improve the BPG compression algorithm. It provides more accurate remote sensing image quality control through a two-step compression strategy, achieving consistency between compression efficiency and image quality [16]. Traditional remote sensing image compression methods can be divided into predictive coding [17], transform coding [18], and vector quantization [19]. For example, 3D-MBLP uses prediction techniques to

* Corresponding author.
E-mail addresses: shicuiping@qqhru.edu.cn (C. Shi), 2022910313@qqhru.edu.cn (K. Shi), 2022910311@qqhru.edu.cn (Z. Zeng), 2022935750@qqhru.edu.cn (F. Zhu).

https://doi.org/10.1016/j.knosys.2025.112996
Received 7 October 2024; Received in revised form 25 December 2024; Accepted 8 January 2025
Available online 9 January 2025

first eliminate image spatial redundancy, then predict the current frequency band content, and finally efficiently encode the prediction error through an entropy decoder [20]. 3D-SPIHT, as a transformation compression method for 3D images, achieves efficient image compression by applying 3D wavelet transform in both spatial and spectral domains [21]. Qian developed an efficient and fast vector quantization compression algorithm for multispectral images, whose core strategy is to directly map the input vector to the best matching codeword index in the codebook, thereby significantly improving the efficiency of data transmission and storage [22]. However, traditional remote sensing image compression methods have the following limitations: under high compression ratios, there are obvious artifacts and block effects in the reconstructed images. Therefore, for remote sensing images, traditional compression methods are difficult to obtain high fidelity images at high compression ratios.

To seek breakthroughs, researchers are focusing on the popular deep learning technology in recent years. Classic deep learning-based image compression frameworks mainly include autoencoders (AE) [23,24] and variational autoencoders (VAE) [25,26]. The SSCNet proposed by the Riccardo team utilizes deep convolutional autoencoder technology to efficiently compress satellite image big data, while demonstrating excellent performance in compression ratio and signal reconstruction [23]. The Alves team designed a simplified version of the variational autoencoder, specifically designed to address the computational resource constraints in satellite image compression. By reducing the network size and optimizing the entropy model, this encoder effectively reduces computational complexity while ensuring compression efficiency [25]. However, compared to AE, the VAE framework exhibits stronger image reconstruction capabilities due to its continuous mapping space, such as generating images with smooth transitions between pixels. In recent years, VAE based baseline networks have shown outstanding performance in image compression, surpassing traditional methods and achieving efficient and high-quality compression [27,28,29,30]. VAE based image compression networks typically include a main encoder, entropy encoding, and a main decoder. They first compress images through neural networks, then quantize pixel data, and finally generate efficient bitstreams using traditional encoding techniques. In addition, to improve modeling accuracy and fully utilize prior information, some compression models introduce entropy models (Laplace entropy model, mixture Gaussian model, layered entropy model, etc.) into the framework [31,32,33,34,35]. The above model provides a strong theoretical basis for the remote sensing image compression task based on deep learning. Based on the above theories, some researchers have developed some deep learning-based compression networks for remote sensing images, and achieved good rate distortion performance [36,37,38,39,40]. Although these methods demonstrate some compression performance, the edges of the reconstructed image are often blurred at low bit rates. The proposed DMENet realizes the reconstruction of clear edges at low bit rate through the guidance of multi-dimensional edge information. In addition, the proposed DMENet also strengthens the extraction of multi-scale features and the feature capture ability of probability models. As a result, DMENet achieves excellent compression performance.

The common deep learning techniques used for remote sensing image compression mainly include three categories: image compression methods based on convolutional neural network (CNN) [41,42,43], image compression methods based on Transformer [44,45,46], and image compression methods based on generative adversarial network (GAN) [47,48]. In the CNN based method, Shao et al. proposed a deep learning based remote sensing image compression method, which decomposes remote sensing image features into high-frequency and low-frequency feature components through discrete wavelet transform (DWT), and enhances the quality of high and low-frequency features respectively [49]. In addition, Shao et al. proposed a compression network for remote sensing images, which effectively captures a wide range of contextual information by integrating long-range convolution and improved non local attention, achieving efficient compression while maintaining a lightweight design with low computational load [50]. In the Transformer based image compression method, Chuan et al. constructed a hyper prior network framework that integrates Transformer and CNN for remote sensing image compression, considering both local and non-local redundancy reduction. Through three-stage training to enhance generalization ability, the compression efficiency and quality were significantly improved [51]. In addition, Li et al. proposed an image compression method that uses deep neural networks to distinguish objects and backgrounds in remote sensing images, and reduces the bit rate by smoothing the background. Furthermore, combining Transformer and patch local attention module to optimize compression, balancing bit allocation through regional differentiation loss [45]. In GAN based image compression methods, Han et al. proposed an edge guided adversarial network aimed at preserving sharp texture information simultaneously [47]. In addition, Kan et al. proposed a remote sensing satellite image compression method based on conditional generative adversarial networks, which improved the details of reconstructed images by introducing Gaussian Laplacian loss and perceptual metrics [48]. Although the above methods have achieved good compression results, there may be artifacts and blurriness in the reconstructed images at high compression ratios. The essence of this phenomenon is the loss of structural features, which also leads to suboptimal rate distortion performance of these methods.

Remote sensing images contain rich structural features. Structural features mainly include edge features, texture features, structural information, etc. In high compression ratios, losing structural features can lead to artifacts, block effects, and blurring, resulting in the loss of important information in the image. Therefore, efficiently aligning the structural features between the original image and the reconstructed image has become a pressing challenge in the field of remote sensing image compression.

To alleviate the above problems, in this paper, a dynamic feature enhancement network guided by multi-dimensional collaborative edge information (DMENet) is proposed for remote sensing image compression. It can achieve high fidelity remote sensing image compression while preserving higher quality structural features. Structural features are divided into three categories: horizontal structural features, vertical structural features, and edge features. Based on this, this paper constructs horizontal attention (HA), vertical attention (VA), and edge feature extraction module (EEM) respectively. A multi-dimensional feature extraction module guided by edge information (MDEI) is designed using HA, VA, and EEM. And based on this, a multi-dimensional synergistic loss guide by edge information ($Loss_{\mathrm{MDSE}}$) is constructed to align the structural features between the original image and the reconstructed image. Secondly, remote sensing images often contain irregular shaped features (such as features where multiple objects overlap, resulting in peculiar shapes) and multi-scale features. Therefore, this paper constructs a slicing strategy (SS) for improving computational efficiency and reducing memory usage, a strange feature extraction block (SEB) for enhancing irregular feature extraction ability, and a pyramid feature enhancement (PEB) for enhancing multi-scale feature capture ability. Based on the above work, a slice dynamic pyramid module (SDPM) is constructed to achieve dynamic extraction of irregular shape features and multi-scale features. Finally, the latent spatial representation ability of conventional probability models is insufficient. The feature maps in probability models are deep features extracted multiple times, which contain a large number of spatial and channel features. However, the convolution blocks in conventional probability models cannot accommodate a large number of features, resulting in feature loss. For this, this paper proposes a latent representation space enhancement module (LSM) to enhance the deep feature representation ability of probability models. In summary, this paper constructs a high-performance DMENet based on the proposed MDEI, SDPM, LSM, and $Loss_{\mathrm{MDSE}}$.

This study conducted extensive experiments on three remote sensing

image datasets: San Francisco [52], NWPU-RESISC45 [53], and UC Merced [54]. The experimental results show that compared to some advanced compression methods, the proposed DMENet performs better in evaluation metrics such as peak signal-to-noise ratio (PSNR) and multiscale structural similarity index metric (MS-SSIM). In addition, the reconstructed images are also used for remote sensing image scene classification to test the impact of compression methods on downstream tasks. The experiment shows that the classification performance of remote sensing images reconstructed by DMENet is the best.

The main contributions of this paper are summarized below:

1) A multi-dimensional feature extraction module guided by edge information (MDEI) is proposed. Based on this, a multi-dimensional synergistic loss guide by edge information ($Loss_{\mathrm{MDSE}}$) is further constructed. It can effectively extract multi-dimensional structural features by aligning horizontal structural features, vertical structural features, and edge features.
2) A slice dynamic pyramid module (SDPM) is designed. It can extract irregular shape features and multi-scale features through a new slicing strategy, a dynamic feature capture mechanism, and the multi-scale feature enhancement block.
3) A latent representation space enhancement module (LSM) is constructed. It extracts multi-level spatial and channel features through spatial attention and channel attention respectively, which can effectively enhance the deep feature representation ability of the probability model.
4) This study effectively integrates MDEI, SDPM, LSM, and a rate distortion optimization strategy for structural feature alignment to construct a dynamic feature enhancement network guided by multi-dimensional collaborative edge information (DMENet) for remote sensing image compression. Extensive experiments on the San Francisco, NWPU-RESISC45, and UC Merced datasets have demonstrated the superior performance of DMENet in multiple evaluation metrics.

The remainder of the study is organized as follows: In Section II, the proposed DMENet framework and the details of each module are elaborated. In Section III, this paper comprehensively analyzes and compares the proposed DMENet and other compression methods through a large number of experiments. In Section IV, conclusions and future work are discussed.

2. Methodology

In this section, the proposed DMENet, as well as the modules MDEI, SDPM, LSM, and a rate distortion optimization strategy for structural feature alignment will be introduced in detail.

2.1. The Overall framework of the proposed DMENet

The proposed DMENet can preserve high-quality structural features and comprehensively improve the compression performance of the model from the perspective of aligning multidimensional structural features between the original image and the reconstructed image. It achieves high-quality remote sensing image compression through MDEI for extracting structural features, SDPM for dynamically extracting irregular features and multi-scale features, LSM for enhancing the deep level feature representation ability of the probability model, and a rate distortion optimization strategy focused on structural feature alignment. MDEI involves three sub modules, including HA capable of extracting horizontal structural features, VA capable of extracting vertical structural features, and EEM capable of extracting edge features. SDPM involves three sub modules, including SS for channel slicing, SEB for extracting irregular features, and PEB for extracting multi-scale features. LSM involves two parts of latent spatial representation enhancement, i.e., spatial feature enhancement and channel feature enhancement. In addition, this paper constructs a method for calculating structural differences using MDEI and proposes a rate distortion optimization strategy focused on structural differences.

The overall structure of the proposed DMENet is shown in Fig. 1. This paper designs compression block (C block 1-4) for compression and reconstruction block (R block 1-4) for reconstruction by reasonably selecting the size of convolution kernels and reallocating the number of channels, achieving excellent rate distortion performance at low complexity. The specific parameters of C block 1-4 and R block 1-4 are shown in Table 1. The probability model is mainly used for probability modeling, which includes a hyperprior network (Hyper encoder, Hyper decoder, LSM), Q (quantizer), AE (arithmetic encoding), and AD (arithmetic decoding). The specific parameters of the Hyper encoder and Hyper decoder are shown in Table 2, and the minimum form of data existence in this model (bit stream) is between AE and AD. The hyperprior network is utilized to learn the probability model (i.e. entropy model) that entropy encoding relies on, and is also used to generate the parameters of the entropy model (i.e. mean parameter μ_i and scale parameter $\sigma_i{}^2$). The entropy model is modeled as a conditional Gaussian. MDEI (where a=0.5, b=0.5, c=0.1, d=0.5, e=0.5, f=0.1) is used to calculate the structural difference $Loss_{\mathrm{MDSE}}$ between the compressed and reconstructed parts, where $Loss_{\mathrm{MDSE}}$ represents the difference between the structural features of the compression and reconstruction parts. The smaller the loss value, the smaller the structural difference between the two sides, that is, the higher the quality of the structural features. The loss used here is mean squared error (MSE), which can be expressed as formula 1. In the rate-distortion optimization, R represents entropy rate, λ represents penalty coefficients used to control different bit rates, and D represents distortion (calculated by MSE).

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(-X)^2 \tag{1}$$

where, m denotes the number of pixels, $\widehat{X}$ denotes reconstructed image, X denotes original image.

In Tables 1 and 2, N represents the number of channels, ↓ represents downsampling, ↑ represents upsampling, and RELU represents the linear rectification function. GDN stands for generalized split normalization function and IGDN stands for its inverse operation, which are nonlinear activation functions and are more suitable for normalizing image data than other normalization functions

The overall working process of DMENet is as follows. (1) Image compression part: Firstly, the remote sensing image data blocks are processed through C block 1-2 to obtain shallow feature maps. Afterwards, dynamic extraction of irregular shaped features and multi-scale features is strengthened through SDPM, and the data is processed through C block 3-4 to obtain a deep level feature map. Then, the initially compressed deep feature maps are further processed through quantization, arithmetic coding, and LSM enhanced probability model to remove statistical redundancy, resulting in the minimum bit stream of the model after data processing. (2) Image reconstruction part: The model combines the obtained bitstream with the mean parameter μ_i and scale parameter $\sigma_i{}^2$ in the probability model to reconstruct the image with high quality using R block4, R block3, SDPM, R block2, and R block1. Finally, the reconstructed and original images are input into MDEI for structural feature alignment. And add the obtained $Loss_{\mathrm{MDSE}}$ to $Loss_{Total}$ for targeted rate distortion optimization of structural features.

