(Received 8 July 2025, Accepted 22 October 2025, Scheduled 19 November 2025)

Hyperspectral Image Denoising Using Spatial Spectral Attention Network Based on Transformer

Xiaozhen Ren[1], Jing Cui[2], Yi Hu[1], Zhipeng Guo[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: Although Transformer models have made significant progress in the field of hyperspectral image denoising, their original architecture still has limitations in processing the spatial and spectral correlations of images. It often results in the loss of details in spatial features and insufficient exploration of the uniqueness of different spectral bands. To overcome these challenges, this paper proposes a Transformer based spatial spectral attention network aimed at enhancing the utilization efficiency of spatial spectral correlations. In response to the common problem of over smoothing in spatial feature processing, a dual channel spatial feature fusion module is introduced, which effectively enhances the capture of spatial details and ensures clear reproduction of image textures and edges. Meanwhile, in the spectral dimension, a multi-scale spectral feature extraction with self-attention mechanism is applied, which can sensitively identify and utilize the differences between spectral bands, thereby achieving more accurate feature extraction at the spectral level. By integrating residual connections in the spatial spectral feature extraction layer, the model can efficiently fuse spatial and spectral information, ultimately achieving high-quality denoising. The experimental results have verified the excellent performance of this method on both the ICVL dataset and Urban real dataset, achieving good denoising results and demonstrating significant advantages in maintaining image details and spectral fidelity.

1. INTRODUCTION

Hyperspectral images are three-dimensional datasets obtained by spectroscopic instruments, surpassing traditional color images with their rich spectral resolution. They can finely capture the spectral characteristics of every pixel in the scene and can almost fully present the reflection spectral curve of objects. This characteristic of hyperspectral images has shown great potential in various applications, including remote sensing detection [1], material identification [2], agricultural production monitoring [3], medical image analysis [4], and classification [5]. However, the hyperspectral imaging process is often compromised by multiple factors, such as insufficient illumination, low reflected energy, atmospheric condition changes, and sensor electronic noise, which often lead to image quality degradation and various noise doping. It poses obstacles for subsequent image analysis and application. Consequently, denoising of hyperspectral images has become a crucial preprocessing step, which directly affects the ability to accurately extract useful information from hyperspectral images, thereby ensuring the accuracy and reliability of subsequent analysis. Therefore, efficient and accurate denoising techniques are the key to promoting the application and expansion of hyperspectral images in various fields.

The denoising process of hyperspectral images is a reverse process aimed at restoring cleaner images from contaminated hyperspectral images. To address this challenge, researchers have delved into the inherent properties of hyperspectral images themselves. The core of such methods lies in utilizing the intrinsic physical properties of hyperspectral images, such as low-rank prior structure [6–8], non-local similarity [9, 10], spatial spectral correlation [11, 12], and global spectral autocorrelation [13], to construct effective constraint conditions for denoising. For example, the K-singular value decomposition (KSVD) [14] uses dictionary learning for denoising, but ignores the inherent spectral information of hyperspectral images. block-matching and three-dimensional (3D) filtering (BM3D) [15] focuses on non-local self-similarity in the spatial dimension, but fails to fully explore information in the spectral dimension. Hyper-Laplacian regularized unidirectional low-rank tensor recovery (LLRT) [16] approximates the structure of the original hyperspectral image through a non-local low rank tensor model. Non-local tensor dictionary learning (LTDL) [17] is a framework based on dictionary learning. Minimum spanning tree singular value decomposition (MStSVD) [18] combines sparse representation and low rank decomposition methods for denoising. Block-matching and four-dimensional filtering (BM4D) [19] adopts threshold filtering technology. Although these methods have utilized the potential features of space and spectrum to varying degrees, they have achieved ideal denoising effects. However, their effectiveness is ultimately limited by the correlation between manually set prior information and essential characteristics of hyperspectral images. On the other hand, BM4D method often involves complex iterative optimization processes for denoising methods of hyperspectral images based on prior constraints. This not only increases the time complexity of the methods but

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

doi:10.2528/PIERC25070802

also limits their application in real-time processing of large-scale datasets. Therefore, it has become a hot and difficult topic in current research for seeking more efficient denoising strategies to adapt to the complex structure and fast processing requirements of hyperspectral images.

In recent years, with the rapid development of deep learning, especially the rise of convolutional neural networks (CNNs), new vitality has been injected into various fields [20–22]. For example, a hyperspectral image denoising method, denoising method with deep image prior and sparse low-rank prior (DIP-SLR) [23], which combines the deep sparse and low-rank prior, is proposed to improve performance. Yuan et al. [24] innovatively proposed mapping method between the noisy and clean hyperspectral images (HSI) with deep convolutional neural network (HSID-CNN), cleverly incorporating a residual connection mechanism to achieve end-to-end learning from noisy inputs to clear images without noise, significantly improving denoising accuracy and efficiency. Dusmanu et al. [25] proposed a trainable convolutional neural network for joint description and detection of local features (D2Net) to achieve better denoising effects. Three-dimensional quasi-recurrent neural network (QRNN3D) [26] effectively captures the intrinsic relationship between spatial and spectral features of hyperspectral images through the combination of three-dimensional (3D) convolution and quasi-recurrent techniques, further enhancing the denoising effect. Maffei et al. [27] proposed HSI single denoising CNN (HSI-SDeCNN), which directly takes 3D image blocks as inputs to the network and uses downsampling to reduce computation time.

