(Received 8 April 2025, Accepted 18 June 2025, Scheduled 22 July 2025)

Hyperspectral Image Denoising Based on Multiscale Spatial-Spectral Feature Fusion in Frequency Domain

Xiaozhen Ren[1], Jing Cui[2], Yi Hu[1], Xiaotian Zhang[1], and Yingying Niu[2,*]

[1] *School of Artificial Intelligence and Big Data, Henan University of Technology, Zhengzhou 450001, China*
[2] *School of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, China*

ABSTRACT: Hyperspectral images often suffer from various types of noise pollution during acquisition and processing, which can significantly affect their application. However, existing denoising methods have limitation in fully utilizing the spatial and spectral correlation of hyperspectral image. In order to take full advantage of the multiscale spatial features and global spectral correlation of hyperspectral image, a hyperspectral image denoising method based on multiscale spatial-spectral feature fusion in frequency domain is proposed in this paper. The proposed method utilizes the structural decomposition of multiscale wavelet transform to transfer the denoising of hyperspectral image to the frequency domain, not only minimizing information loss, but also decomposing noise into small scales, making it easier to remove in the frequency domain. Moreover, a cross-multiscale fusion attention is designed to improve the model performance by considering multiscale information and cross-space learning. A spectral position-aware self-attention module is proposed to more fully exploit the spectral correlation in hyperspectral image. A multiscale fusion of spatial-spectral feature module is introduced to merge the different spatial and spectral features, thereby enhancing the denoising performance of the model. The experimental results demonstrate that the proposed method outperforms mainstream denoising methods in terms of performance. In addition, it exhibits better visual quality in texture details and edge protection.

1. INTRODUCTION

Hyperspectral images (HSIs) have rich information in space and spectrum, which makes it widespread in the fields of material classification, change detection, semantic segmentation, and object detection [1–4]. However, during the collection, HSIs are susceptible to noise contamination from various issues such as temperature change, illumination inconsistency, atmosphere absorption, and sensor breakdown [5–7]. Noise is an inevitable factor that can severely impact the quality of acquired HSIs. Therefore, it is of utmost importance for effectively reducing noise in HSIs as it serves as a fundamental requirement for various remote sensing applications.

HSI denoising can be considered as an inverse problem. Its goal is to restore the original HSI from the observed image corrupted by noise. Traditional denoising methods in HSI rely on prior knowledge and mathematical formulation to effectively denoise the image. These methods encompass algorithms such as tensor decomposition [8, 9], local and non-local similarity [10–12], and sparse low-rank techniques [13–15]. These algorithms have seen continuous refinement and improvement in recent years. Traditional denoising methods are known for their interpretability, generative nature, and reduced reliance on training data. However, these approaches often involve solving optimization problems, which require time-consuming numerical iterations and parameter tuning to achieve satisfactory denoising outcomes. Furthermore, for complex scenarios, it is challenging to find an accurate model that can effectively associate the observed HSI with the desired noise-free HSI. The absence of a precise model often leads to denoising failure and limits the effectiveness of traditional denoising methods.

Deep learning has gained significant popularity in the last few years, particularly in the field of HSI denoising [16–18]. Convolutional neural network (CNN) based approaches have shown substantial improvement compared to traditional denoising techniques, marking a significant advancement in the field. The success and widespread adoption of CNN-based methods in HSI denoising can be attributed to their robust learning capacity and enhanced representation capability. These methods excel at capturing complex data spatial dependency by using the powerful capability of convolutional filters. Unlike traditional denoising methods, CNN-based approaches directly model the relationship between noisy and clean HSI through the learning of convolutional filters. Chang et al. [19] introduced an HSI denoising method, which employs a tensor to learn the filters in each layer, thereby extracting spatial information in the local receiver domain while maintaining the integrity of the spectral spatial structure. Nguyen et al. [20] provided an HSI denoising method that combines the sparse low-order prior with the deep image prior. The sparse low-order prior is obtained by singular value decomposition, while the deep prior is provided by a convolutional neural network to restore the image. Do and Vetterli [21] proposed the use of wavelet transform for noise removal, but the effectiveness of the wavelet transform is highly dependent on the choice of the selected wavelet basis function.

* Corresponding author: Yingying Niu (niuyy@haut.edu.cn).

If the selected wavelet basis function is too large, more details will be retained, but it is not conducive to the removal of low-frequency noise. However, if the selected wavelet basis function is too small, the effect of processing low-frequency noise is better, but the performance in preserving image details is poor, which is easy to lead to excessive image smoothing and loss of details. To sum up, experiments are needed to further validate the size of the wavelet basis function selection.

In addition, due to the particularity of HSI, it is a critical challenge to conserve the spectral-spatial structure of HSI during denoising [22–25]. Zhang et al. [22] proposed a spatial-spectral gradient network (SSGN) method, which uses spatial spectral gradient learning to better extract intrinsic and deep features of HSI. However, the network structure is complex and requires more computing resource and training time. Wang et al. [23] proposed a novel convolutional network using united octave and attention mechanism (UOANet). It mainly embeds negative residual mapping in the Unet architecture to extract features in frequency space and spectrum, but the network has large computational cost and long processing time. Pan et al. [25] designed a spatial spectral quasi-attention recurrent network (SQAD), whose core objective is to construct a spatial spectral quasi-recurrent recursive unit to maintain spatial and spectral information, but ignores non-local similarity across the bands. Yuan et al. [26] proposed a residual learning based denoising method for HSI, which considers both spatial and spectral information, and does not require manual adjustment of hyperparameters for different HSIs. However, the correlation between the spectra is not fully utilized, and the use of a small receptive field neglects the global spatial correlation.

Since there is correlation between different bands in HSI, transformer model can capture long-range dependency through self-attention mechanism [27, 28]. It enables effective interaction and integration of information between different bands. Therefore, transformer architectures are adopted for HSI denoising. Lai et al. [29] introduced a hybrid spectral denoising network incorporating guided self-attention mechanism. Li et al. [30] designed a multi-hierarchical cross transformer network for HSI denoising, which denoises in the spectral direction using long-range dependency between bands. Sun et al. [31] designed a multi-scale 3D-2D hybrid convolutional neural network based on a CNN and transformer for feature extraction. Hu et al. [32] used a combined architecture of convolutional neural network and transformer to extract global features and enhance local features.

Xiong et al. [33] used a combination of transformer and recurrent neural network (RNN) to perform recurrent computation across bands, thus allowing global spectral correlation. However, the transformer model is complex with more parameters and higher computational requirements, and the computational cost is higher when dealing with large-scale HSI. Meanwhile, although transformer has strong global correlation capturing capability, most of the existing work does not fully consider the close connection between spatial and spectral properties of hyperspectral images, which may limit its performance in specific scenes. In summary, existing denoising methods have limitation in effectively using the spatial and spectral correlation of HSI. Although these methods can achieve good results in specific noise cases, they lack universality and are difficult to address the challenge of mixed noise.

In summary, a multiscale spatial-spectral feature fusion network in frequency domain for HSI denoising is proposed in this paper, which fully explores multiscale spatial features and global spectral correlation of HSI. Firstly, a multi-feature decomposition discrete wavelet transform is performed on the observed noisy image to obtain different frequency subcomponents, where the low-frequency component contains most of the energy in the image, while the high-frequency component usually includes some important detailed information such as object edge and texture. Since the low-frequency component usually contains the overall structure and global features of the image, we only utilize simple $3 \times 3 \times 3$ convolution and rectified linear unit (ReLU) to remove noise contained in low-frequency, and this paper primarily focuses on denoising for the high-frequency part of the HSI. Ultimately, the complementarity between spatial and spectral features is enhanced by the multiscale fusion of spatial-spectral feature module to improve the denoising ability of the model. The main contributions of this paper are as follows.

(1) In order to take full advantage of the multiscale spatial features and global spectral correlation of HSI, we introduce an HSI denoising method based on multiscale spatial-spectral feature fusion in frequency domain. This method utilizes the structural decomposition of multi-scale wavelet transform to replace traditional down-sampling operation, minimizing the information loss and decomposing noise into small scales.