2.2. MDEI

Remote sensing images contain rich structural features, including edge features, texture features, and structural information in various directions. The loss of structural features can lead to block effects and blurring effects. Therefore, a three-branched MDEI is designed to extract structural features at different levels. In addition, this paper divides

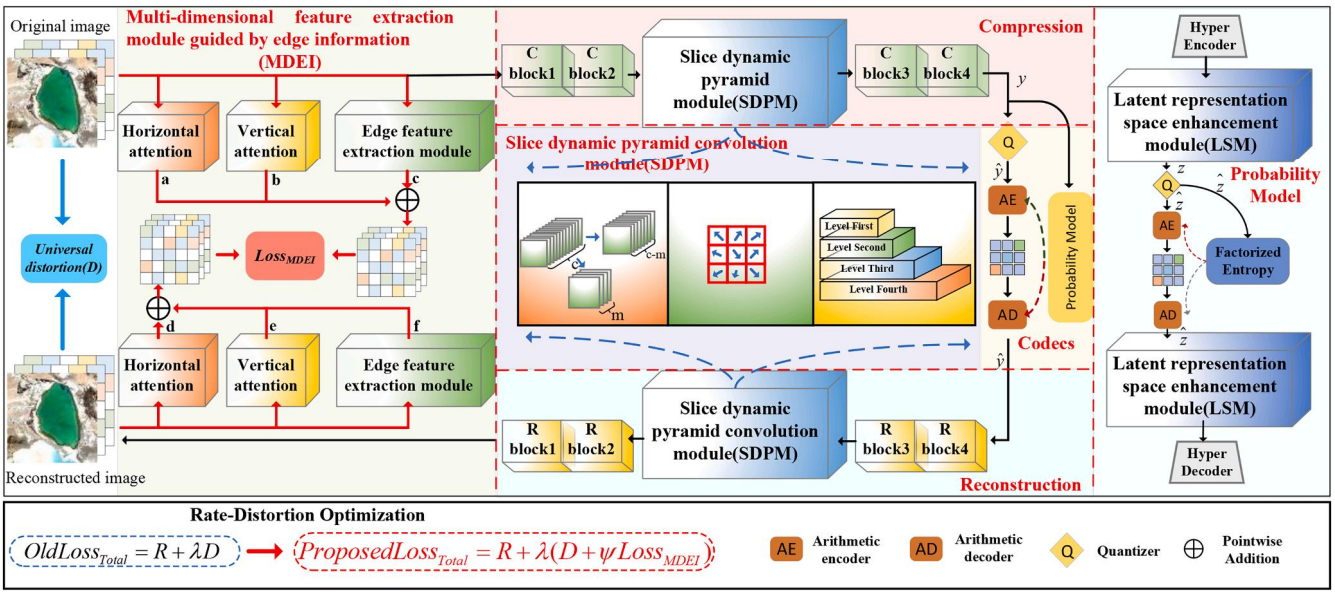


Fig. 1. The overall structure of the proposed DMENet.

Table 1
Specific parameters of C block and R block.

Block	First layer	Second layer
C block 1	Conv2D 7×7 3 N/4 2↓	GDN
C block 2	Conv2D 3×3 N/4 N/2 2↓	GDN
C block 3	Conv2D 3×3 N/2 3N/4 2↓	GDN
C block 4	Conv2D 3×3 3N/4 N 2↓	GDN
R block 1	Conv2D 3×3 N/4 3 2↑	IGDN
R block 2	Conv2D 3×3 N/2 N/4 2↑	IGDN
R block 3	Conv2D 3×3 3N/4 N/2 2↑	IGDN
R block 4	Conv2D 3×3 N 3N/4 2↑	IGDN

Table 2
Specific parameters of Hyper encoder and Hyper decoder.

	Hyper encoder	Hyper decoder
Layer1	Conv2D 3×3 N N 1	Conv2D 3×3 N N 2↑
Layer2	RELU	RELU
Layer3	Conv2D 3×3 N N 2↓	Conv2D 3×3 N N 2↑
Layer4	RELU	RELU
Layer5	Conv2D 3×3 N N 2↓	Conv2D 3×3 N N 1
Layer6	-	RELU

structural features into horizontal structural features, vertical structural features, and edge features. Based on these three structural features, this paper designs HA, VA, and EEM respectively, and fuses the obtained features by multiplying them with weight coefficients to obtain feature maps for calculating structural loss $Loss_{\mathrm{MDSE}}$. Finally, a rate-distortion optimization strategy focusing on structural feature alignment is constructed through $Loss_{\mathrm{MDSE}}$.

MDEI is essentially a module that aligns the structural features between the original image and the reconstructed image through loss. It mainly consists of three parts: the module MDEI (Origin) for extracting the structural features of the original image, the loss and the module MDEI (Reconstruction) for extracting the structural features of the reconstructed image. Due to the symmetrical structure of MDEI (Origin) and MDEI (Reconstruction), only MDEI (Origin) will be introduced here.

MDEI (Origin) is a three-branched structure that includes HA for capturing horizontal structural features, EEM for capturing edge features, and VA for capturing vertical structural features. Firstly, input the data block into MDEI (Origin) and use the permute to redirect it, resulting in three data blocks $X4$, $X1$, and $X5$. In HA, $X4$ is transformed into a vector of shape $1 \times 1 \times W$ using Avgpool and Stdpool, preserving only the horizontal features of the data. Afterwards, stripe convolution is adopted to extract the horizontal structural features of the data under low complexity conditions. Afterwards, a series of deformation operations are performed to restore the original shape of the data block. The process of HA can be expressed as: (Fig. 2).

$$X6_{H\times W\times C} = \mathrm{Reconstruction}(BarConv_{1\times k}(a(Avgpool(X4_{H\times C\times W})) + b(Stdpool(X4_{H\times C\times W})))) \quad (2)$$

Here, $BarConv_{1\times k}$ represents a stripe convolution with a kernel shape of $1 \times$ k, where a and b represent weight coefficients of 0.5 and 0.5, respectively. *Reconstruction* includes *Sigmoid*, *Expand*, and *Permute*, mainly used to restore the data block to its original shape and fuse it with the other two branches.

The working principle of VA is similar to that of HA, with the difference being that it extracts vertical structural features. The process of VA can be expressed as:

$$X7_{H\times W\times C} = \mathrm{Reconstruction}(BarConv_{1\times k}(c(Avgpool(X5_{C\times W\times H})) + d(Stdpool(X5_{C\times W\times H})))) \quad (3)$$

Here, $BarConv_{1\times k}$ represents a stripe convolution with a kernel shape of $1 \times$ k, where c and d represent weight coefficients of 0.5 and 0.5, respectively. *Reconstruction* includes *Sigmoid*, *Expand*, and *Permute*, mainly used to restore the data block to its original shape and fuse it with the other two branches.

The EEM branch is designed for edge feature extraction, and its structural diagram is shown in Fig. 3. The convolution kernel of Gaussian convolution is initialized with Gaussian to smooth and denoise the image. The downsample used here is interval sampling downsampling. Upsample uses interpolation with a value of 0 inserted. After the first Gaussian convolution, downsampling, upsampling, and second Gaussian convolution, the differences in features in the input data are greatly reduced, resulting in all features tending towards contextualization. Then, the original data block (containing rich edge features) is used to subtract the data block whose features have been contextualized to obtain the edge features of the image. It is worth mentioning that the initial values of M and N here are 1 and 256, respectively. In this way, the values of each point in the convolution kernel can conform to the Gaussian distribution, to better extract the edge features. M and N here are not fixed values, just initial values, and they will automatically fit the network during the network training process. In addition, the

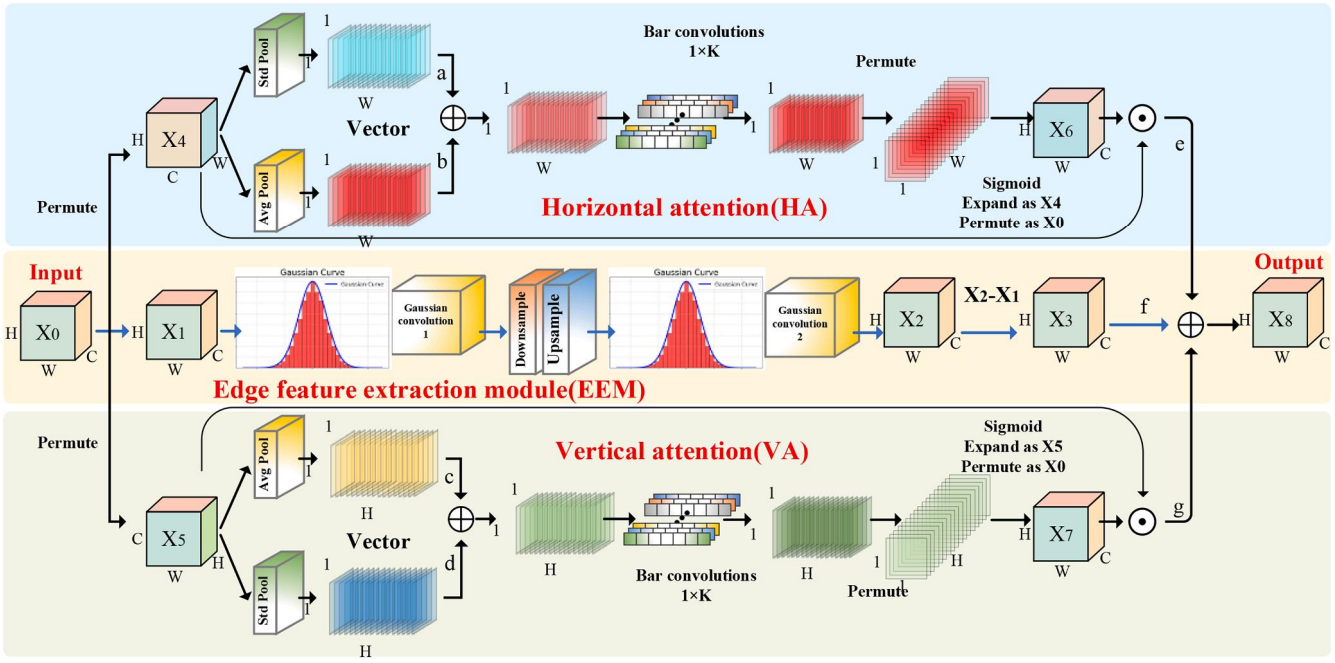


Fig. 2. Schematic diagram of MDEI (Origin), where Input represents the input feature map, Output represents the output feature map, and weight coefficients a, b, c, d, e, f, g represents 0.5, 0.5, 0.5, 0.5, 0.5, 0.1, 0.5, respectively. Here, Avgpool represents global average pooling, and Stdpool represents global standard deviation pooling. C. H and W respectively represent the number of channels, width, and height of the data block. The convolution kernel shape of Bar convolution is set to 1 × 3.

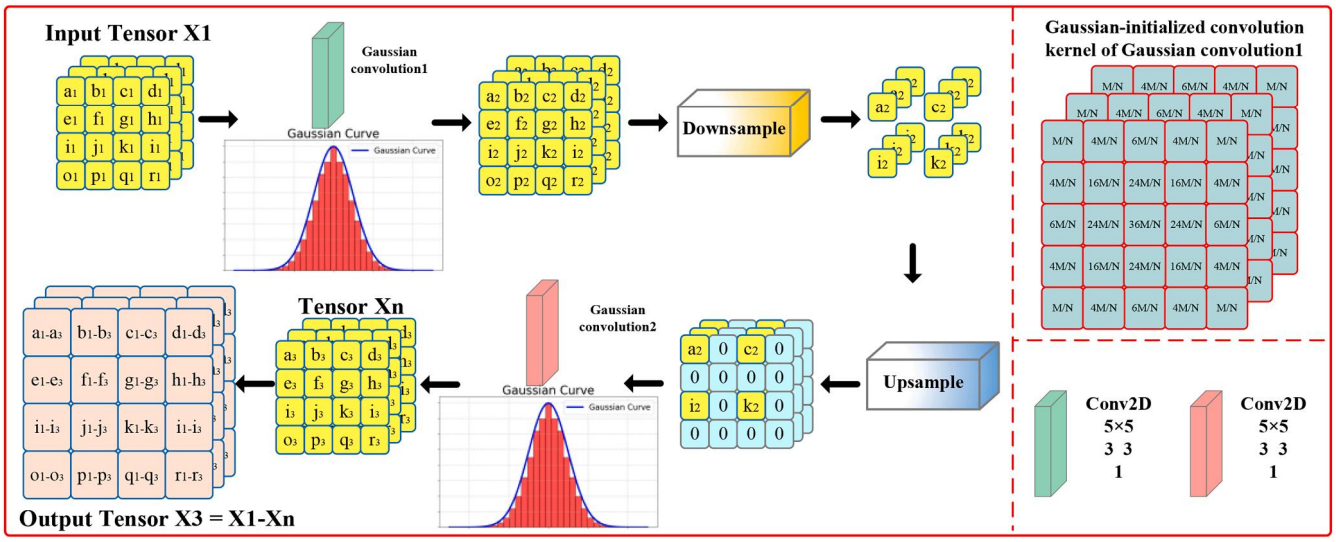


Fig. 3. The schematic diagram of EEM structure, where input tensor represents the input feature map, output tensor represents the output feature map, and *M*and*N*are 1 and 256, respectively. The initial convolution kernel of Gaussian convolution 1 is shown in the upper right. The value of the initialized convolution kernel of Gaussian convolution 2 is 4 times that of Gaussian convolution 1. In Conv2D 5 × 5 3 3 1, 5 × 5 represents the shape of the convolution kernel, the first 3 represents the number of input channels, the last 3 represents the number of output channels, and 1 represents the number of channel groups. The parameter settings for other convolutions are also the same. The Input data is 3 bands.

convolution kernel size is set to 5 because the input data has a large spatial size. Therefore, larger convolutional kernels are adopted to extract edge features over long distances. The process of EEM can be expressed as:

$$X3_{H\times W\times C} = X1_{H\times W\times C} - GaussianConv_2(Upsample(Downsample(GaussianConv_1(X1_{H\times W\times C})))) \quad (4)$$

The structural features of the original image are obtained by fusing the data of the three branches:

$$Output_{MDEI(Origin)} = e(X6_{H\times W\times C}) + f(X3_{H\times W\times C}) + g(X7_{H\times W\times C}) \quad (5)$$

where e, f, and g are the weight coefficients of the corresponding branches, which are 0.5, 0.1, and 0.5, respectively.

Similarly, the structural feature $Output_{MDEI(Reconstruction)}$ of the reconstructed image is also obtained.

$Loss_{MDSE}$ can be calculated from $Output_{MDEI(Origin)}$ and $Output_{MDEI(Reconstruction)}$:

$$Loss_{MDSE} = L_{MSE}\left(Output_{MDEI(Origin)}, Output_{MDEI(Reconstruction)}\right) \quad (6)$$

where L_{MSE} denotes the loss measured using *MSE*.

Finally, $Loss_{\mathrm{MDSE}}$ aligns the complex structural features between the original image and the reconstructed image through rate distortion optimization to achieve high-quality image compression. This is described in more detail in the rate-distortion optimization part.