At the same time, the Transformer architecture, which originally shone in the field of natural language processing, has also begun to cross over into the field of computer vision [28]. Li et al. [29] constructed spectral enhanced rectangle transformer (SERT) model, which utilizes the spatial similarity of hyperspectral images and the low rank of spectra to design the Transformer structure for denoising purposes. Zhang et al. [30] proposed three-dimension spatial-spectral attention transformer (TDSAT), which combines three-dimensional convolution and spectral-spatial attention Transformer blocks to denoise HSI with an arbitrary number of bands. The U-former designed by Wang et al. [31], with powerful capabilities of the Transformer, successfully captured the global dependency in the image spatial domain, greatly improving the denoising effect. Three-dimensional quasi recurrent and transformer based network (TRQ3D) [32] further integrates structures of U-former and QRNN3D, using a dual-branch approach to deeply extract image features from both spatial and spectral levels, and then fuse them to achieve better denoising results. Although Transformer architecture has shown great potential in denoising hyperspectral images, its utilization of spatial and spectral correlations still needs to be improved. The existing methods tend to be overly smooth when processing spatial features, which may fail to extract small-scale information features. In the spectral dimension, there is also a drawback of overly focusing on channel characteristics and neglecting the unique differences between different spectral bands.

To address the above issues, a spatial spectral attention network based on Transformer is proposed for hyperspectral image denoising (SSAT-HSI), which effectively integrates spatial and spectral features through a hierarchical architecture. It mainly consists of two modules: dual channel spatial feature fusion module (DC-SFM) and multi-scale spectral feature extraction (MSFE) with self-attention module. In the spatial dimension, the dual channel spatial feature fusion module DC-SFM is used to extract features, consisting of two branches. This module integrates a local processing branch, which utilizes convolution and channel shuffling, with a global processing branch based on self-attention. This hybrid design enables the simultaneous capture of fine-grained textures and long-range dependencies, more effectively preserving image edges and structural details. In the spectral dimension, convolutions of different scales are first used to obtain features at multiple scales, and then self-attention is used to select spectral bands for denoising. Finally, extensive experiments are conducted on both simulated and real-world hyperspectral datasets, including ICVL and Urban. The results demonstrate that the proposed SSAT-HSI outperforms state-of-the-art traditional and deep learning-based methods in terms of peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and spectral angular mapping (SAM) metrics, particularly under complex noise conditions, validating its robustness and effectiveness in maintaining both spatial fidelity and spectral consistency.

2. PROPOSED DENOISING METHOD

Figure 1 shows the architecture of the spatial spectral attention network based on the Transformer designed for hyperspectral image denoising SSAT-HSI in this paper. It consists of six concatenated spatial spectral denoising blocks (SSDBs). Each SSDB is embedded with four Transformer sub-modules, each of which integrates a spatial spectral feature extraction layer (SSFEL) internally. With this hierarchical network design, the proposed model is able to efficiently capture and process the complex relationships between spatial and spectral features in hyperspectral images. The brilliance of this architecture lies in its ability to significantly improve the quality of denoised images while ensuring fine preservation of detailed information.

Specifically, the noisy hyperspectral image can be seen as a clean hyperspectral image corrupted by various types of noise. This image degradation process can be modeled as

$$Y = X + N \tag{1}$$

where N represents the noise, $X \in R^{H\times W\times B}$ the clean hyperspectral image, $Y \in R^{H\times W\times B}$ the actual observed hyperspectral image after being contaminated by noise N, where H and W represent the height and width of the spatial dimension, respectively. And B represents the spectral dimension of the hyperspectral image, that is, the number of bands.

Before entering SSDB for deep processing, a 3×3 convolutional layer is first used to extract preliminary features F_0 from the noisy image Y. Then, the initially extracted features are input into the SSDB module with a fixed size for deeper analysis and denoising of spatial spectral features.