(2) A cross-multiscale fusion attention (C-MFA) is proposed by considering multi-scale information and cross-space learning. It can better focus on key details in high-frequency components and improve image denoising effect.

(3) To effectively utilize spectral correlation, a spectral position-aware self-attention module (SPA-SA) is designed, which could capture the relative position relationship between different spectra and enhance the model ability of long-range spectral dependency.

(4) A multiscale fusion of spatial-spectral feature module (Ms-FSS) is designed to facilitate the fusion of different spatial and spectral features. This module enhances the complementarity between spatial and spectral features, thereby improving the denoising performance.

2. PROPOSED DENOISING METHOD

In this section, a novel multiscale spatial-spectral feature fusion network MSF-Net in frequency domain is proposed for HSI denoising. As shown in Fig. 1, the input noisy image is first decomposed into low-frequency and high-frequency components by Haar wavelet transform. Here, we use wavelet decomposition instead of down-sampling in traditional network to obtain multi-scale features. Due to the orthogonality of the Haar wavelet transform, no information from the original image is lost during this operation. Meanwhile, the time-frequency domain gradual segmentation property of wavelet transform helps to preserve image details and texture information. The introduction of this structure enables the network to have better

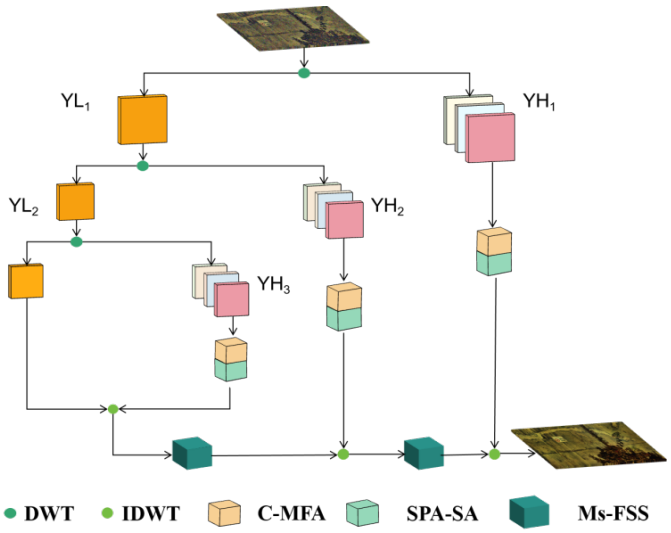


FIGURE 1. The overall network structure of MSF-Net.

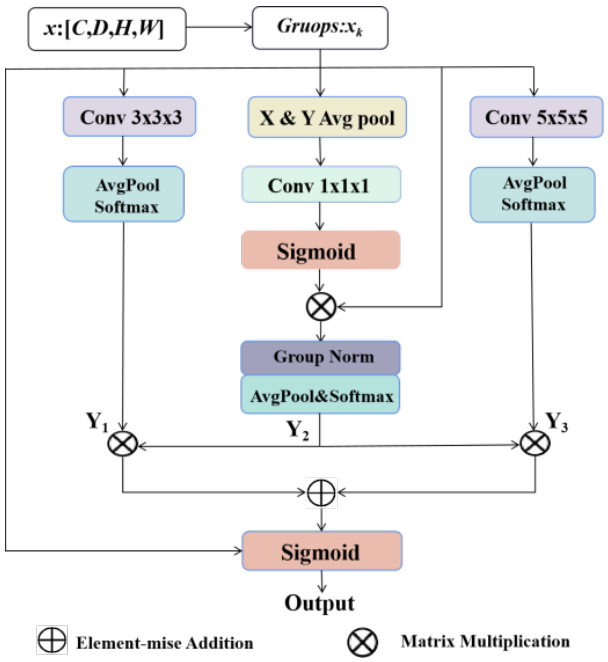


FIGURE 2. The structure of cross-multiscale fusion attention module.

performance and characterization on the task of HSI denoising. Furthermore, the noise can also be decomposed into small scales, making them easier to remove in frequency domain.

Since the low-frequency components usually contain the overall structure and basic features of the image, they are denoised by simple $3 \times 3 \times 3$ convolution and rectified linear unit (ReLU) convolution blocks. The different scales of high-frequency components are fed into the C-MFA module to preserve details and texture information while suppressing the influence of noise. Then, the enhanced features of the high-frequency components are fed into the SPA-SA module, which exploits the spectral correlation in HSI to improve denoising performance. Later, denoised HSI is obtained by performing inverse wavelet transform, and an Ms-FSS module is added after each level of inverse wavelet transform to effectively fuse different spatial and spectral features of HSI while reducing artifacts after denoising.

2.1. Cross-Multiscale Fusion Attention Module (C-MFA)

The input hyperspectral image is first subjected to the wavelet transform because it not only provides a powerful multi-resolution analysis capability to decompose the signal into components at different scales, but also efficiently handles noise in both spatial and spectral dimensions while retaining important structural information. After wavelet transform and convolution, the HSI is decomposed into low-frequency and high-frequency components. The high-frequency component contains some key detailed information that is essential for preserving the detailed features of the image. In contrast, the low-frequency component covers more of the overall structure and global features of the image. Therefore, when processing HSI, applying the attention mechanism to low-frequency component may have less impact on the overall feature extraction results. We only apply the attention mechanism for high-frequency component. In addition, traditional methods often deal with single-scale features, making it difficult to fully utilize the rich multi-scale information in HSI. Furthermore, they usually lack effective fusion mechanisms in the feature extraction process, which affects the denoising results. Therefore, a cross-multiscale fusion attention (C-MFA) is proposed in this paper, as shown in Fig. 2. By introducing this module, the multi-scale information in HSI can be utilized more comprehensively, and cross-space contextual information can also be introduced to improve the perception and representation of detailed features. The proposed C-MFA module has the capacity to better focus on the key detailed information in the high-frequency component, improving the denoising result and visual quality of the image.

Specifically, the C-MFA module uses a network structure that includes three parallel branches. For the input feature map $x \in R^{C\times D\times H\times W}$, C, D, H, and W denote the number of channels, bands, height and width of the feature map, respectively. We first divide x into K sub-features by channel dimension C, where $x = [x_0, x_1, \ldots, x_k, \ldots, x_{K-1}]$, $x_k \in R^{C/K\times D\times H\times W}$. This feature grouping operation could effectively enhance feature learning. Each sub-feature x_k is then input into three parallel branches. In the first branch, a $3 \times 3 \times 3$ convolutional kernel and an average pooling operation combined with a softmax activation function is employed to extract and weight the spatial features of the HSI. This operation is not only effective in capturing local details and enhancing the focus on specific regions, but also lighter on the input data and better preserving the detail information in the original data. The second branch is performed by applying pooling operations in the H and W directions, respectively, so that the dependency among all channels can be captured. Subsequently, we normalize and weight the features by applying group normalization and average pooling operations, combined with a softmax activation function. This step aims to extract and weight the global features of the HSI so as to better understand the structure and content of the whole image. In addition, the global perception is enhanced by averaging pooling in the vertical and horizon-