2.3. SDPM

Remote sensing images often contain a lot of irregular features (such as features formed by multiple objects overlapping) and multi-scale features (such as short-range information such as ships, long-range information such as rivers crossing the entire city). Therefore, an SDPM is designed to achieve dynamic extraction of irregular features and multi-scale features. The overall structure of SDPM is shown in Fig. 4. It mainly includes three core components: SS for channel slicing, SEB for enhancing irregular feature extraction capability, and PEB for enhancing multi-scale feature capture capability.

The remote sensing image data block is input into SDPM, and SS will slice it along the channel dimension. Afterwards, some channels will remain unchanged, while others will be feature extracted through convolution. Finally, the two sets of features will be concatenated in the channel dimension. This strategy uses a channel preservation method like residual connections to make the model more cautious and faster in adjusting parameters, significantly improving the model's understanding of complex channel features. The process of SS can be represented as:

$$Output_{SS} = F_{concat}(Tensor_{H\times W\times m}(Input_{H\times W\times C}), \mathrm{Conv}_{3\times 3}(Tensor_{H\times W\times(c-m)}(Input_{H\times W\times C}))) \quad (7)$$

Here, $Tensor_{H\times W\times m}(Input_{H\times W\times C})$ is a tensor of shape $H\times W\times c$ obtained by slicing the input features, $Tensor_{H\times W\times(c-m)}(Input_{H\times W\times C})$ is a tensor of shape $H\times W\times(c-m)$ obtained by slicing the input features, $\mathrm{Conv}_{3\times 3}$ represents a convolution with a kernel size of 3×3, $F_{concat}(,)$ represents a functional function that concatenates two tensors in the channel dimension, and $Output_{SS}$ represents the output feature map of SS.

Secondly, $Output_{SS}$ will be input into SEB for extracting irregular features. In remote sensing images, many features are composed of multiple overlapping objects, as well as many bizarre features. These shape of these features are irregular, and conventional square convolution kernels are difficult to effectively sample their features. Therefore, deformable convolution is introduced to construct SEB, thereby enhancing the extraction of irregular features. The convolution here is combined with offset learning, which allows the size and position of the convolution kernel to be dynamically adjusted according to the image. This allows the shape of the convolution kernel to adapt to the shape and size of different objects. Here, $Output_{\mathrm{SEB}}$ represents the output feature map of SEB.

Finally, $Output_{\mathrm{SEB}}$ is input into PEB for multi-scale feature extraction, and the problem of insufficient receptive field and loss of detail features is solved. Four sets of well-designed convolutions reduce the dimensionality of the feature map to obtain high-quality spatial features. Then, the four dimensionally reduced features are spliced together in the channel dimension, and the channel features are fused through point convolution. Finally, residual connections are introduced to speed up the training of the network. The process of PEB can be represented as:

$$Output_{PEB} = Conv_{1\times 1}(F_{concat}(A, B, C, D) + Output_{SEB}) \quad (8)$$

Here, $Conv_{1\times 1}$ represents point convolution, $F_{concat}(,)$ represents a functional function that splices two tensors in the channel dimension, A, B, C, D represent four sets of well-designed convolutions, $Output_{PEB}$ represents the output feature map of the PEB (i.e., the final output feature map of the SDPM).

2.4. LSM

The latent spatial representation ability of conventional probability models is insufficient. The feature maps in probability models are deep features extracted multiple times, which contain a large number of spatial and channel features. However, the convolution blocks in conventional probability models cannot accommodate a large number of features, resulting in feature loss. Based on this, this paper designs LSM to enhance the deep feature representation ability of probability models, thereby greatly improving the accuracy of probability models (Fig. 5).

The feature extraction process of LSM is divided into two parts, including large kernel bar convolution for spatial feature enhancement (LKE) and channel feature extraction. LKE adopts a four-branched structure, and the convolutions selected are all large kernel convolutions, which can greatly improve the latent spatial representation ability. However, large kernel convolution can bring huge computational burden. Therefore, strip convolution is adopted to obtain the same receptive field with half the computational complexity. In addition, a residual branch is introduced to accelerate the training of the network. It should be noted that the convolution kernels in LKE are all single channel patterns, and there is no interaction of channel features.

The process of LKE can be represented as:

$$Output_{LKE} = E + Conv_{1\times 7}(Conv_{7\times 1}(E)) + Conv_{1\times 11}(Conv_{11\times 1}(E)) + Conv_{1\times 21}(Conv_{21\times 1}(E)) \quad (9)$$

Here, E represents the feature of the Input in LKE after a 5×5 convolution.

Then, a point convolution is adopted to extract the channel features

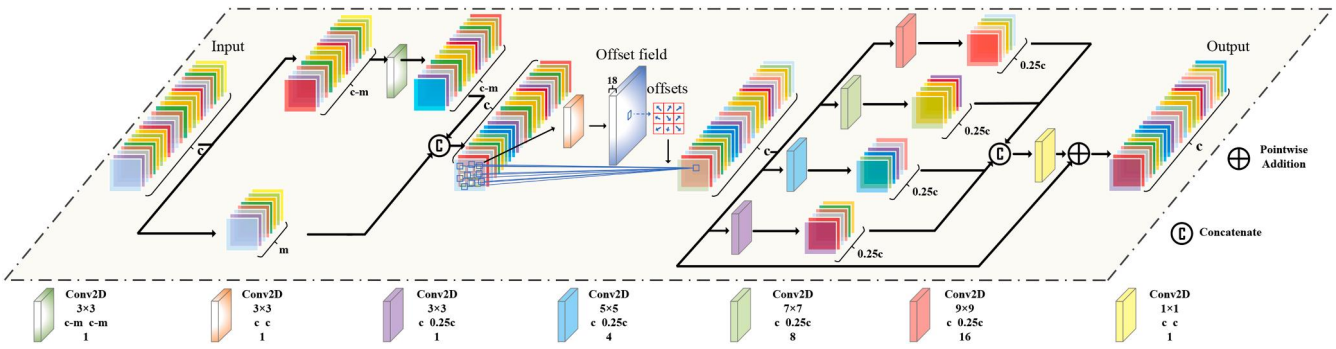


Fig. 4. Schematic diagram of SDPM. Input represents the input feature map, and Output represents the output feature map. In Conv2D 3×3 c-m c-m 1, 3×3 represents the shape of the convolution kernel, the first c-m represents the number of input channels, the last c-m represents the number of output channels, and 1 represents the number of channel groups. The parameter settings for other convolutions are also the same. The input data here is mapped to high-dimensional data and therefore is multi band.

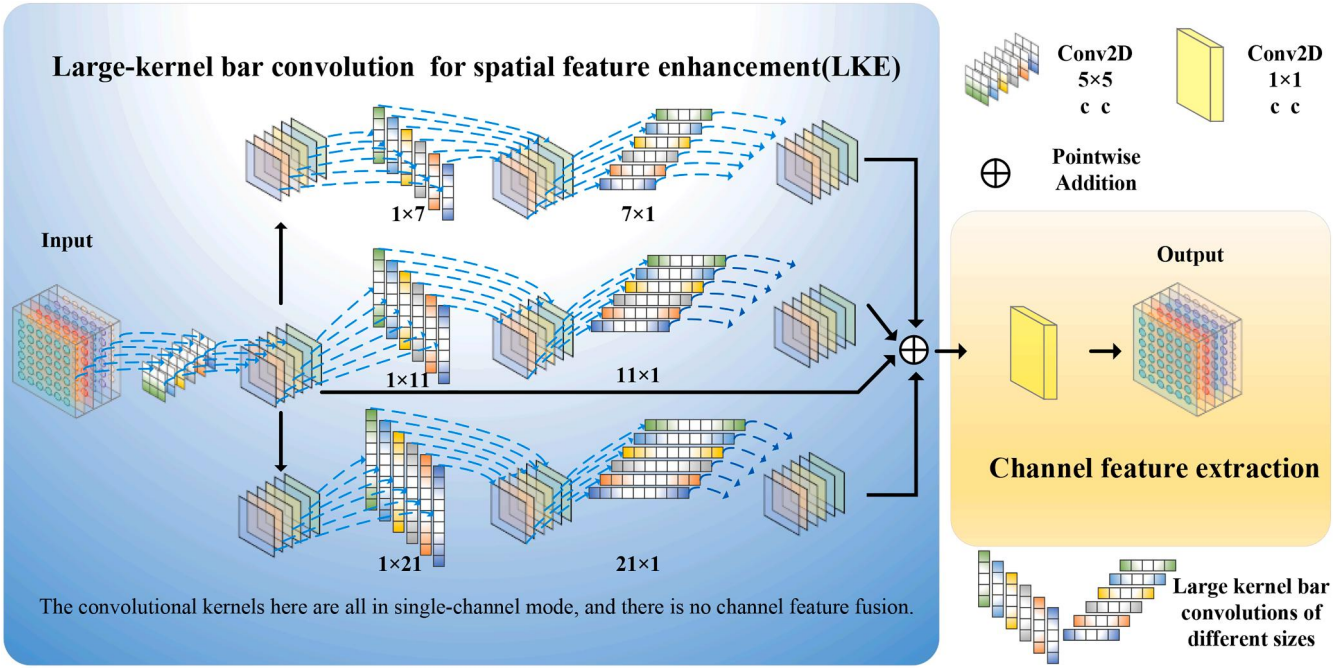


Fig. 5. Schematic diagram of LSM.

and the final feature map is obtained. That is

$$Output_{LSM} = Conv_{1\times1}(Output_{LKE}) \tag{10}$$

In general, LSM separates spatial feature extraction from channel feature extraction. It achieves high latent spatial representation capability at low complexity through the strip convolution of large kernels. As a result, the modeling accuracy of the probabilistic model is greatly improved.

2.5. Rate-distortion optimization

The goal of the compression framework is to achieve a balance between compression and distortion. To achieve this, a rate distortion optimization strategy is often added to the compression framework to guide the model for efficient training. In short, the strategy is designed to ensure that the data is compressed with as little information loss as possible. The rate distortion optimization strategy can be represented as

$$\text{argmin} Loss_{Total} = R + \lambda D \tag{11}$$

Here, R represents entropy rate, which is the cross-entropy between the latent edge distribution and the learning entropy model. D represents distortion between the original image and the reconstructed image. Different bitrates can be controlled by adjusting the penalty coefficient λ.

$$R = R_{\hat{y}} + R_{\hat{z}} \tag{12}$$

Here, the bitrate consists of the latent representation information $\hat{y}$ together with the side information $\hat{z}$.

$$R_{\hat{y}} = -\sum_i \log_2(p_{\hat{y}}(\hat{y})) \tag{13}$$

$$R_{\hat{z}} = -\sum_i \log_2(p_{\hat{z}}(\hat{z})) \tag{14}$$

Here, $p_{\hat{y}}$ is an entropy model that can be learned, $p_{\hat{z}}$ represents Hyper encoder.

To further improve the quality of image compression, a novel rate distortion optimization strategy is proposed in this paper. $Loss_{\text{MDSE}}$ is introduced into $Loss_{Total}$, that is, the ability of the whole network to reconstruct structural features is improved by aligning the structural features between the original image and the reconstructed image. The new rate distortion optimization strategy can be expressed as:

$$\text{argmin} ProposedLoss_{Total} = R + \lambda(D + \psi Loss_{MDEI}) \tag{15}$$

Here, ψ represents the coefficient of $Loss_{\text{MDSE}}$.

3. Experimental results and analysis

Sufficient experiments have been carried out on some remote sensing image datasets, including San Francisco [52], NWPU-RESISC45 [53], and UC-Merced [54]. These datasets contain a wealth of ground object information, which can effectively evaluate the performance of DMENet. In this paper, DMENet is compared with some excellent compression methods, including traditional codecs and deep learning-based compression models, to verify the superiority of the proposed method. Traditional image compression methods include JPEG2000 [11], BPG [55], and WebP [14]. Compression models based on deep learning include Minnen et al. [56], Balle et al. (hyperprior) [57], Balle et al. (factorized-relu) [57], Tong 2023 [58] and Shi2024[46]. Experimental results show that the proposed DMENet has the best compression performance in both PSNR and MS-SSIM evaluation indicators. In addition, the quality of the reconstructed images obtained by different compression methods is evaluated through the classification task, which further verifies the superiority of DMENet.

3.1. Experimental setting

3.1.1. Introduction to remote sensing image dataset

1. Dataset San Francisco: San Francisco is a dataset of remotely sensed images from [52]. It is a remote sensing image with a resolution of 17408×17408, covering a variety of feature information such as buildings, coasts, highways, ports, lakes, etc. In this paper, it is cropped to 256×256 pixel images, and 3000 valid images are selected to form the dataset. These images are divided into a training set, a validation set, and a test set at a ratio of 8:1:1. Fig. 6 shows some of the samples.

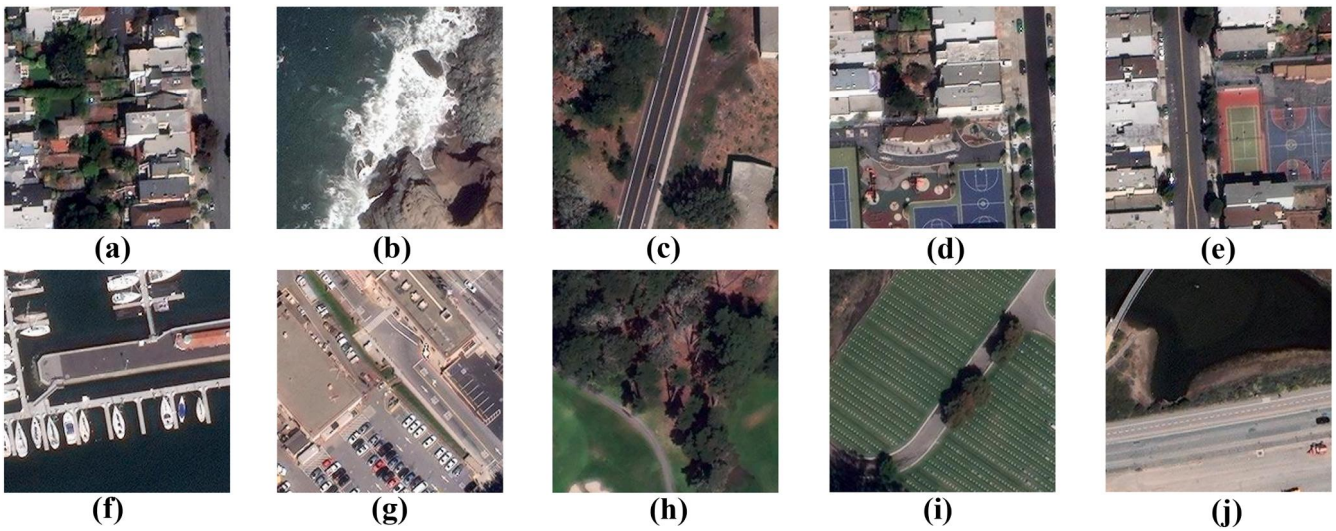


Fig. 6. Some images from San Francisco dataset. (a) Buildings (b) Coastline (c) Highway (d) Basketball court (e) Tennis court (f) Harbour (g) Parking lot (h) Forest (i) Farmland (j) Lake.