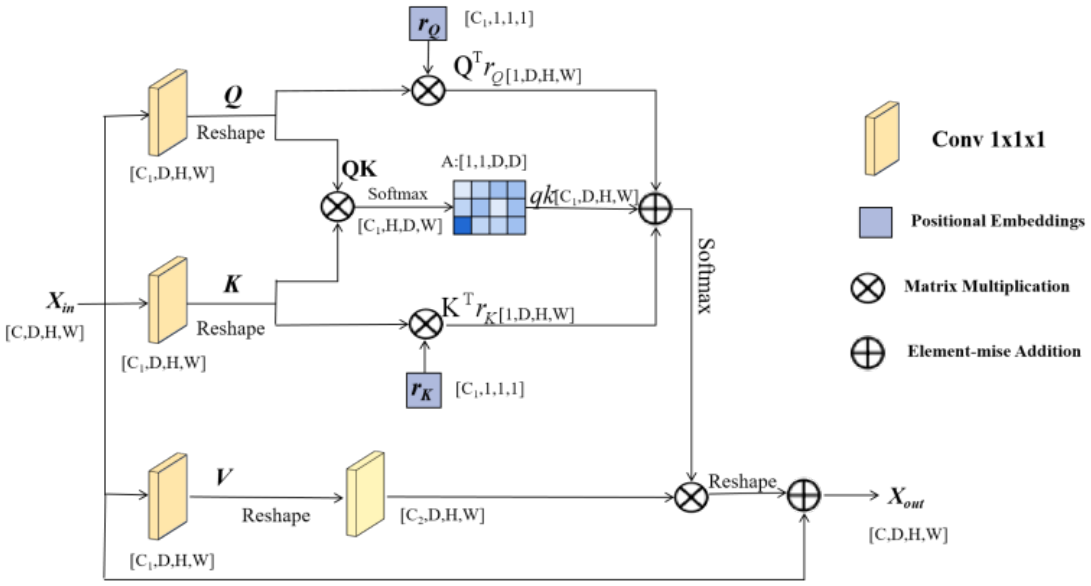


FIGURE 3. The structure of spectral position-aware adaptive attention module.

tal directions, respectively. This approach helps to capture the dependencies among all channels, thus enhancing the model's understanding of the entire image structure. The third branch employs a $5\times5\times5$ convolutional kernel and an average pooling operation to extract a wider range of local features from the HSI. This helps to capture a wider range of contextual information in the image, which improves the perception and representation of detailed features.

Next, the C-MFA module fuses the output feature maps of the three branches by means of cross-space learning. Each branch is responsible for information extraction at a specific scale, which is eventually integrated through a fusion layer. This multi-scale fusion strategy allows the network to adaptively select the most relevant feature representations based on the characteristics of the input data, thus improving the denoising effect. Furthermore, in addition to standard three-dimensional convolutional operation, the processing specifically for horizontal and vertical directions can help the model better understand and remove the noise propagating along specific directions. This approach enhances the sensitivity of the model to directional noise, making it effective in removing not only randomly distributed noise, but also those disturbances with specific directionality. Such a feature fusion strategy can effectively enhance the model ability to perceive subtle changes in a complex scene, capture pixel-level relationships of high-frequency features, and further improve the discriminative power of feature representations. The output feature map for C-MFA can be expressed by

$$Y_1 = \text{softmax}(\text{avg}(\text{conv}_{3\times3\times3}(x_k))) \quad (1)$$

$$Y_2 = \text{softmax}(\text{avg}(\text{GN}(x_k \times \text{sigmoid}(\text{conv}_{1\times1\times1}(\text{avg}(x_k)))))) \quad (2)$$

$$Y_3 = \text{softmax}(\text{avg}(\text{conv}_{5\times5\times5}(x_k))) \quad (3)$$

$$\text{Output}_{\text{C-MFA}} = \text{sigmoid}((Y_1 \times Y_2 + Y_2 \times Y_3)) \times x_k \quad (4)$$

where x_k is the k-th sub-feature, and avg and GN denote average pooling and group normalization operations, respectively. $\text{Conv}_{3\times3\times3}$, $\text{Conv}_{1\times1\times1}$, and $\text{Conv}_{5\times5\times5}$ represent $3\times3\times3$ convolution, $1\times1\times1$ convolution, and $5\times5\times5$ convolution, respectively. Y_1, Y_2, and Y_3 denote the outputs of the three branches. Overall, the C-MFA module improves denoising in several ways. The first point is multi-scale feature extraction, which utilizes convolutional kernels of different sizes to capture information from local to global. The second point is to extract global features and weight these features through average pooling and group normalization operations. The third point is feature fusion strategy. Fusing feature maps with different scales provide richer and more detailed feature representation, and retain the key details of the image while removing noise.

2.2. Spectral Position-Aware Self-Attention Module (SPA-SA)

The spatial attention C-MFA can improve the model performance by considering multi-scale information and cross-space learning, but it does not fully utilize spectral information of the HSI. Although there has been works that utilize attention mechanisms to extract feature from the spectral dimension of HSI [34, 35], they suffer from the drawbacks such as a large number of parameters and high demand for training data. To effectively utilize spectral correlation and improve denoising performance, a spectral position-aware self-attention (SPA-SA) module is proposed, as shown in Fig. 3. By applying the self-attention mechanism along the spectral dimension rather than the spatial or channel dimension, we are able to more fully exploit the spectral correlation in HSI and enhance the model ability of long-range spectral dependency. Furthermore, the relative position encoding is introduced to better capture the inter-spectral relationships in HSI, which significantly improves the denoising performance of the model.

In SPA-SA module, the input feature map $x \in R^{C\times D\times H\times W}$ first passes through three 3D convolution kernels of size $1\times1\times$

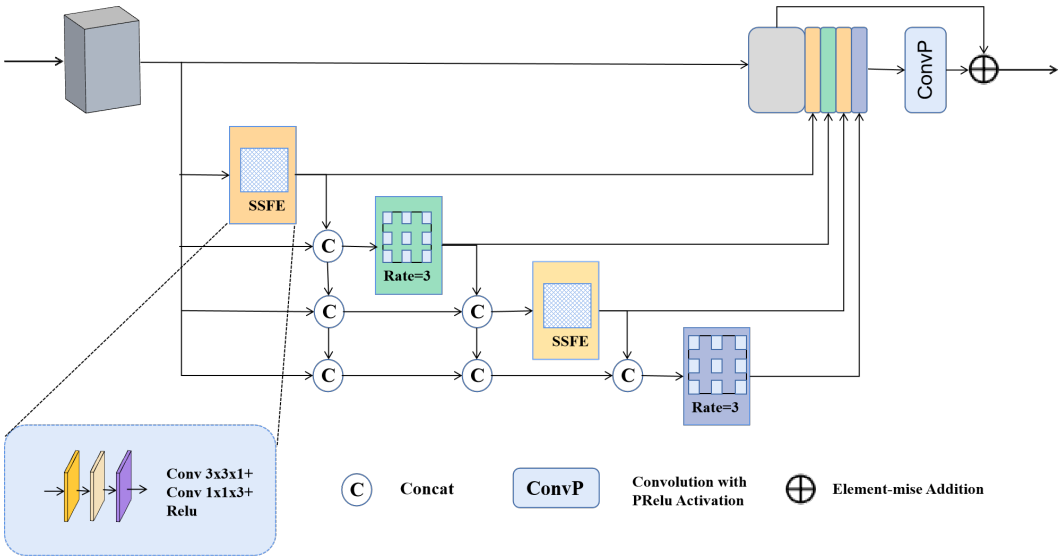


FIGURE 4. The structure of multiscale fusion of spatial spectrum features module.

1 to obtain $Q \in R^{C1\times D\times H\times W}$, $K \in R^{C1\times D\times H\times W}$, and $V \in R^{C1\times D\times H\times W}$, where C_1 is the number of channels after the convolution kernel. They are then subjected to reshape operation for subsequent operation and computation. In order to better handle the correlation and relative position information of different spectra in HSI, the relative position coding information $r_Q \in R^{C1\times 1\times 1\times 1}$ and $r_K \in R^{C1\times 1\times 1\times 1}$ are introduced in SPA-SA module. r_Q and r_K are multiplied with Q and K to obtain $Q^{\mathrm{T}}r_Q \in R^{1\times D\times H\times W}$ and $K^{\mathrm{T}}r_K \in R^{1\times D\times H\times W}$, respectively. Therefore, the position coding is effectively integrated into the attention mechanism. Relative position coding is introduced to allow the model to more effectively understand and utilize the dependency and correlation between different spectral bands in hyperspectral images, especially when dealing with remote spectral dependency. This improvement is helpful for enhancing the overall performance of the model and achieving better results in denoising tasks. Next, QK are normalized using softmax to obtain the attention graph A. A multiplied by QK gives qk. qk is then concatenated with $Q^{\mathrm{T}}r_Q$ and $K^{\mathrm{T}}r_K$ in the channel dimension. For V, we only perform the reshaping operation to ensure that the multiplication operation can be performed correctly in the subsequent computation.