2. Dataset NWPU-RESISC45: NWPU-RESISC45 is provided by Northwestern Polytechnical University (NWPU). The dataset contains a total of 45 different remote sensing image scene categories. Each category contains 700 images, each with a resolution of 256×256 pixels. The dataset contains a variety of geographical environments and scenarios, including airports, deserts, churches, forests, etc. The 140 images in each category were selected to form a dataset of 6,300 remote sensing images, which was then divided into a training set, a validation set, and a test set at a ratio of 8:1:1. Fig. 7 gives some of the samples.
3. Dataset UC-Merced: UC-Merced is a remote sensing image dataset provided by the University of California, Merced. The UC-Merced dataset consists of 21 different categories, each consisting of 100 images. A total of 2100 images are included, each with a resolution of 256×256 pixels. The images include farmland, airports, forests and other landform scenes. The dataset UC-Merced is divided into a training set, a validation set, and a test set at a ratio of 8:1:1. Fig. 8 shows some of the samples.

3.1.2. Evaluation indicators

To evaluate the quality of reconstructed images, two commonly used evaluation metrics are adopted, i.e., peak signal-to-noise ratio (PSNR) and multi-scale structural similarity index measurement (MS-SSIM). In the part of remote sensing scene image classification, the overall accuracy (OA) and confusion matrix (CM) are also used to measure the classification performance.

1)PSNR: PSNR compares the reconstructed image to the original image from the point of view of the mean square error. The higher the PSNR value, the higher the fidelity of the reconstructed image. The peak signal-to-noise ratio can be expressed as:

$$PSNR(X,\widehat{X}) = \frac{1}{C}\sum_{i=1}^{C} 10\log_{10}\left(\frac{\max^2(X^i)}{MSE_i}\right) \tag{16}$$

Here, *MSE* represents the mean square error between the original image and the reconstructed image. $\max^2(X^{(i)})$ represents the square of the largest pixel in band *i.C* represents the number of bands.

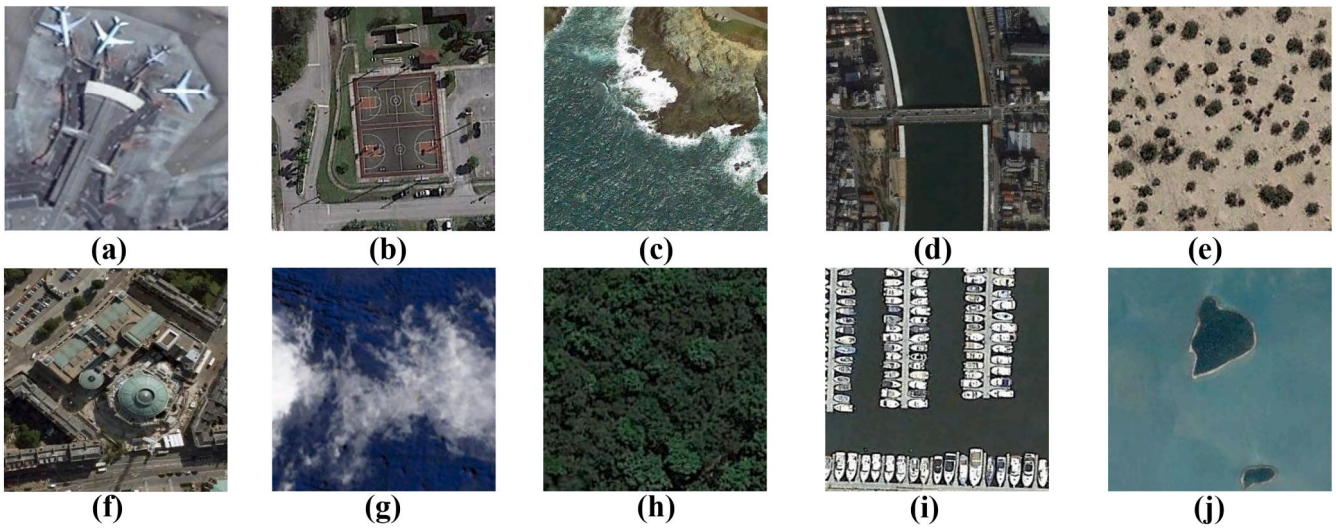


Fig. 7. Some images from NWPU-RESISC45 dataset. (a) Airport, (b), Basketball court, (c), Beach, (d) Bridge, (e), Desert, (f) Church, (g) Clouds, (h) Forest, (i) Port, (j) Island.

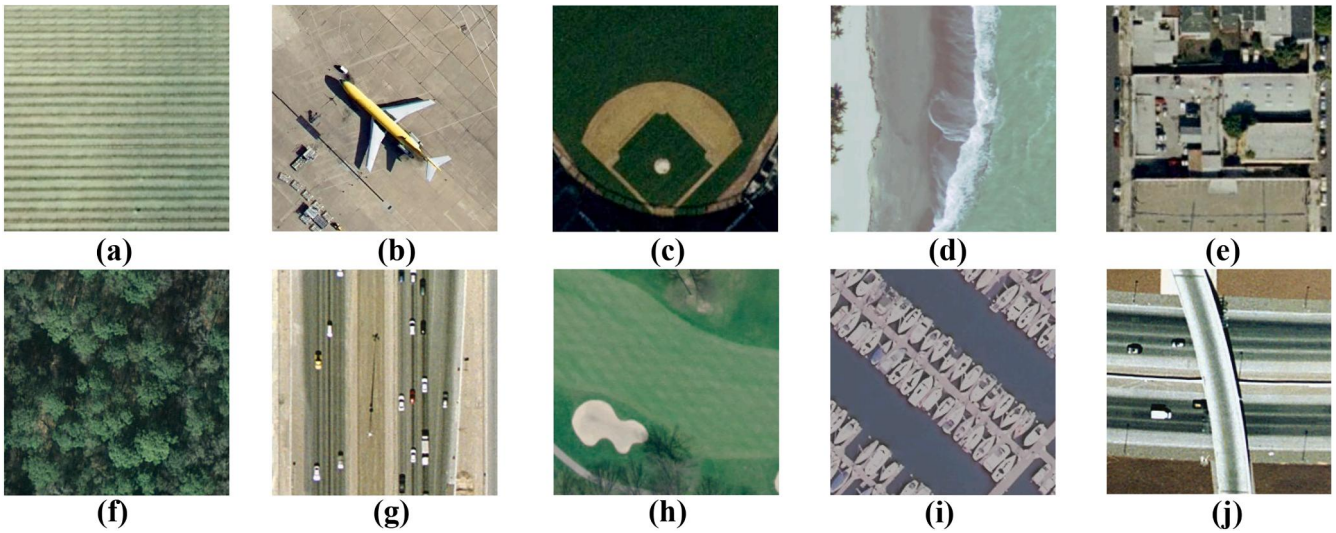


Fig. 8. Some images from UC-Merced dataset (a) Farmland (b) Airplanes (c) Baseball Stadiums (d) Beaches (e) Buildings (f) Forests (g) Roads (h) Golf Courses (i) Ports (j) Overpasses.

2)MS-SSIM: MS-SSIM is a multi-scale structural similarity index. It measures the difference between the original image and the reconstructed image by merging image details at different resolutions. The value ranges from 0 to 1, with higher values indicating higher similarity and higher quality of the reconstructed image. The formula for MS-SSIM can be expressed as:

$$D_{MS-SSIM} = 1 - \prod_{m=1}^{M} \left(\frac{2\mu_X \mu_{\widehat{X}} + C_1}{{\mu_X}^2 + {\mu_{\widehat{X}}}^2 + C_1} \right)^{\alpha_m} \left(\frac{2\sigma_{X\widehat{X}} + C_2}{{\sigma_X}^2 + {\sigma_{\widehat{X}}}^2 + C_2} \right)^{\zeta_m} \tag{17}$$

Here, M represents different resolutions, 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, $\sigma_{X\widehat{X}}$ represent the covariance between the original image and the reconstructed image, α_m and ζ_m represent the relative importance between the two terms, C_1 and C_2 are constant terms to prevent the divisor from being 0.

To clearly compare the differences in MS-SSIM values, they are converted into decibel values. This process can be expressed as:

$$MS - SSIM = -10\log_{10}(1 - D_{MS-SSIM}) \tag{18}$$

3)Classification indicators of remote sensing scenes: In this paper, two widely used remote sensing scene classification evaluation indicators are selected to measure the quality of the reconstructed image, including OA and CM. The OA value is obtained by dividing the number of correctly classified images by the total number of test images, and it reflects the overall performance of a classification model. CM reflects the degree of confusion and detailed classification errors between different scene categories. Each row in the CM represents the true category, and each column represents the predicted category.

3.1.3. Experimental environment and parameter settings

In this study, the proposed DMENet is implemented by PyTorch. The Adam optimizer was chosen. In this network, two optimizers are used, one is the main optimizer between the main encoder (Compression) and the main decoder (Reconstruction), and the other is the auxiliary optimizer between the hyper encoder and the hyper decoder. For the main optimizer, the initial learning rate is set at 10^{-4}, and the optimal model of DMENet will be stored when the learning rate decays to 10^{-6} during network training. For the auxiliary optimizer, its initial learning rate is set at 10^{-3}. During training, the batch size is set to 8. In this experiment, the neural network models are trained on an NVIDIA GeForce RTX 3090, and the traditional codecs are performed on a CPU (i9-9900K CPU@3.60GHz). For the sake of fairness, all experiments in this paper were conducted in the above environment. The penalty coefficient λ used in this paper is [0.660, 0.508, 0.211, 0.072, 0.033, 0.013, 0.007]. In the proposed rate distortion optimization strategy, the coefficient ψ of $Loss_{\mathrm{MDSE}}$ is set to 0.0185. In C block1-4, R block1-4, Hyper encoder, and Hyper decoder, N is set to 256. In the classification of remote sensing scenes, the benchmark model used for testing was EMTCAL (Efficient Multiscale Transformer and Cross-Level Attention Learning) [59]. The dataset used for training is NWPU-RESISC45, and the training-to-test ratio is 10%-90%. The images used for compression and the images used for remote sensing scene classification training are not crossed. The reconstructed images are only used for testing the classification performance, not for the training of the classification network.

3.2. Rate distortion performance

In this experiment, PSNR and MS-SSIM were used to evaluate the rate distortion performance of the model. Fig.s 9-11 show the PSNR and MS-SSIM rate distortion performance curves for all model experiments on the San Francisco, NWPU-RESISC45, and UC-Merced datasets, respectively. In traditional codec-based image compression methods, BPG exhibits excellent rate distortion performance, which is better than WebP and JPEG2000 in most cases. This significant advantage is mainly due to BPG's multi-channel encoding technology, which allows different color channels to be encoded independently, allowing for fine control of image detail features. In the image compression method based on deep learning, Balle et al. (factorized-relu) exhibits relatively poor rate distortion performance, mainly because it only uses a simple convolutional layer and has limited feature extraction ability. While this can be improved by increasing the number of convolutional layers, this significantly increases the model parameters and lengthens the inference time. In Fig. 10 and Fig. 11, the Shi2024 and Tong2023 methods achieve relatively good compression performance due to their excellent attention mechanism design and reasonable residual convolution modules. But they perform poorly on the San Francisco dataset, suggesting that they are less robust. The other methods are mediocre in rate distortion performance, mainly because they lack a strong attention mechanism and excellent rate distortion optimization strategy. On the

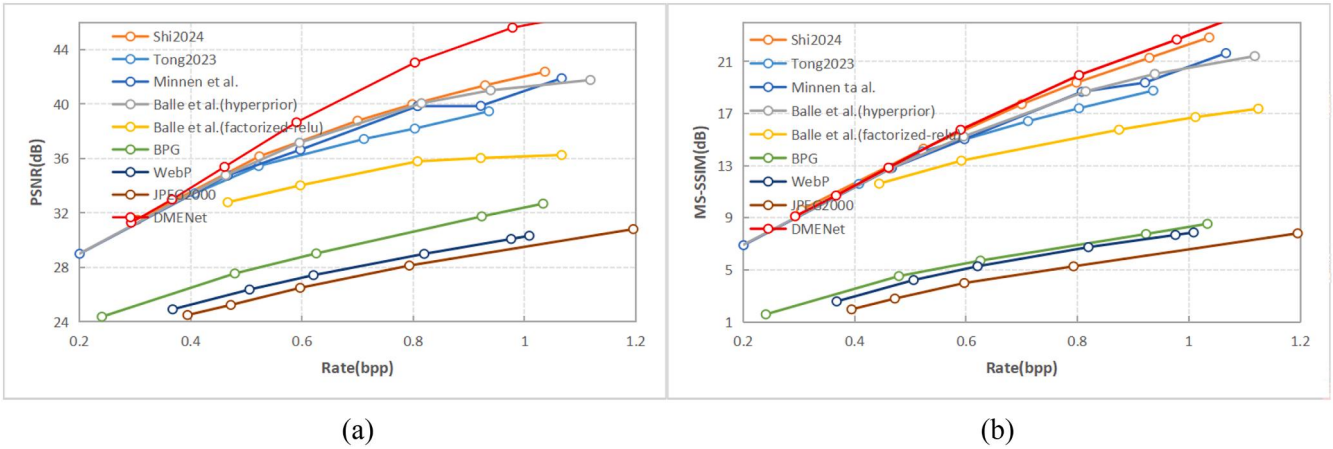


Fig. 9. Rate distortion curves on San Francisco. (a) PSNR (b) MS-SSIM.