In addition, to stabilize the training process and optimize feature extraction, residual connection is introduced. The original input feature is added with the feature encoded by relative position, so that the module can better retain the original information during the learning process and avoid the problem of gradient disappearance or gradient explosion. This could effectively improve the training stability and convergence speed of the model. In general, it improves the model's ability to understand and process complex spectral information by efficiently utilizing the spectral properties of hyperspectral images. Specifically, it is achieved by applying a self-attention mechanism along the spectral dimension to capture and utilize the dependency between different bands. Then, the relative position coding is introduced to enhance the ability to model remote spectral dependency and to better understand the relative relationship between spectra. Finally, it ensures that the denoising process does not lose critical spectral details and structural features.

2.3. Multiscale Fusion of Spatial-Spectral Feature Module (Ms-FSS)

Due to the particularity of HSI, it is a critical challenge to maintain the spectral-spatial structure of HSI during denoising. The existing deep learning based denoising methods for HSI may lack effective feature fusion mechanism, resulting in insufficient extracted feature information and affecting the denoising effect. To address this challenge, we design the multiscale fusion of a spatial-spectral feature module (Ms-FSS), which aims at effectively fusing different spatial and spectral features of HSI while reducing artifacts after denoising.

The Ms-FSS module is shown in Fig. 4. After processing the input feature map with inverse discrete wavelet transform (IDWT), it is first input into the network and processed by the spatial and spectral feature extraction module (SSFE). The SSFE consists of convolutional blocks of size $3\times3\times1$, $1\times1\times3$, and ReLU activation function. The $3 \times 3 \times 1$ convolutional block is mainly concerned with extracting the spatial information of the input feature, while the $1 \times 1 \times 3$ convolutional block is utilized to extract the spectral information of the input feature. With this design, the SSFE module better captures and utilizes the spatial and spectral information of hyperspectral images. It could retain more detailed information and enhance the overall performance and adaptability of the model. Finally, non-linearization is performed by ReLU activation function in the SSFE module to accelerate the model training.

After the feature extraction by the SSFE module, the output of SSFE module and the original input feature are used as inputs for the next branch, and the receptive field is enlarged by dilated convolution with a dilation rate of 3 to capture multiscale contextual information. The choice of dilation rate 3

is based on experimentally validated results. The results show that this setting maximizes the receptive field of the model without over-sparse the input signal. The main purpose of introducing this module is to better capture multi-scale contextual information in hyperspectral images without increasing the computational burden. Next, the original input feature, the feature extracted from the previous SSFE module, and the feature extracted from the dilated convolution with a dilation rate of 3 are concatenated together and passed as input to the next SSFE module for feature extraction. This step fuses features from different layers to obtain richer and more diverse information, thereby better capturing key features in the HSIs. In addition, this structure also promotes information exchange and interaction between features, which helps the model to better learn and understand the structural and semantic information in the HSIs, thereby improving the denoising performance. Overall, the module specifically improves denoising performance in several ways. Firstly, convolution kernels of different sizes are utilized to capture information from local to global, and these features are effectively fused. Secondly, the receptive field is expanded by dilated convolution to capture long-range dependency while retaining more detailed information. Finally, SSFE module is introduced to jointly process spatial and spectral information to enhance the feature representation and ensure that key spectral details and structural features are retained in the denoising process.

3. EXPERIMENTS AND ANALYSIS

In this section, various denoising experiments were performed on both simulated and real datasets. The denoising results were quantitatively analyzed using evaluation metrics, and typical denoised images were provided for subjective visual assessment. In addition, ablation studies were conducted to investigate the effectiveness of individual modules in the proposed model.

3.1. Datasets

Two datasets of ICVL [36] and Urban [37] were utilized to evaluate the denoising performance of the proposed MSF-Net method.

The ICVL dataset is composed of 201 images with a spatial resolution of 1392×1300 and spectral bands ranging from 400 to 700 nm. 100 of these HSIs are randomly selected as the training set, 50 HSIs as the test set, and the rest are used for validation to assure the authenticity of the experiment. In the experimental procedure, the images in the training set were uniformly cropped to 1024×1024 and normalized to $[0, 1]$. In addition, to expand the training set, the HSIs in the training set are processed into multiple overlapping image patches by cropping. The spatial resolution of each image patch is 64×64, and the spectral resolution is kept constant. At the same time, rotation and other operations were also used to further enhance the training set. Moreover, the hyperspectral images in the test set were cropped into 512×512 image patches for better visual effects.

To further demonstrate the generalization ability of our model, a real hyperspectral dataset named Urban was also tested, which consists of hyperspectral image with 307×307 pixels and 210 bands.

3.2. Experimental Details

The MSF-Net was implemented on the PyTorch platform and the network was trained by NVIDIA GeForce RTX 3060. The network parameters were initialized by Kaiming, and the network was optimized using the Adam optimizer. The number of epochs was set to 80. The initial learning rate was set to 0.001 and the batch size set to 8.

Furthermore, peak signal-to-noise-ratio (PSNR), structural similarity (SSIM), and spectral angular mapping (SAM) are used as the corresponding evaluation metrics to assess the performance.

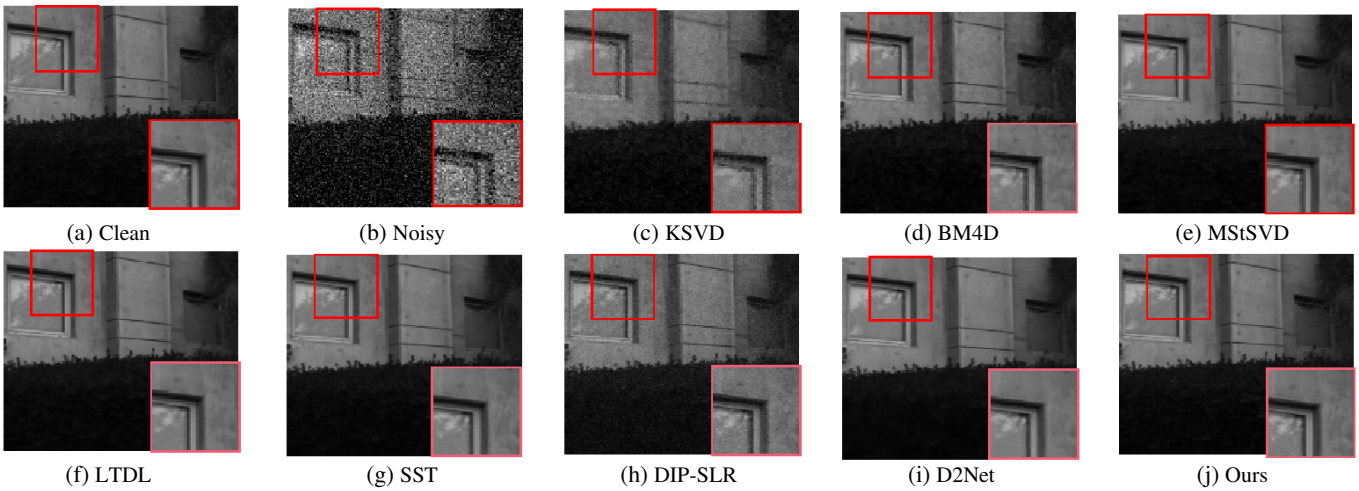
3.3. Evaluation on the ICVL Dataset

3.3.1. Results on the Hyperspectral Image with Gaussian Noise

In this experiment, the noisy observations were obtained by adding i.i.d Gaussian noise with different variances to the hyperspectral image. To evaluate the performance of the proposed MSF-Net in the task of hyperspectral image denoising, the comparative analysis was conducted against several existing denoising methods, including KSVD [38], BM4D [12], MStSVD [14], LTDL [15], SST [30], DIP-SLR [20], and D2Net [23]. For most methods, the study replicated their implementation using publicly available source code, along with the provided parameters or training models from the respective authors. However, for the SST method, we re-implemented it within a PyTorch based framework based on available source code and model parameters, and retrained it using synthetic data.