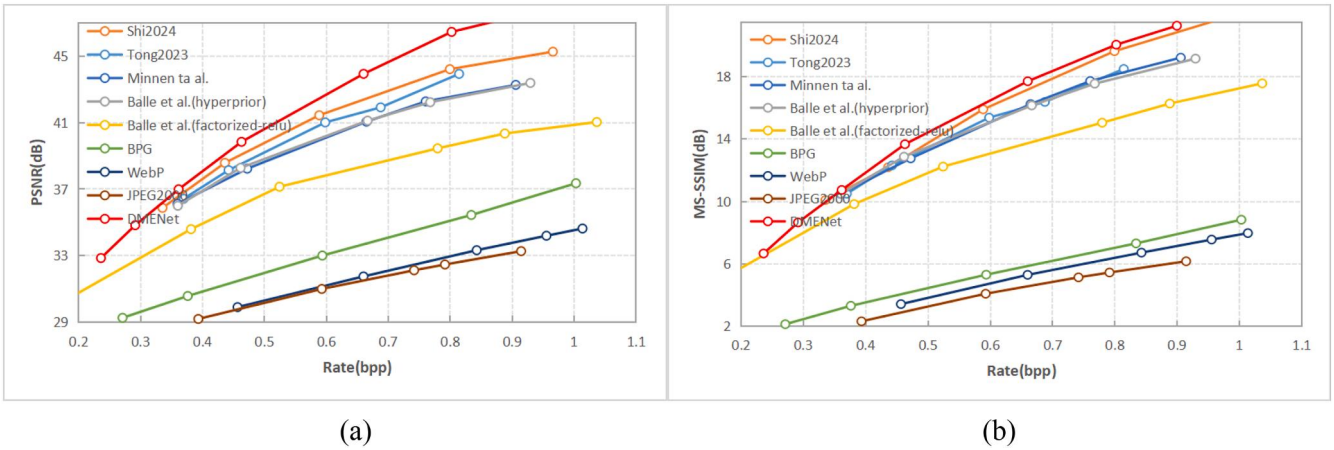


Fig. 10. Rate distortion curves on NWPU-RESISC45. (a) PSNR (b) MS-SSIM.

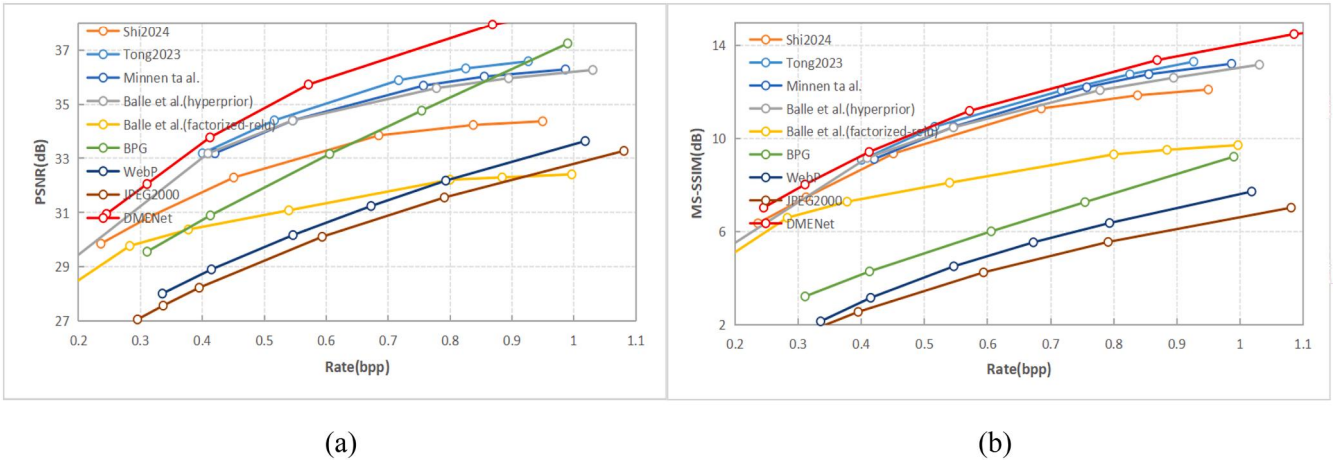


Fig. 11. Rate distortion curves on UC-Merced. (a) PSNR (b) MS-SSIM.

dataset San Francisco, specifically, at 0.46 bpp, DMENet achieves PSNR improvements of 1.8%, 7.2 %, 1.8%, 2.4%, and 1.5% compared to that of Balle et al. (hyperprior), Balle et al. (factorized-relu), Minnen et al., Tong2023, and Shi2024, respectively. In addition, DMENet achieves MS-SSIM improvements of 1.0%, 9.1%, 0.6%, 0.9%, and 0.5% compared to that of Balle et al. (hyperprior), Balle et al. (factorized-relu), Minnen et al., Tong2023, and Shi2024, respectively. This performance advantage becomes even greater at higher bpp. For example, at 1.1bpp, DMENet achieves PSNR improvements of 11.3%, 28.3%, 11.0%, 13.4%, and 7.9% compared to that of Balle et al. (hyperprior), Balle et al. (factorized-relu), Minnen et al., Tong2023, and Shi2024, respectively. In addition, DMENet achieves MS-SSIM improvements of 15.0%, 41.0 %, 13.9%, 20.2%, and 6.7% compared to that of Balle et al. (hyperprior), Balle et al. (factorized-relu), Minnen et al., Tong2023, and Shi2024,

respectively. The DMENet proposed in this paper achieves the best rate distortion performance on three datasets at the same time. This superior performance not only strongly proves the robustness of DMENet, but also clearly validates the effectiveness of MDEI, SDPM, LSM and the proposed rate distortion optimization strategy in DMENet. In particular, multi-dimensional edge information plays a significant role in improving the performance of the model at different bpp.

3.3. Visualization comparison experiment of reconstructed images

To further verify the effectiveness of DMENet, a visual comparison experiment was carried out in this paper. Fig. 12 shows the reconstructed images of each method on the dataset San Francisco at 0.25 bpp, along with their partial enlargements. Fig. 13 shows the reconstruction results of each method on the dataset UC-Merced at 0.28bpp. Taking Fig. 12 as an example, for the traditional image compression method based on codec, the zebra crossing of the BPG method retains more texture information than that of JPEG2000 and WebP. The zebra crossings in the JPEG2000 and WebP reconstruction areas have lost their clear edges and are blurred. The main reason is that the BPG method has a multi-channel coding technology, which has a stronger ability to reconstruct detailed features. The image compression and comparison method based on deep learning generally achieves better visualization results than traditional image compression methods, but it is still worse than DMENet. In Fig. 12, some artifacts and noise are prevalent in the reconstructed images of Minnen et al., Balle et al., and Balle et al. (factorized-relu). This results in a blurry image in some areas. The transitions between pixels in the reconstructed images of these three contrast methods are too coarse, which leads to color flattening and distortion. Finally, comparing Tong2023, Shi2024 and DMENet, the roof of the partially enlarged image in DMENet leaves more texture features and clearer edges of the object. The magnified area of the Tong 2023 and Shi 2024 reconstruction images is too smooth, and some detail features are lost. As a result, DMENet achieves the best visualization on the dataset San Francisco. In addition, in Fig. 13, DMENet also achieves the best visualization on the dataset UC-Merced. This also verifies that the proposed method has strong robustness. From the perspective of visualization, the above experiments fully prove that the proposed DMENet can reconstruct more complete and clearer images of structural features, and also prove that the new rate distortion optimization strategy plays a key role in aligning the structure.

In addition, the visualization results of the feature maps are provided to demonstrate the effectiveness of the MDEI in extracting structure features and edge features. Here, the three branches of MDEI (HA, VA, EEM) and the overall feature map are visualized. Since the input images are all three bands, each band of the feature map is visualized. As shown in Fig. 14, HA extracts the horizontal structural features, VA extracts the vertical structural features, and EEM extracts the edge features. Finally, the overall feature map provides high-quality structural and edge information, which strongly illustrates the effectiveness of MDEI.

3.4. Ablation experiments

In this paper, sufficient ablation experiments are carried out to verify the effectiveness of the proposed components such as MDEI, SDPM and LSM. Fig.s 15-17 are the results of ablation experiments on the dataset

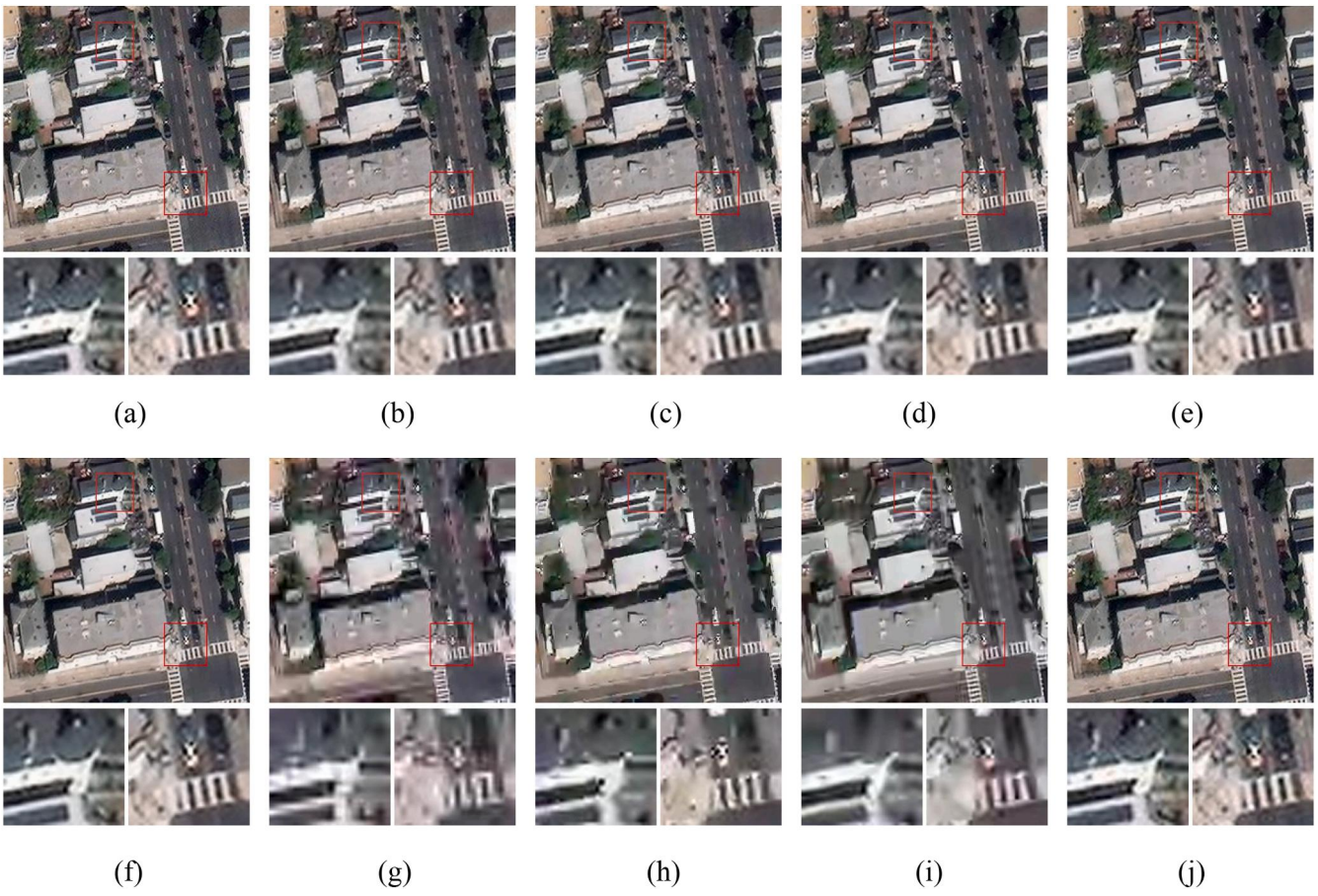


Fig. 12. Visual comparison of reconstructed images obtained by different methods on the dataset San Francisco. (a) Original (b) Minnen et al.(bpp:0.251;PSNR: 29.66;MS-SSIM: 7.55) (c) Balle et al.(hyperprior) (bpp:0.251;PSNR:29.74;MS-SSIM:7.77) (d) Balle et al.(factorized-relu) (bpp: 0.250 PSNR: 29.06MS-SSIM: 7.27) (e) Tong2023 (bpp:0.249;PSNR: 30.25;MS-SSIM: 8.08) (f) Shi2024 (bpp: 0.251;PSNR: 30.43;MS_SSIM: 8.27) (g) JPEG2000(bpp: 0.265; PSNR:22.79; MS-SSIM:1.25) (h) Webp(bpp: 0.38; PSNR:24.81; MS-SSIM:2.78) (i) BPG (bpp:0.244; PSNR:24.36; MS-SSIM:1.94) (j) DMENet (bpp: 0.250; PSNR:30.39; MS-SSIM:8.24).