The proposed MSF-Net method is quantitatively compared with other methods on the ICVL dataset, and the results are shown in Table 1. The best results are shown in bold, while the second-best results are indicated by an underline. Among the comparison methods, the first four methods are based on traditional method, while the last four methods rely on deep learning. From Table 1, it is obvious that in the comparison of performance indexes, the denoising methods based on deep learning are significantly superior to that of the traditional methods. MSF-Net method can outperform all the other methods when the noise variance σ is 30. Even the variance σ increases to 70, MSF-Net can still robustly remove noise and obtain the highest PSNR and SSIM. Especially in blind noise scenarios, i.e., the HSI is contaminated by Gaussian noise with randomly distributed variance between 10 and 70, and the noise distribution is different in different bands. The proposed MSF-Net still achieves the best performance in all the metrics. These results show that MSF-Net method can yield good results under low noise condition, demonstrating its robustness and superior denoising ability.

In order to better evaluate the performance of the proposed MSF-Net method in HIS denoising, we show the denoising results of various algorithms on noisy HSI under different noise conditions. Fig. 5 presents the denoising results under i.i.d



(a) Clean (b) Noisy (c) KSVD (d) BM4D (e) MStSVD

(f) LTDL (g) SST (h) DIP-SLR (i) D2Net (j) Ours

FIGURE 5. The denoising results on the ICVL dataset under Gaussian noise with variance $\sigma = 30$.

TABLE 1. Comparison of denoising performance for Gaussian noise with different variances on ICVL dataset.

σ	Index	Noisy	KSVD	BM4D	MStSVD	LTDL	SST	DIP-SLR	D2Net	Ours
30	PSNR	18.58	29.65	36.41	38.35	39.80	42.35	42.13	42.51	**43.86**
	SSIM	0.121	0.630	0.917	0.938	0.950	0.991	0.947	0.962	**0.994**
	SAM	31.013	9.172	4.242	2.981	2.350	1.973	2.183	1.840	**1.741**
50	PSNR	13.15	28.23	35.35	37.30	38.88	40.93	38.51	40.83	**41.33**
	SSIM	0.042	0.537	0.891	0.928	0.944	0.984	0.952	0.974	**0.989**
	SAM	56.637	18.859	7.452	5.217	3.439	3.271	3.310	3.581	**3.169**
70	PSNR	11.27	25.89	32.05	34.03	35.26	39.44	39.31	39.64	**40.28**
	SSIM	0.035	0.491	0.844	0.906	0.918	0.987	0.947	0.950	**0.991**
	SAM	53.885	13.758	6.707	4.643	3.210	**3.107**	3.271	3.260	3.186
Blind	PSNR	16.83	29.79	38.20	40.33	41.54	42.24	40.73	42.01	**43.28**
	SSIM	0.103	0.514	0.903	0.939	0.955	0.958	0.943	0.990	**0.995**
	SAM	72.516	33.936	13.873	9.172	5.847	3.011	3.270	3.105	**2.833**

Gaussian noise with variance $\sigma = 30$, and Fig. 6 presents the denoising results under i.i.d Gaussian noise with variance $\sigma = 50$. From Figs. 5 and 6, it can be observed that the KSVD algorithm causes serious distortion of the texture information in the image, which results in poor image quality.

The BM4D, MStSVD and DIP-SLR algorithms have better noise reduction effects, but they have limitation in preserving the details of the image. The LTDL, SST and D2Net algorithms preserve the detailed information of the image as much possible, but there may be prominence in some subtle edge parts. Compared with other methods, our proposed algorithm has better noise suppression and detail preservation ability for hyperspectral image. This is mainly because the cross-multiscale fusion attention module and spectral position-aware self-attention module can effectively utilize cross-space contextual information and spectral correlation to enhance the noise suppression capability. Meanwhile, the multi-scale fusion of spatial-spectral feature module is introduced to effectively fuse different spatial and spectral features of HSI.

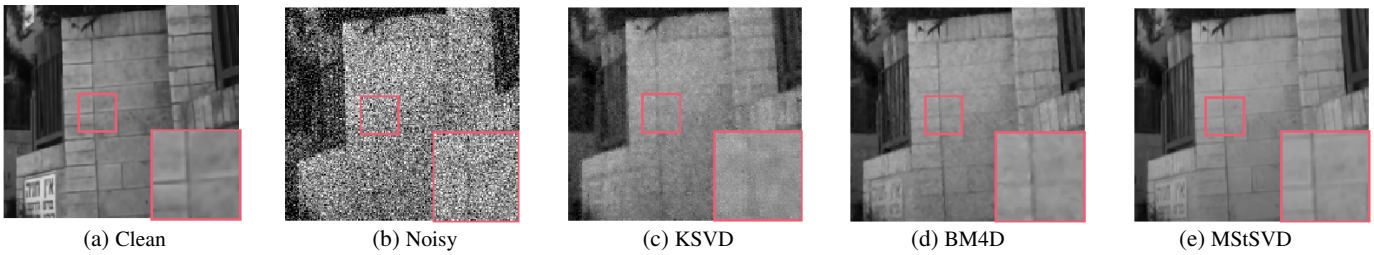
3.3.2. Results on the Hyperspectral Image with Complex Noise

In order to provide additional validation for the performance of the introduced MSF-Net method under various complex noises, the denoising performance comparison of our MSF-Net with other excellent denoising methods in four complex noise cases is given in Table 2. In Table 2, noise cases 1 to 4 represent non-i.i.d Gaussian noise, stripe noise, deadline noise, and their mixed noise, respectively. It can be clearly seen from Table 2 that the proposed MSF-Net method outperforms all the other methods in the stripe and mixture noise cases. In the non-i.i.d Gaussian noise case, the proposed MSF-Net has the highest PSNR and SSIM, and second best SAM. The comparison of the above experimental results proves the robustness and applicability of the hyperspectral denoising method proposed in this paper.

To demonstrate the superiority of the proposed method over other methods, Fig. 7 presents the noise reduction results of hyperspectral image in the stripe noise using different methods. It

TABLE 2. Comparison of denoising performance for complex noise on ICVL dataset.

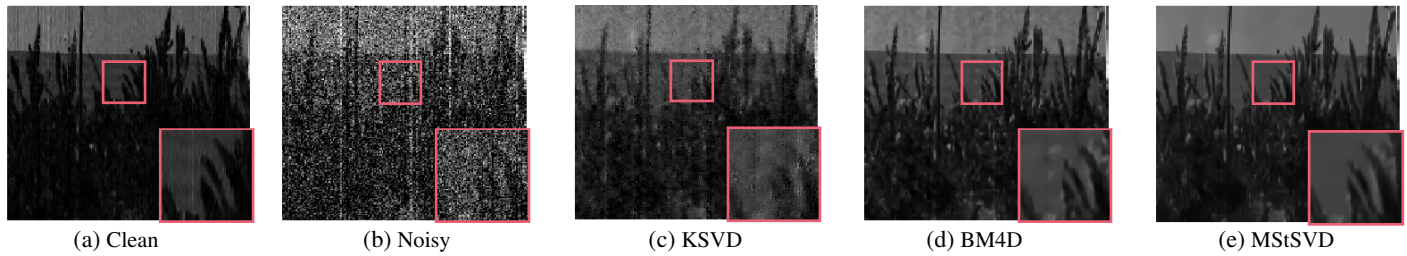
Case	Index	Noisy	KSVD	BM4D	MStSVD	LTDL	SST	DIP-SLR	D2Net	Ours
Case1	PSNR	17.86	26.78	34.73	36.46	37.51	42.53	42.03	42.56	**42.91**
	SSIM	0.175	0.538	0.881	0.913	0.925	0.984	0.970	0.989	**0.996**
	SAM	36.211	18.564	4.526	3.151	2.577	**2.074**	3.192	3.108	2.430
Case2	PSNR	17.30	26.91	25.95	36.05	37.21	38.66	35.49	39.95	**41.13**
	SSIM	0.159	0.509	0.569	0.914	0.947	0.965	0.970	0.972	**0.987**
	SAM	47.384	8.361	6.986	6.532	5.157	2.539	4.103	3.174	**2.372**
Case3	PSNR	18.86	26.93	28.06	32.17	33.92	37.83	38.62	**40.54**	40.16
	SSIM	0.214	0.566	0.645	0.829	0.872	0.948	0.961	0.973	**0.985**
	SAM	56.924	18.965	17.532	6.646	5.500	2.462	3.101	2.516	**2.355**
Case4	PSNR	13.99	20.38	22.44	28.60	32.30	38.78	36.15	38.97	**39.25**
	SSIM	0.108	0.363	0.551	0.701	0.840	0.947	0.926	0.959	**0.989**
	SAM	48.04	47.039	27.158	6.307	5.146	2.836	2.639	2.851	**2.582**



(a) Clean (b) Noisy (c) KSVD (d) BM4D (e) MStSVD

(f) LTDL (g) SST (h) DIP-SLR (i) D2Net (j) Ours

FIGURE 6. The denoising results on the ICVL dataset under Gaussian noise with variance $\sigma = 50$.



(a) Clean (b) Noisy (c) KSVD (d) BM4D (e) MStSVD

(f) LTDL (g) SST (h) DIP-SLR (i) D2Net (j) Ours

FIGURE 7. The denoising results for stripe noise on the ICVL dataset.

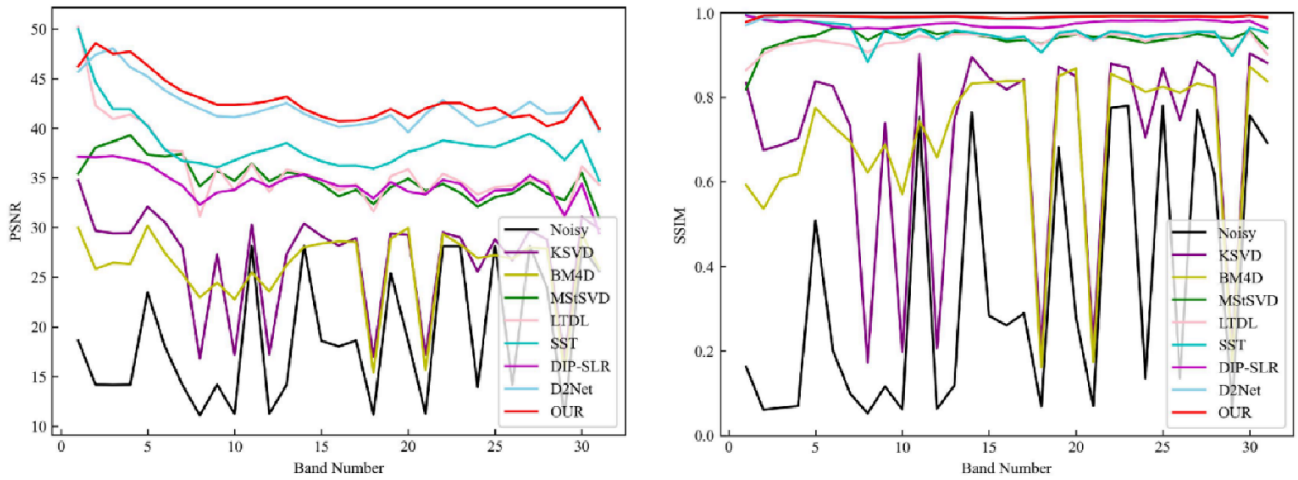


FIGURE 8. Comparison of PSNR and SSIM for all bands of hyperspectral image after stripe noise reduction.

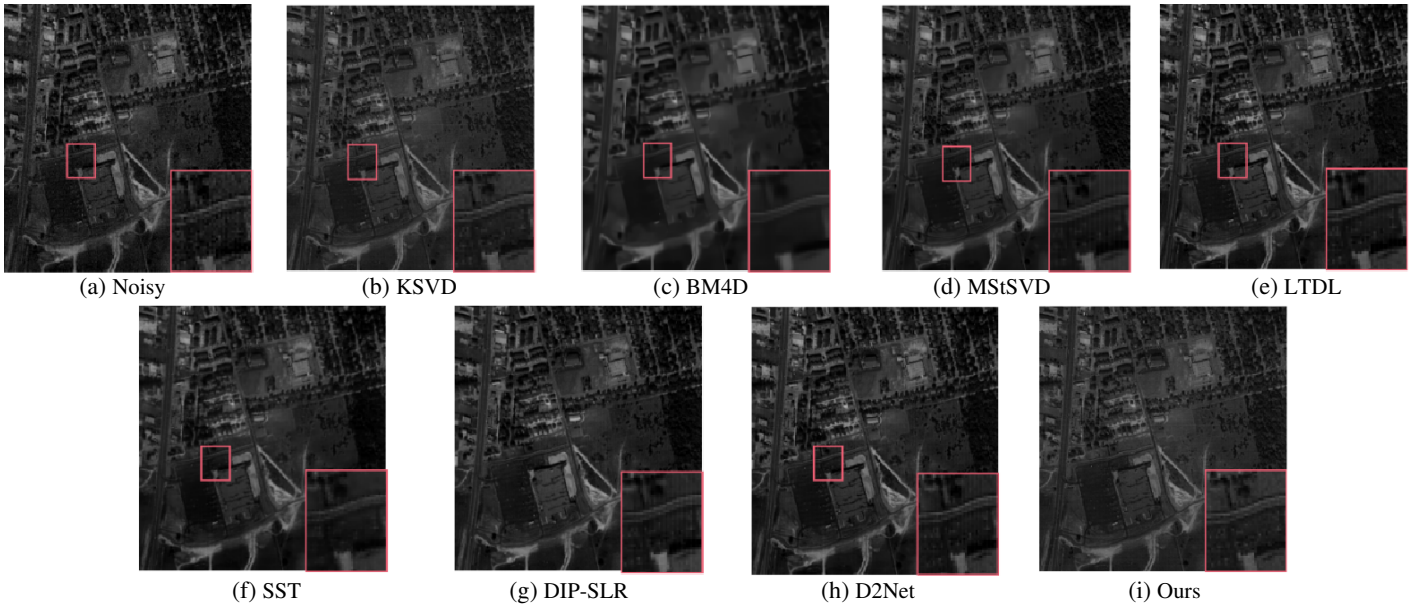


FIGURE 9. Noise reduction results of pseudo-color images synthesized by different methods for the real dataset Urban.

can be seen from Fig. 7 that the results of KSVD and DIP-SLR still have obvious noise traces. BM4D, MStSVD, LTDL and SST do not have obvious stripe noise, but the overall and local details are more blurred. Compared with other methods, D2Net and our proposed method have better noise suppression ability. However, our proposed method is clearer in local areas. Overall, these results demonstrate that our method has excellent denoising performance for hyperspectral image. To further quantify the denoising performance of the proposed method, the comparison of PSNR and SSIM for all bands of hyperspectral image after stripe noise reduction is plotted in Fig. 8. From Fig. 8, it can be seen that compared to other methods, the proposed method has better performance. It is evident that the proposed MSF-Net method exhibits robustness for noise reduction in the spectral dimension.

3.4. Evaluation on the Real Dataset

To comprehensively validate the performance of the proposed method, the hyperspectral image denoising experiments conducted on Urban dataset with real-world noise are shown in this section.