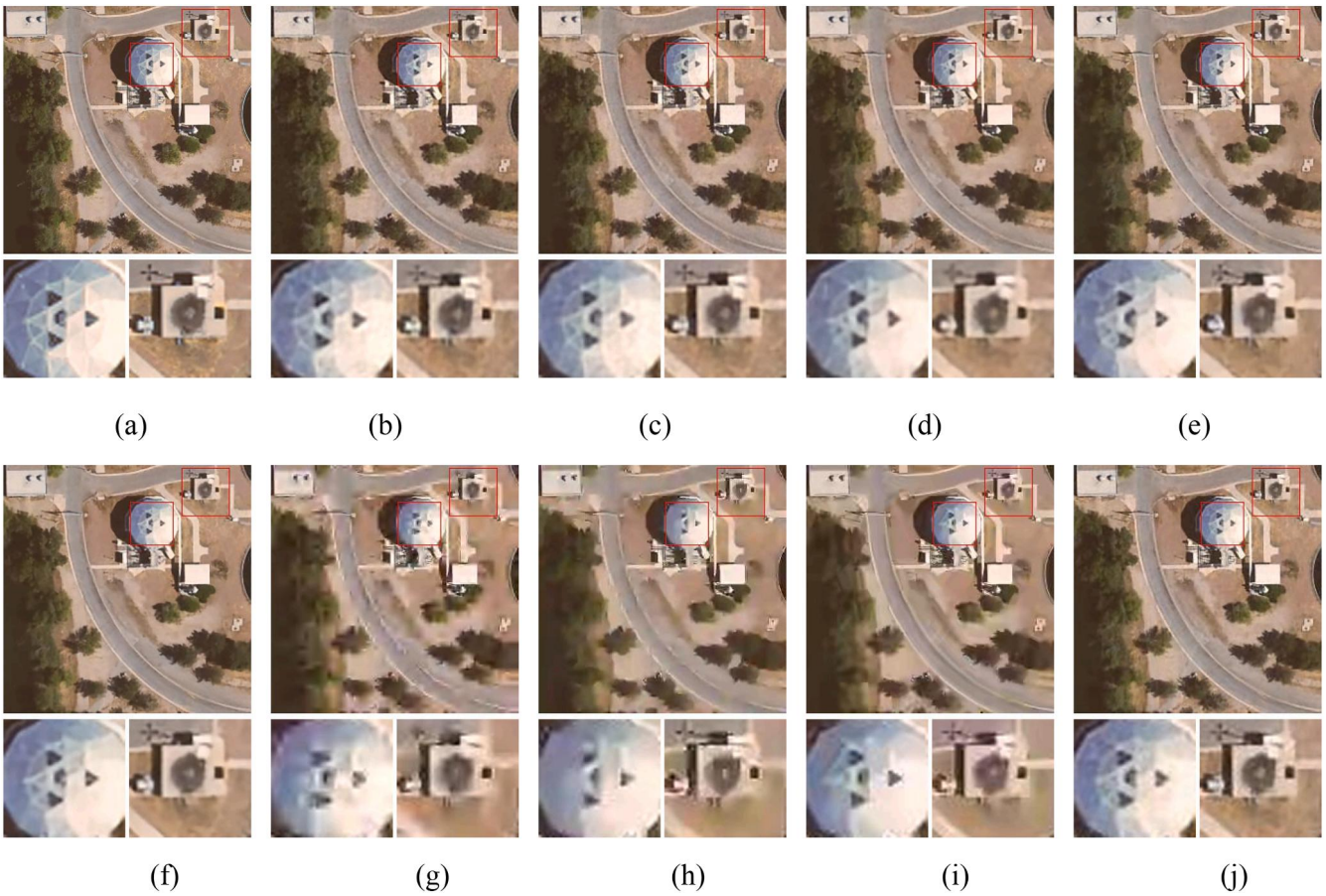


Fig. 13. Visual comparison of reconstructed images obtained by different methods on the dataset UC-Merced. (a) Original (b) Minnen et al.(bpp:0.280;PSNR:30.79;MS-SSIM:6.42) (c) Balle et al.(hyperprior) (bpp: 0.278;PSNR:30.90;MS-SSIM:6.81) (d) Balle et al.(factorized-relu) (bpp:0.281;PSNR:29.27;MS-SSIM:5.75) (e) Tong2023 (bpp:0.281;PSNR:29.74;MS-SSIM:5.07)(f) Shi2024(bpp:0.278;PSNR:30.43;MS-SSIM:6.83)(g)JPEG2000(bpp:0.268;PSNR:26.47;MS-SSIM:1.21)(h)Webp (bpp:0.260;PSNR:26.88;MS-SSIM:1.88)(i)BPG(bpp:0.271;PSNR:27.19;MS-SSIM:2.90)(j) DMENet (bpp:0.281;PSNR:31.55;MS-SSIM:7.49).

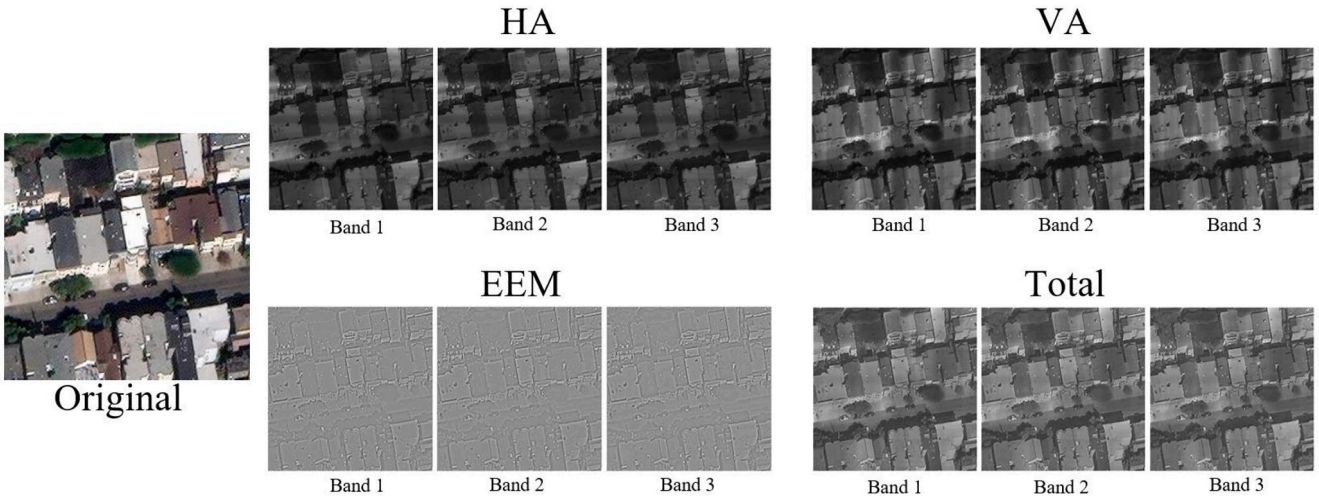


Fig. 14. Visualization of feature maps in MDEI

San Francisco, the dataset NWPU-RESISC45, and the dataset UC-Merced, respectively. The baseline represents the baseline network, and MDENet (MDEI), MDENet (SDPM), MDENet (LSM),MDENet (SDPM+ MDEI) ,MDENet (SDPM+ LSM) and MDENet (MDEI + LSM) represent the network after different modules are added. MDENet stands for complete network.. Fig.s 15-17 show that the compression performance of MDENet (MDEI), MDENet (SDPM), MDENet (LSM), MDENet (SDPM+ MDEI) ,MDENet (SDPM+ LSM) and MDENet (MDEI + LSM) exceeds that of baseline at different bit rates. This fully illustrates the effectiveness of aligning structural features between the original image and the reconstructed image. It also fully shows that the enhancement of irregular features and multi-scale features is conducive to the compression of remote sensing images. And the probability model with higher latent spatial representation capabilities is of great significance

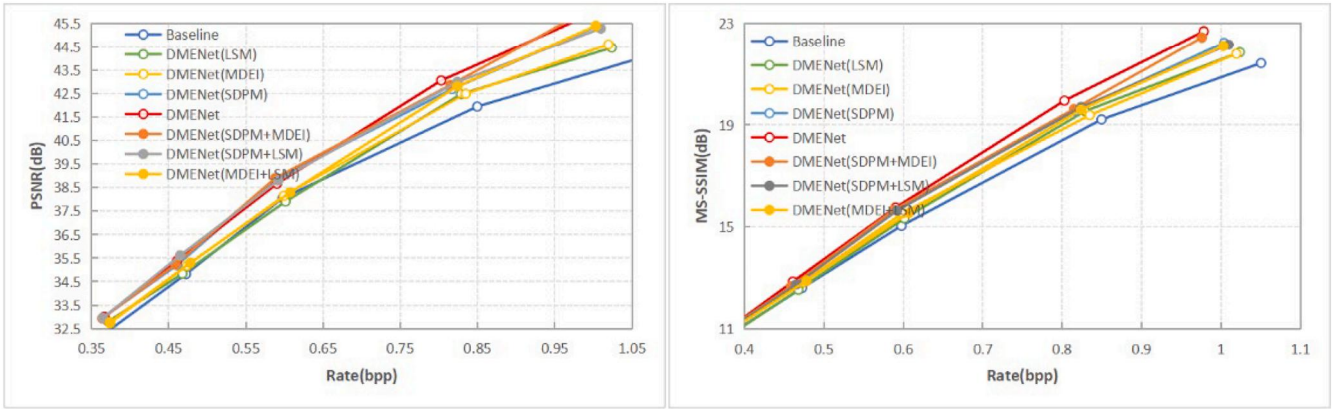


Fig. 15. Some ablation results of the proposed method on the San Francisco dataset. (a) PSNR, (b) MS-SSIM.

for image reconstruction. MDENet achieves the best rate distortion performance at the same bit rate. This shows that the network, module, and rate distortion optimization strategies can work together efficiently to achieve excellent compression performance.(Fig. 16).

3.5. Verification of the generalization of the proposed method

To verify the generalization of the proposed DMENet, the rate distortion performance was tested on a dataset of natural scenes. Here, a dataset (Natural scenes) is constructed, which includes five scenes: bathroom, bookstore, classroom, elevator, and kitchen. A total of 500 images are included, each with a resolution of 256×256 pixels. The dataset is divided into a training set, a validation set, and a test set at a ratio of 8:1:1. In addition, two image compression methods with excellent performance were selected for comparison. As shown in Fig. 18, DMENet still achieves significant performance advantages over PSNR and MS-SSIM. This fully demonstrates that DMENet is also suitable for the compression of natural images. The reason is that there are also a large number of structural features and edge features in the natural image, and the proposed DMENet strongly aligns these features. This further proves that the proposed DMENet method has good generalization ability.

3.6. Classification of remote sensing scene images

In this paper, the reconstructed images obtained by different compression methods are used for remote sensing scene classification, to verify the effectiveness of the proposed DMENet method from the perspective of application. The dataset selected is NWPU-RESISC45. The benchmark model for remote sensing scene classification is EMTCAL. To ensure the fairness of the experiment, the reconstructed images of different methods were obtained at a bit rate of 0.7 bpp. Fig. 19 shows the overall accuracy (OA) of the reconstructed images obtained by different methods when classifying the scenes of remote sensing images. The proposed DMENet obtained the highest OA value and achieved the best classification performance, which was more than 1.31% compared to Balle et al. (factorized-relu). In addition, it also has certain advantages over other methods. The classification accuracy of the proposed DMENet on each category is shown in Fig. 20. In particular, it has achieved 100% classification accuracy in many categories such as airplane, bridge, and freeway. This outstanding performance is mainly attributed to the two core advantages of DMENet: powerful structural feature extraction capability and edge feature extraction ability. These two capabilities enable DMENet to efficiently capture and reconstruct the rich edge structure information contained in objects such as bridges and highways, ensuring high-precision classification on these categories. Fig. 21 shows the confusion matrix of Minnen et al., Balle et al. (hyperprior), Balle et al. (factorized-relu), Tong2023, Shi2024, and DMENet for

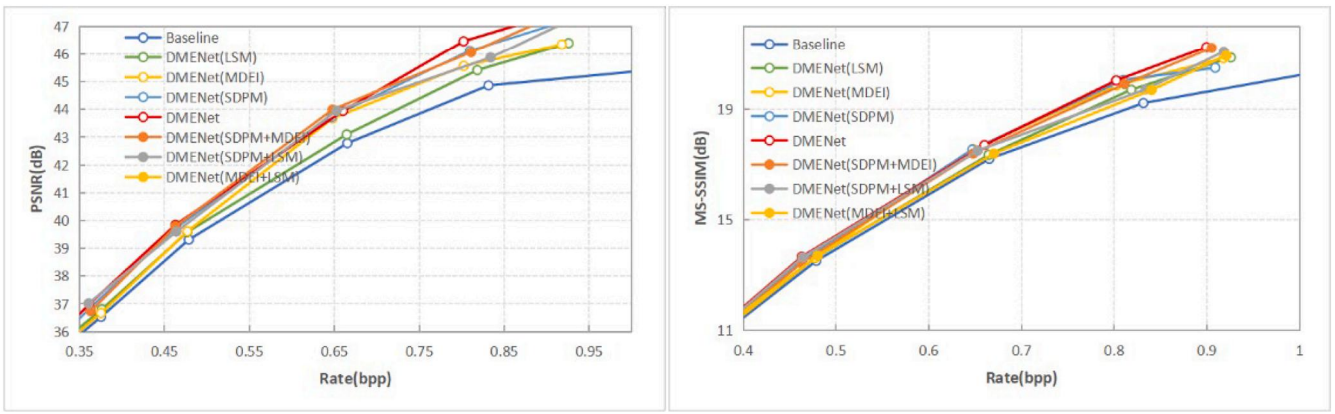


Fig. 16. Some ablation results of the proposed method on the NWPU-RESISC45 dataset. (a) PSNR, (b) MS-SSIM.