Due to the lack of clear images in the real dataset Urban, the effectiveness of the proposed method on the Urban dataset is intuitively evaluated by comparing the pseudo-color images before and after denoising. The denoising results are presented in Fig. 9. As can be seen from Fig. 9, our proposed method can effectively suppress mixed noise while recovering the detailed information in HSI such as edges and texture, which confirms the superiority of the proposed method in mixed noise removal. Furthermore, to quantitatively compare the denoising performance of the proposed method with other methods, we

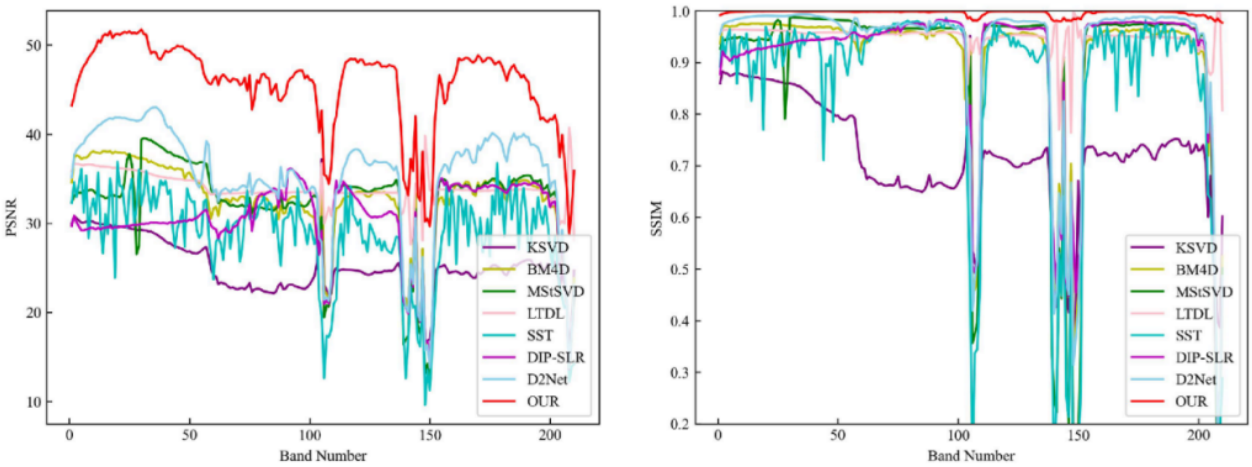


FIGURE 10. Comparison of PSNR and SSIM for all bands of hyperspectral image after noise reduction on real dataset.

TABLE 3. The ablation experiment results on ICVL dataset.

Baseline	C-MFA	SPA-SA	Ms-FSS	PSNR	SSIM	SAM
√	-	-	-	29.85	0.643	5.381
√	√	-	-	38.57	0.714	4.514
√	-	√	-	40.31	0.970	3.172
√	-	-	√	41.65	0.979	2.946
√	√	√	√	43.28	0.995	2.833

compared the PSNR and SSIM for all bands of hyperspectral image after noise reduction in Fig. 10. From Fig. 10, it is clear that the proposed method significantly improves PSNR performance in all bands, and the SSIM of the proposed method is superior to all other methods. This demonstrates the advantage of our model in HSI denoising.

3.5. Ablation Experiment

To demonstrate the impact of the proposed modules within the network, this section explores the ablation experiment implemented using the ICVL dataset.

3.5.1. Validity of Each Sub-Module

In each ablation experiment, independent identically distributed Gaussian noise with randomly distributed variance σ between 10 and 70 is added in the hyperspectral images. The ablation experiment results are presented in Table 3. In Table 3, √ represents including the module and — represents not including the module. In Table 3, the baseline model is denoised using only the basic wavelet transform and inverse wavelet transform. The PSNR, SSIM, and SAM values of the baseline model are calculated. Then, the performance metrics under each sub-module are analyzed. Firstly, it can be seen that the PSNR is 29.85 dB, SSIM 0.643 and SAM 5.38 under the baseline model. In the following, the sub-modules are added in order. The C-MFA module is first added. From Table 3, it can be seen that the PSNR improves by 8.72 dB; SSIM improves by 0.071; and SAM decreases by 0.867, which shows that it has a better denoising effect for the space. Secondly, the SPA-SA sub-module is added. From Table 3, we can see that the SAM is significantly reduced to 3.172, which shows that this module has a better denoising effect for the spectral dimension. Then, the Ms-FSS module is added. From Table 3, it can be seen that the SSIM is getting closer to 1, which means that our image is getting closer to the original image after the denoising model. Finally, according to the results in Table 3, it is evident that the proposed method, incorporating cross-multiscale fusion attention, spectral position-aware self-attention, and multi-scale fusion of spatial-spectral feature, yields superior denoising outcomes. It improves the PSNR from 29.85 dB to 43.28 dB, increases SSIM from 0.643 to 0.995 when the modules are added. The experimental results validate the effectiveness of each proposed module.

3.5.2. Effect of Different Dilation Rates in Ms-FSS

Table 4 shows the comparative experiments of dilated convolution on ICVL dataset for different dilation rates. From Table 4, it can be seen that when the dilation rate $= 1$, this is the standard convolution. When the dilation rate $= 2$, the receptive field is slightly enlarged. When the dilation rate $= 3$, this is the optimal equilibrium, which both enlarges the receptive field and is able to retain the details of the image better. When the dilation rate $= 4$, the receptive field is further enlarged, but it may lead

TABLE 4. Comparative experiments of dilated convolution on ICVL dataset for different dilation rates.

Dilation rate	PSNR (dB)	SSIM	SAM
1	35.2	0.92	3.4
2	37.5	0.94	3.3
3	38.6	0.96	2.9
4	37.8	0.95	3.1

to excessive smoothing. Therefore, the choice of the dilation rate of 3 is the best practice based on experimental validation and can effectively improve the denoising performance of the model.

4. CONCLUSION

In this paper, a multiscale spatial-spectral feature fusion network in frequency domain for HSI denoising is proposed. It effectively separates the signal and noise features through multi-scale discrete wavelet cascade decomposition. The proposed method utilizes the structural decomposition of multi-scale wavelet transform to replace traditional down-sampling operation, which could decompose noise into smaller scales, making it easier to remove in the frequency domain while minimizing information loss. In the network structure, cross-multiscale fusion attention C-MFA is introduced to improve the model performance by considering multiscale information and cross-space learning. And spectral position-aware self-attention module SPA-SA is designed to more fully exploit the spectral correlation in HSI and enhance the model ability of long-range spectral dependency. Furthermore, the multiscale fusion of spatial-spectral feature module Ms-FSS is proposed to effectively fuse different spatial and spectral features of HSI, thereby improving the denoising performance. Experimental results confirm that the proposed method outperforms mainstream denoising methods in terms of denoising performance and provides clearer visual results, such as texture details and edges.

ACKNOWLEDGEMENT

This work was supported by Natural Science Project of Science and Technology Department of Henan Province under Grant 252102211029; Natural Science Project of Zhengzhou Science and Technology Bureau under Grant 22ZZRDZX31; Project of Henan Key Laboratory of Superhard Abrasives under Grant JDKFJJ2023010; High-Performance Computing Platform of Henan University of Technology.

REFERENCES

[1] Wang, Y., X. Chen, F. Wang, M. Song, and C. Yu, "Meta-learning based hyperspectral target detection using siamese network," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 60, 1–13, 2022.

[2] Liu, C., L. Yang, Z. Li, W. Yang, Z. Han, J. Guo, and J. Yu, "Multi-level relation learning for cross-domain few-shot hyperspectral image classification," *Applied Intelligence*, Vol. 54, No. 5, 4392–4410, 2024.

[3] Ma, Z. and B. Yang, "Spatial-spectral hypergraph-based unsupervised band selection for hyperspectral remote sensing images," *IEEE Sensors Journal*, Vol. 24, No. 17, 27 870–27 882, 2024.

[4] Qu, X., J. Zhao, Y. Cheng, H. Tian, and G. Cui, "Compressed hyperspectral imaging based on residual-spectral attention mechanism and similar image prior," *Optics and Lasers in Engineering*, Vol. 180, 108330, 2024.

[5] Rasti, B., P. Scheunders, P. Ghamisi, G. Licciardi, and J. Chanussot, "Noise reduction in hyperspectral imagery: Overview and application," *Remote Sensing*, Vol. 10, No. 3, 482, 2018.

[6] Zhang, Q., Q. Yuan, J. Li, X. Liu, H. Shen, and L. Zhang, "Hybrid noise removal in hyperspectral imagery with a spatial–spectral gradient network," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 57, No. 10, 7317–7329, 2019.

[7] Bahraini, T., A. Ebrahimi-Moghadam, M. Khademi, and H. S. Yazdi, "Bayesian framework selection for hyperspectral image denoising," *Signal Processing*, Vol. 201, 108712, 2022.

[8] Li, H.-C., X.-R. Feng, R. Wang, L. Gao, and Q. Du, "Superpixel-based low-rank tensor factorization for blind nonlinear hyperspectral unmixing," *IEEE Sensors Journal*, Vol. 24, No. 8, 13 055–13 072, 2024.

[9] Fu, H., G. Sun, A. Zhang, B. Shao, J. Ren, and X. Jia, "Tensor singular spectrum analysis for 3-D feature extraction in hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 61, 1–14, 2023.

[10] Wang, Z., M. K. Ng, L. Zhuang, L. Gao, and B. Zhang, "Nonlocal self-similarity-based hyperspectral remote sensing image denoising with 3-D convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 60, 1–17, 2022.

[11] He, C., L. Sun, W. Huang, J. Zhang, Y. Zheng, and B. Jeon, "TSLRLN: Tensor subspace low-rank learning with non-local prior for hyperspectral image mixed denoising," *Signal Processing*, Vol. 184, 108060, 2021.

[12] Maggioni, M., V. Katkovnik, K. Egiazarian, and A. Foi, "Nonlocal transform-domain filter for volumetric data denoising and reconstruction," *IEEE Transactions on Image Processing*, Vol. 22, No. 1, 119–133, 2013.

[13] Peng, J., W. Sun, H.-C. Li, W. Li, X. Meng, C. Ge, and Q. Du, "Low-rank and sparse representation for hyperspectral image processing: A review," *IEEE Geoscience and Remote Sensing Magazine*, Vol. 10, No. 1, 10–43, 2021.

[14] Gong, X., W. Chen, and J. Chen, "A low-rank tensor dictionary learning method for hyperspectral image denoising," *IEEE Transactions on Signal Processing*, Vol. 68, 1168–1180, 2020.

[15] Zhang, F., K. Zhang, W. Wan, and J. Sun, "3D geometrical total variation regularized low-rank matrix factorization for hyperspectral image denoising," *Signal Processing*, Vol. 207, 108942, 2023.

[16] Murugesan, R., N. Nachimuthu, and G. Prakash, "Attention based deep convolutional U-Net with CSA optimization for hyperspectral image denoising," *Infrared Physics & Technology*, Vol. 129, 104531, 2023.

[17] Fu, H., Z. Ling, G. Sun, J. Ren, A. Zhang, L. Zhang, and X. Jia, "Hyperdehazing: A hyperspectral image dehazing benchmark dataset and a deep learning model for haze removal," *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 218, 663–677, 2024.

[18] Sun, L., Q. Cao, Y. Chen, Y. Zheng, and Z. Wu, "Mixed noise removal for hyperspectral images based on global tensor low-rankness and nonlocal SVD-aided group sparsity," *IEEE Trans-*

actions on Geoscience and Remote Sensing, Vol. 61, 1–17, 2023.

[19] Chang, Y., L. Yan, H. Fang, S. Zhong, and W. Liao, "HSI-DeNet: Hyperspectral image restoration via convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 57, No. 2, 667–682, 2018.

[20] Nguyen, H. V., M. O. Ulfarsson, J. Sigurdsson, and J. R. Sveinsson, "Deep sparse and low-rank prior for hyperspectral image denoising," in *IGARSS 2022 — 2022 IEEE International Geoscience and Remote Sensing Symposium*, 1217–1220, Kuala Lumpur, Malaysia, 2022.

[21] Do, T. and T. Vetterli, "Wavelet-based denoising methods: A review," *IEEE Signal Processing Magazine*, Vol. 22, No. 6, 84–93, 2005.

[22] Zhang, Q., Q. Yuan, J. Li, X. Liu, H. Shen, and L. Zhang, "Hybrid noise removal in hyperspectral imagery with a spatial-spectral gradient network," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 57, No. 10, 7317–7329, 2019.

[23] Wang, S., L. Li, X. Li, J. Zhang, L. Zhao, X. Su, and F. Chen, "A denoising network based on frequency-spectral-spatial-feature for hyperspectral image," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 16, 6693–6710, 2023.

[24] Dusmanu, M., I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-net: A trainable cnn for joint description and detection of local features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8092–8101, 2019.

[25] Pan, E., Y. Ma, X. Mei, F. Fan, J. Huang, and J. Ma, "SQAD: Spatial-spectral quasi-attention recurrent network for hyperspectral image denoising," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 60, 1–14, 2022.

[26] Yuan, Q., Q. Zhang, J. Li, H. Shen, and L. Zhang, "Hyperspectral image denoising employing a spatial-spectral deep residual convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 57, No. 2, 1205–1218, 2018.

[27] Carion, N., F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*, 213–229, Springer International Publishing, Cham, 2020.

[28] Wang, M., W. He, and H. Zhang, "A spatial-spectral transformer network with total variation loss for hyperspectral image denoising," *IEEE Geoscience and Remote Sensing Letters*, Vol. 20, 5503105, 2023.

[29] Lai, Z., C. Yan, and Y. Fu, "Hybrid spectral denoising transformer with guided attention," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13 065–13 075, Paris, France, 2023.

[30] Li, M., Y. Fu, and Y. Zhang, "Spatial-spectral transformer for hyperspectral image denoising," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, No. 1, 1368–1376, Washington, USA, 2023.

[31] Sun, L., X. Wang, Y. Zheng, Z. Wu, and L. Fu, "Multiscale 3-D-2-D mixed CNN and lightweight attention-free transformer for hyperspectral and LiDAR classification," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 62, 2100116, 2024.

[32] Hu, S., F. Gao, X. Zhou, J. Dong, and Q. Du, "Hybrid convolutional and attention network for hyperspectral image denoising," *IEEE Geoscience and Remote Sensing Letters*, Vol. 21, 5504005, 2024.

[33] Xiong, F., J. Zhou, Q. Zhao, J. Lu, and Y. Qian, "MAC-Net: Model-aided nonlocal neural network for hyperspectral image denoising," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 60, 1–14, 2021.

[34] Wang, D., J. Zhang, B. Du, L. Zhang, and D. Tao, "DCN-T: Dual context network with transformer for hyperspectral image classification," *IEEE Transactions on Image Processing*, Vol. 32, 2536–2551, 2023.

[35] Xue, X., H. Zhang, B. Fang, Z. Bai, and Y. Li, "Grafting transformer on automatically designed convolutional neural network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 60, 1–16, 2022.

[36] Bodrito, T., A. Zouaoui, J. Chanussot, and J. Mairal, "A trainable spectral-spatial sparse coding model for hyperspectral image restoration," *Advances in Neural Information Processing Systems*, Vol. 34, 5430–5442, 2021.

[37] Kalman, L. S. and E. M. B. III, "Classification and material identification in an urban environment using HYDICE hyperspectral data," in *Imaging Spectrometry III*, Vol. 3118, 57–68, San Diego, CA, United States, 1997.

[38] Aharon, M., M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, Vol. 54, No. 11, 4311–4322, 2006.