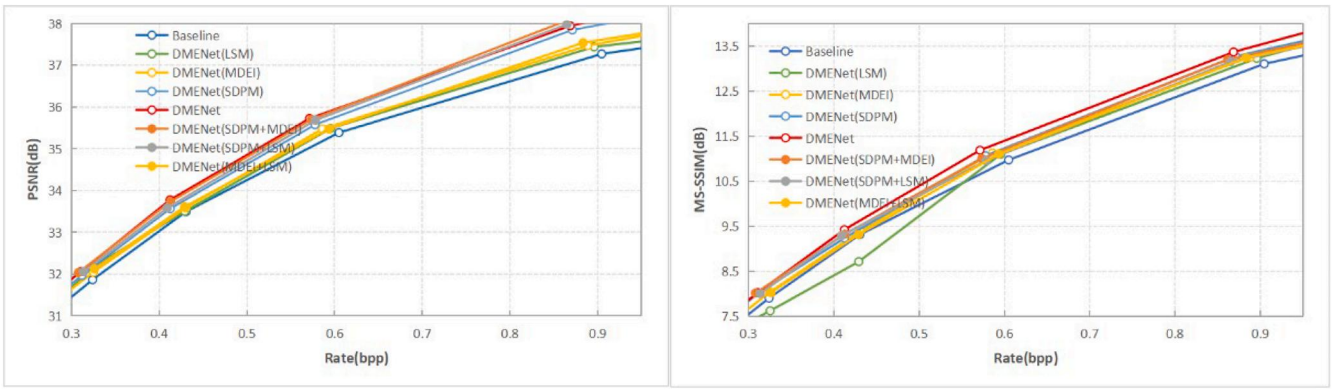


Fig. 17. Some ablation results of the proposed method on the UC-Merced dataset. (a) PSNR, (b) MS-SSIM.

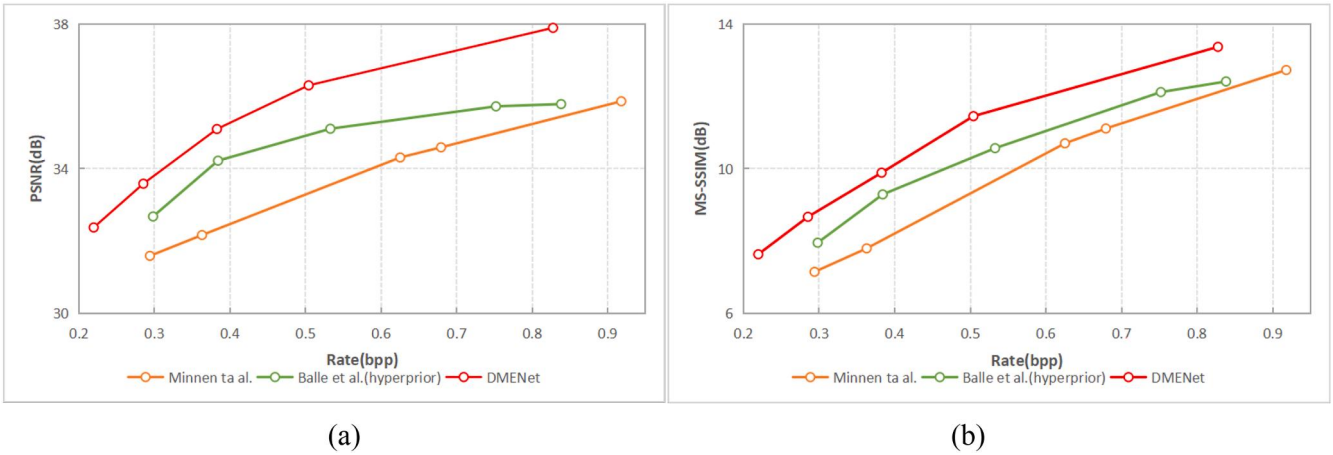


Fig. 18. Rate distortion curves on Natural scenes dataset. (a) PSNR (b) MS-SSIM.

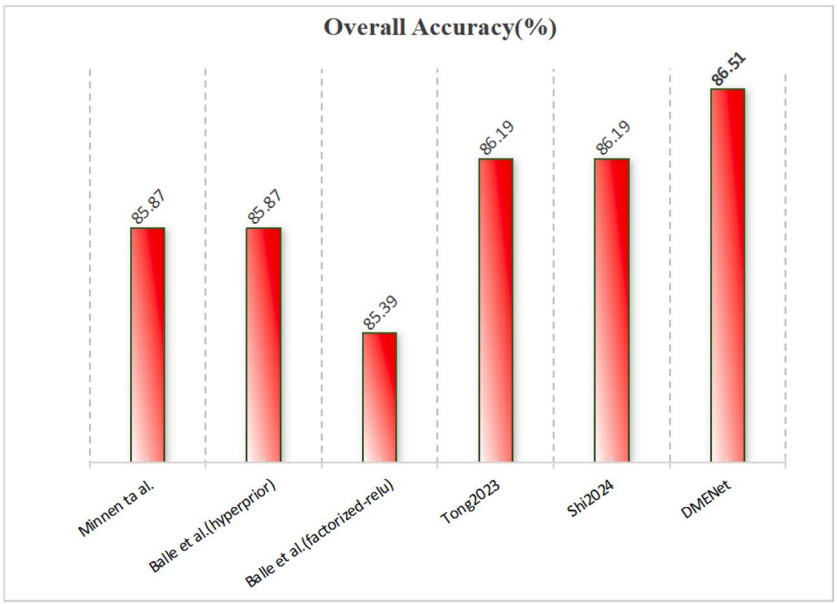


Fig. 19. OA of the reconstructed image obtained by different compression methods in remote sensing scene classification.

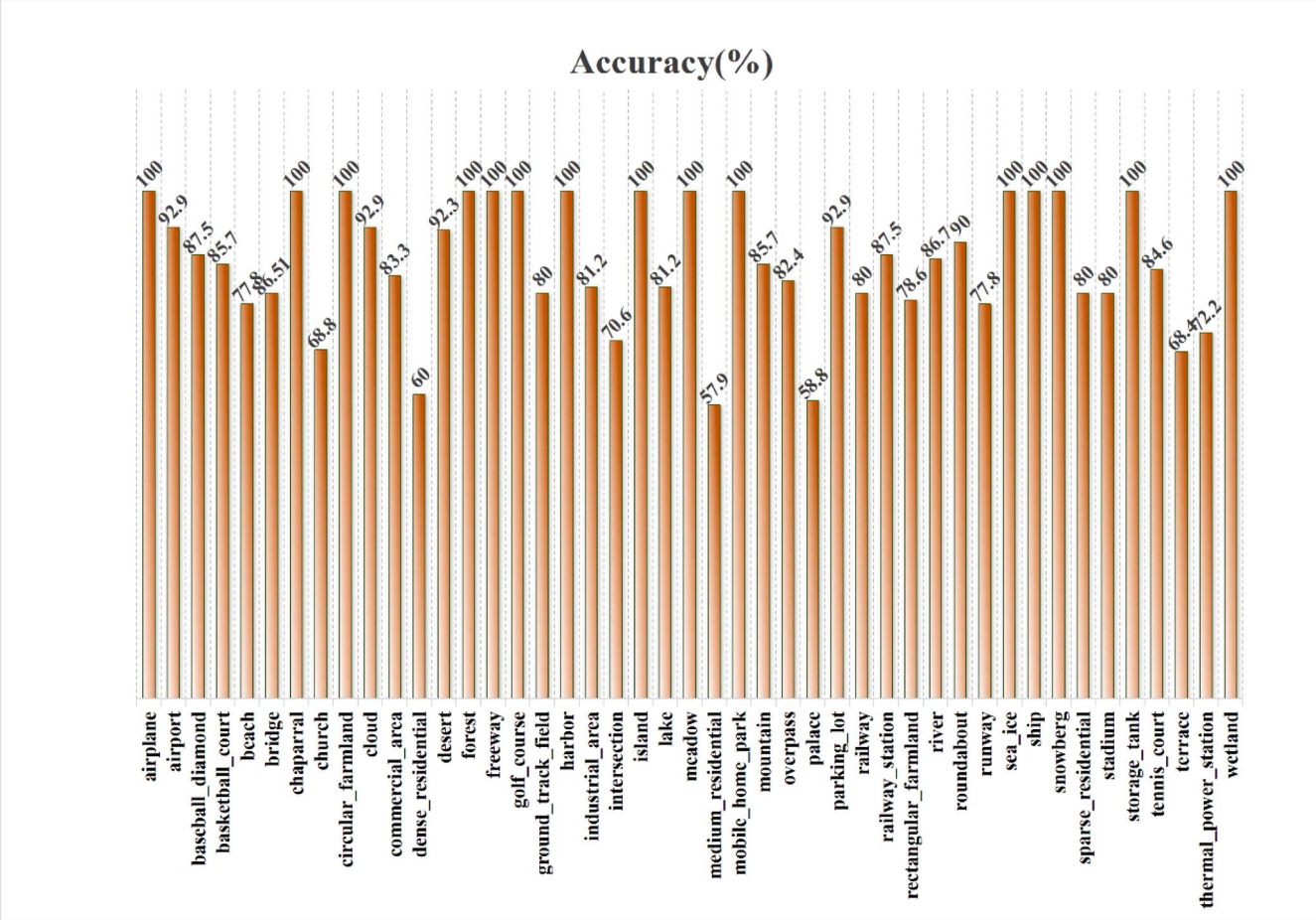


Fig. 20. Classification accuracy for each category (DMENet).

remote sensing image classification. In the confusion matrix in Fig. 21, DMENet has a better classification effect than other comparison methods in the classes of sparse_residential, cloud, lake, and river. This is mainly because these types of scenes have more structural features (such as the edges of lakes and rivers). DMENet enhances the enhancement of a variety of structural features, so that the high-quality structural features greatly improve the quality of the final discriminant features. This is the reason why the reconstructed image obtained by the proposed DMENet achieves the best performance in remote sensing scene classification.

3.7. Complexity analysis

In this section, the complexity analysis is carried out, and the evaluation indicators mainly include Parameter, FLOPs, GPU Memory, Compression time, and Reconstruction time. The input image size is $3\times256\times256$. Through comparison, it is found that the proposed DMENet has great advantages in complexity. By comparing the parameters, it can be found that DMENet has obtained the fewest parameters, which are only 42.40%, 53.48%, 95.32%, 19.23%, and 14.37% of the methods of Minnen et al., Balle et al. (hyperprior), Balle et al. (factorized-relu), Tong2023, Shi2024, etc., respectively. This fully illustrates the superiority of DMENet. By comparing FLOPs, it can be found that DMENet still achieves the fewest FLOPs, which are only 20.37%, 20.80%, 21.23%, 8.20%, and 4.18% of Minnen et al., Balle et al. (hyperprior), Balle et al. (factorized-relu), Tong2023, and Shi2024, respectively. Compared to GPU Memory, DMENet is in the middle of the pack. The reason for this is that some multi-branch structures are used in the network, resulting in more parallel computing, resulting in relatively large GPU memory. Comparing the compression time and the Reconstruction time, it can be found that DMENet's time consumption is in the third place, but at the same bit rate, DMENet's PSNR and MS-SSIM are significantly higher than those of other comparison methods. These experiments strongly demonstrate that DMENet can achieve very good rate distortion performance at low complexity (Table 3).

4. Conclusions

In this paper, DMENet is proposed to achieve high-fidelity remote sensing image compression while retaining higher-quality structural features. First, MDEI is designed to extract multi-dimensional structural features in images. These features are structurally aligned through loss, with the aim of restoring high-quality structural features. Secondly, SDPM is constructed to achieve dynamic extraction of irregular features and multi-scale features. Thirdly, LSM is proposed to solve the problem of deep feature loss caused by low information capacity in probability models. Finally, the whole network is guided by a rate distortion optimization strategy that pays more attention to multi-dimensional structural features. Compared with other methods, the proposed DMENet achieves the best rate distortion performance. Classification is used to evaluate the influence of the reconstructed images obtained by different compression methods on the application, and it is proved that the proposed DMENet method can provide the best classification performance, which indicates that the proposed method can retain the important information in remote sensing images more effectively. In future work, we will explore a variable rate remote sensing image compression method, which can greatly reduce the training cost of the model. In addition, we will further perform more detailed hierarchical processing on the compression and reconstruction process of remote sensing images to

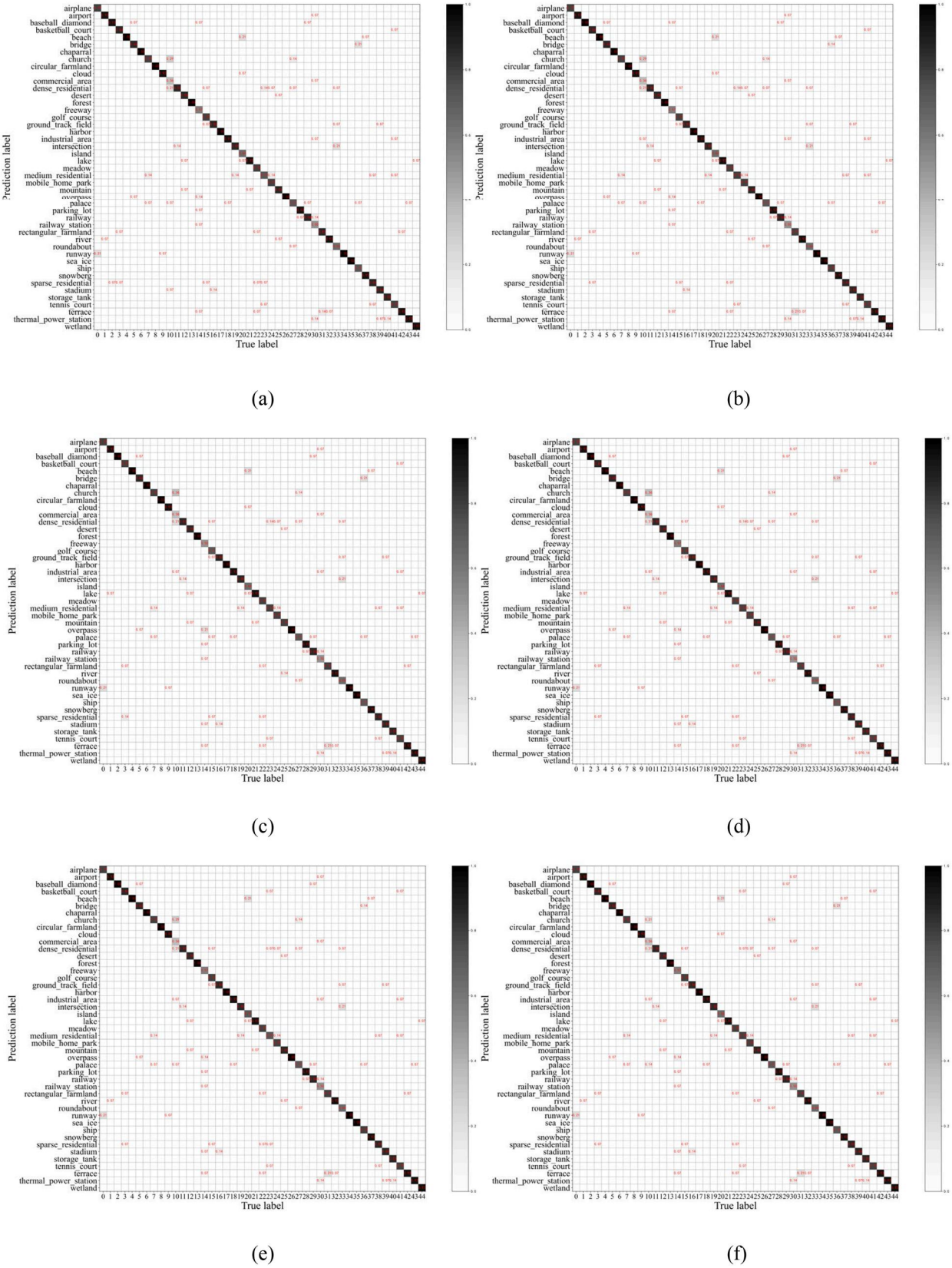


Fig. 21. Confusion matrix of the reconstructed image by different methods. (a), (b), (c), (d), (e), and (f) correspond to Minnen et al., Balle et al. (hyperprior), Balle et al. (factorized-relu), Tong2023, Shi2024 and DMENet, respectively.

Table 3
Complexity parameters for different compression methods.

	Minnen et al.	Balle et al.(hyperprior)	Balle et al. (factorized-relu)	Tong2023	Shi2024	DMENet
Parameter	12.05M	9.91M	5.56M	27.55M	36.87M	5.30M
FLOPs	27.04G	26.49G	25.95G	67.21G	131.91G	5.51G
GPU Memory	3.0GB	2.8GB	1.6GB	4.2GB	7.7GB	3.1GB
Compression time	0.593s	0.073s	0.032s	0.701s	1.156s	0.166s
Reconstruction time	1.078s	0.077s	0.039s	1.321s	1.921s	0.083s

reduce the information gap between latent representation features and specific tasks.

CRediT authorship contribution statement

Cuiping Shi: Methodology, Funding acquisition, Data curation. **Kaijie Shi:** Writing – original draft, Validation, Software, Formal analysis, Data curation, Conceptualization. **Zexin Zeng:** Writing – review & editing. **Fei Zhu:** Validation, Software.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Cuiping Shi reports financial support was provided by National Natural Science Foundation of China (42271409). Cuiping Shi reports financial support was provided by Fundamental Research Funds in Heilongjiang Provincial Universities of China under Grant 145409207. Cuiping Shi reports financial support was provided by the Science and Technology Plan Project of Huzhou under Grant 2024GZ36. Kaijie Shi reports financial support was provided by Qiqihar University Graduate Innovative Research Project under Grant QUZLTS_CX2023025. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (42271409), the Fundamental Research Funds in Heilongjiang Provincial Universities (145409207), in part by the Science and Technology Plan Project of Huzhou under Grant 2024GZ36, and in part by the Qiqihar University Graduate Innovative Research Project under Grant QUZLTS_CX2023025.

Data availability

Data will be made available on request.

References

[1] W Tang, F He, A K Bashir, et al., A remote sensing image rotation object detection approach for real-time environmental monitoring[J], Sustain. Energy Technol. Assess. 57 (2023) 103270.
[2] W Han, X Zhang, Y Wang, et al., A survey of machine learning and deep learning in remote sensing of geological environment: challenges, advances, and opportunities [J], ISPRS J. Photogr. Remote Sens. 202 (2023) 87–113.
[3] C. Tao, S. Fu, J. Qi, H. Li, Thick cloud removal in optical remote sensing images using a texture complexity guided self-paced learning method, IEEE Trans. Geosci. Remote Sens. 60 (2022) 5619612. Art. no.
[4] C Benitez, M Beland, S Esaian, et al., High-resolution remotely sensed data characterizes indices of avifaunal habitat on private residential lands in a global metropolis[J], Ecol. Indic. 160 (2024) 111900.
[5] Cuiping Shi, Xinlei Zhang, Liguo Wang, Zhan Jin, A lightweight convolution neural network based on joint features for Remote Sensing scene image classification, Int. J. Remote Sens. 44 (21) (2023) 6615–6641.
[6] Y Zhang, X Zheng, X. Lu, Remote sensing image retrieval by deep attention hashing with distance-adaptive ranking[J], IEEe J. Sel. Top. Appl. Earth. Obs. Remote Sens. 16 (2023) 4301–4311.
[7] J Núñez, O Fors, X Otazu, et al., A wavelet-based method for the determination of the relative resolution between remotely sensed images[J], IEEE Trans. Geosci. Remote Sens. 44 (9) (2006) 2539–2548.
[8] W Du, J Sun, Q. Ni, Fast and efficient rate control approach for JPEG2000[J], IEEE Trans. Consumer Electron. 50 (4) (2004) 1218–1221.
[9] G Ginesu, M Pintus, D D Giusto, Objective assessment of the WebP image coding algorithm[J], Signal process. Image Commun. 27 (8) (2012) 867–874.
[10] G K Wallace, The JPEG still picture compression standard[J], Communications of the ACM 34 (4) (1991) 30–44.
[11] JPEG2000 official software OpenJPEG,https://jpeg.org/jpeg2000/software.html.
[12] F Li, V Lukin, O Ieremeiev, et al., Quality control for the BPG lossy compression of three-channel remote sensing images[J], Remote Sens. (Basel) 14 (8) (2022) 1824.
[13] B Kovalenko, V Lukin, S Kryvenko, et al., BPG-Based Automatic Lossy Compression of Noisy Images with the Prediction of an Optimal Operation Existence and Its Parameters[J], Applied Sciences 12 (15) (2022) 7555.
[14] M Maldonado, J. WebP, a new web oriented image format, Universitat Oberta de Catalunya (2010) [J].
[15] D Báscones, C González, D. Mozos, Hyperspectral image compression using vector quantization, PCA and JPEG2000[J], Remote Sens. (Basel) 10 (6) (2018) 907.
[16] F Li, V Lukin, O Ieremeiev, et al., Quality control for the BPG lossy compression of three-channel remote sensing images[J], Remote Sens. (Basel) 14 (8) (2022) 1824.
[17] Y Hu, W Yang, Z Ma, et al., Learning end-to-end lossy image compression: a benchmark[J], IEEe Trans. Pattern. Anal. Mach. Intell. 44 (8) (2021) 4194–4211.
[18] F Aulí-Llinàs, M W Marcellin, J Serra-Sagrista, et al., Lossy-to-lossless 3D image coding through prior coefficient lookup tables[J], Inf. Sci. 239 (2013) 266–282.
[19] Z Wang, N M Nasrabadi, T S Huang, Spatial–spectral classification of hyperspectral images using discriminative dictionary designed by learning vector quantization [J], IEEE Trans. Geosci. Remote Sens. 52 (8) (2013) 4808–4822.
[20] R Pizzolante, B. Carpentieri, Multiband and lossless compression of hyperspectral images[J], Algorithms. 9 (1) (2016) 16.
[21] L Thornton, J Soraghan, R Kutil, et al., Unequally protected SPIHT video codec for low bit rate transmission over highly error-prone mobile channels[J], Signal Process. Image Commun. 17 (4) (2002) 327–335.
[22] S.E. Qian, Hyperspectral data compression using a fast vector quantization algorithm[J], IEEE Trans. Geosci. Remote Sens. 42 (8) (2004) 1791–1798.
[23] R La Grassa, C Re, G Cremonese, et al., Hyperspectral data compression using fully convolutional autoencoder[J], Remote Sens. (Basel) 14 (10) (2022) 2472.
[24] J Liu, F Yuan, C Xue, et al., An efficient and robust underwater image compression scheme based on autoencoder[J], IEEE J. Ocean. Eng. (2023).
[25] V Alves de Oliveira, M Chabert, T Oberlin, et al., Reduced-complexity end-to-end variational autoencoder for on board satellite image compression[J], Remote Sens. (Basel) 13 (3) (2021) 447.
[26] Q Xu, Y Xiang, Z Di, et al., Synthetic aperture radar image compression based on a variational autoencoder[J], IEEE Geosci. Remote Sens. Lett. 19 (2021) 1–5.
[27] Z Cheng, H Sun, M Takeuchi, et al., Learned image compression with discretized gaussian mixture likelihoods and attention modules[C], in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 7939–7948.
[28] Z Guo, Z Zhang, R Feng, et al., Causal contextual prediction for learned image compression[J], IEEE Trans. Circ. Syst. Video Technol. 32 (4) (2021) 2329–2341.
[29] T Chen, H Liu, Z Ma, et al., End-to-end learnt image compression via non-local attention optimization and improved context modeling[J], IEEE Transactions on Image Processing 30 (2021) 3179–3191.
[30] M Cao, W Dai, S Li, et al., End-to-end optimized image compression with deep Gaussian process regression[J], IEEE Trans. Circ. Syst. Video Technol. (2022).
[31] F Kong, T Cao, Y Li, et al., Multi-scale spatial-spectral attention network for multispectral image compression based on variational autoencoder[J], Signal. Processing. 198 (2022) 108589.
[32] C Fu, B Du, L. Zhang, Sar image compression based on multi-resblock and global context[J], IEEE Geosci. Remote Sens. Lett. 20 (2023) 1–5.
[33] L Zhang, X Hu, T Pan, et al., Global priors with anchored-stripe attention and multiscale convolution for remote sensing images compression[J], IEEe J. Sel. Top. Appl. Earth. Obs. Remote Sens. (2023).
[34] J Gao, Q Teng, X He, et al., Mixed entropy model enhanced residual attention network for remote sensing image compression[J], Neural Process. Lett. 55 (7) (2023) 10117–10129.
[35] S Xiang, Q Liang, L. Fang, Discrete wavelet transform-based Gaussian mixture model for remote sensing image compression[J], IEEE Trans. Geosci. Remote Sens. (2023).
[36] Y Guo, Y Chong, Y Ding, et al., Learned hyperspectral compression using a student's T hyperprior[J], Remote Sens. (Basel) 13 (21) (2021) 4390.
[37] M Zhao, R Yang, M Hu, et al., Deep learning-based technique for remote sensing image enhancement using multiscale feature fusion[J], Sensors 24 (2) (2024) 673.

[38] G Sumbul, J Xiang, B. Demir, Towards simultaneous image compression and indexing for scalable content-based retrieval in remote sensing[J], IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–12.
[39] W Ye, W Lei, W Zhang, et al., GFSCompNet: remote sensing image compression network based on global feature-assisted segmentation[J], Multimed. Tools. Appl. (2024) 1–25.
[40] S Xiang, Q Liang, P. Tang, Task-oriented compression framework for remote sensing satellite data transmission[J], IEEe Trans. Industr. Inform. (2023).
[41] M Liu, L Tang, L Fan, et al., CARNet: Context-aware residual learning for JPEG-LS compressed remote sensing image restoration[J], Remote Sens. (Basel) 14 (24) (2022) 6318.
[42] L Zhang, X Hu, T Pan, et al., Global priors with anchored-stripe attention and multiscale convolution for remote sensing images compression[J], IEEe J. Sel. Top. Appl. Earth. Obs. Remote Sens. (2023).
[43] H Wang, L Liao, J Xiao, et al., Uplink-Assist Downlink Remote Sensing Image Compression via Historical Referecing[J], IEEE Trans. Geosci. Remote Sens. (2023).
[44] C Fu, B. Du, Remote sensing image compression based on the multiple prior information[J], Remote Sens. (Basel) 15 (8) (2023) 2211.
[45] J Li, X. Hou, Object-fidelity remote sensing image compression with content-weighted bitrate allocation and patch-based local attention[J], IEEE Trans. Geosci. Remote Sens. (2024).
[46] C Shi, K Shi, F Zhu, et al., Multi-head global attention and spatial spectral information fusion for remote sensing image compression[J], IEEe J. Sel. Top. Appl. Earth. Obs. Remote Sens. (2024).
[47] P Han, B Zhao, X. Li, Edge-guided remote sensing image compression[J], IEEE Trans. Geosci. Remote Sens. (2023).
[48] K Cheng, Y Zou, Y Zhao, et al., A remote sensing satellite image compression method based on conditional generative adversarial network[C]//Image and Signal Processing for Remote Sensing XXIX, SPIE 12733 (2023) 322–331.
[49] S Xiang, Q. Liang, Remote sensing image compression based on high-frequency and low-frequency components[J], IEEE Trans. Geosci. Remote Sens. (2024).
[50] S Xiang, Q. Liang, Remote sensing image compression with long-range convolution and improved non-local attention model[J], Signal. Processing. 209 (2023) 109005.
[51] C Fu, B. Du, Remote sensing image compression based on the multiple prior information[J], Remote Sens. (Basel) 15 (8) (2023) 2211.
[52] https://resources.maxar.com/product-samples/analysis-ready-data-san-francisco-california.
[53] G Cheng, J Han, X. Lu, Remote sensing image scene classification: benchmark and state of the art[J], Proceedings of the IEEE 105 (10) (2017) 1865–1883.
[54] Y. Yang, S. Newsam, Bag-of-visual-words and spatial extensions for land-use classification, in: ACMSIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS), 2010.
[55] F. Bellard. Bpg image format. [Online]. Available:http://bellard.org/bpg/.
[56] D Minnen, J Ballé, G D Toderici, Joint autoregressive and hierarchical priors for learned image compression[J], Adv. Neural Inf. Process. Syst. (2018) 31.
[57] J Ballé, D Minnen, S Singh, et al., Variational image compression with a scale hyperprior[J], arXiv preprint arXiv:1802.01436 (2018).
[58] K Tong, Y Wu, Y Li, et al., QVRF: A Quantization-error-aware variable rate framework for learned image compression[J], arXiv preprint arXiv:2303.05744 (2023).
[59] X Tang, M Li, J Ma, et al., EMTCAL: Efficient multiscale transformer and cross-level attention learning for remote sensing scene classification[J], IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–15.