$$F_l = SSDB_l(F_{l-1}), \quad l = 1, 2, \cdots, 6 \tag{2}$$

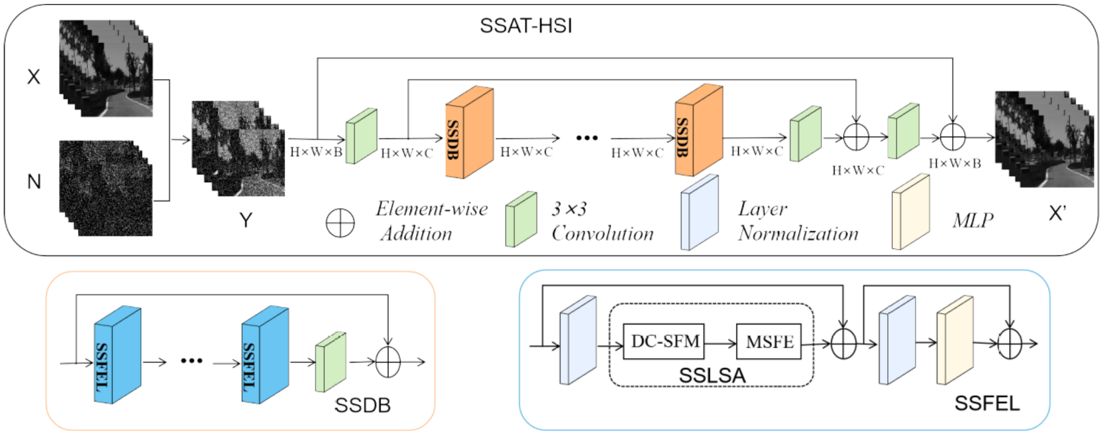


FIGURE 1. The architecture of SSAT-HSI.

where $SSDB_l$ represents the spatial spectral denoising block of the lth layer.

To enhance the expressive power of the network, a 3×3 convolutional layer is embedded in each spatial spectral denoising unit, and these units are concatenated through a skip connection. In addition, to prevent overfitting of the model during training, the residual connection is added before each unit. Moreover, in order to recover a cleaner hyperspectral image from the final feature F_6, the shallow feature is connected through two 3×3 convolutional layers and the skip connection. Finally, the original noisy image is reintroduced to ensure the integrity and accuracy of the restored image.

2.1. Spatial Spectral Feature Extraction Layer SSFEL

In the absence of the residual connection and the multi-layer perceptron (MLP) architecture, self-attention networks exhibit a nonlinear, sharp decline as the network depth increases. This will ultimately result in the model being unable to effectively capture most of the detailed features of the input image. To solve this problem, the spatial spectral feature extraction layer (SSFEL) is proposed in this section. This design principle closely follows the basic architecture of visual ViT [33]. After completing feature processing through the self-attention module, normalization, residual connection, and MLP processing layers are constructed.

Specifically, assume that Z_l represents the input feature of the lth concatenated spatial spectral feature extraction layer SSFEL, the output of the lth SSFEL can be written as

$$Z_l^{'} = SSLSA\left(LN(Z_{l-1})\right) + Z_{l-1} \tag{3}$$

$$Z_l = MLP\left(LN(Z_l^{'})\right) + Z_l^{'} \tag{4}$$

where $Z_l^{'}$ denotes the output of spatial spectral multi-head self-attention (SSLSA), and Z_l represents the output of the lth SSFEL.

SSLSA includes two parts: dual-channel spatial feature fusion module (DC-SFM) and multi-scale spectral feature extraction with self-attention (MSFE). Assuming that the input feature of SSLSA is $Z^{in} \in R^{H\times W\times C}$, after normalization, the feature is first divided in the DC-SFM. In DC-SFM, the input feature is divided into several sizes of $M \times M$ non-overlapping blocks, so that the entire input feature is segmented into HW/M^2 blocks. The feature within each block is processed independently to capture local spatial information more finely. This partitioning method not only helps reduce computational complexity but also enhances the model's perception of local image structures. The extraction of local and global features for each block through the DC-SFM can be expressed as

$$\begin{aligned} Z_i^{in} &= WinPartition(Z^{in}), \quad i = 1, \cdots, L \qquad (5)\\ Z_i^{spa} &= DC{-}SFM(Z^{in}), \quad i = 1, \cdots, L \qquad (6)\\ Z_{spa} &= WinReverse(Z_i^{spa}), \quad i = 1, 2, L \qquad (7) \end{aligned}$$

where $WinPartition(\cdot)$ represents the division of the input feature; $DC{-}SFM(\cdot)$ represents the dual-channel spatial feature fusion module; $WinReverse(\cdot)$ is the inverse transformation of the $WinPartition$ processing, which restores the features processed by the dual channel spatial feature fusion module to the same size as the original input features through upsampling or interpolation operations. This step ensures that the spatial structure of features is preserved, while providing rich information input for the next stage of the network. Then, these processed output features are passed on to multi-scale spectral feature extraction with self-attention module MSFE for further integration and analysis, and this process can be expressed as

$$Z_{spe} = MSFE(Z_{spa}). \tag{8}$$

2.2. Dual-Channel Spatial Feature Fusion Module (DC-SFM)

Due to the limitations of local characteristics and a limited receptive field, convolution operation has insufficient performance in constructing global features. In contrast, the Transformer architecture excels in extracting global features and analyzing long-range dependency in images through its attention mechanism. Convolution and attention mechanism are not mutually exclusive but complementary. They work together to extract both local detailed features and capture global features. Therefore, the dual-channel spatial feature fusion module DC-

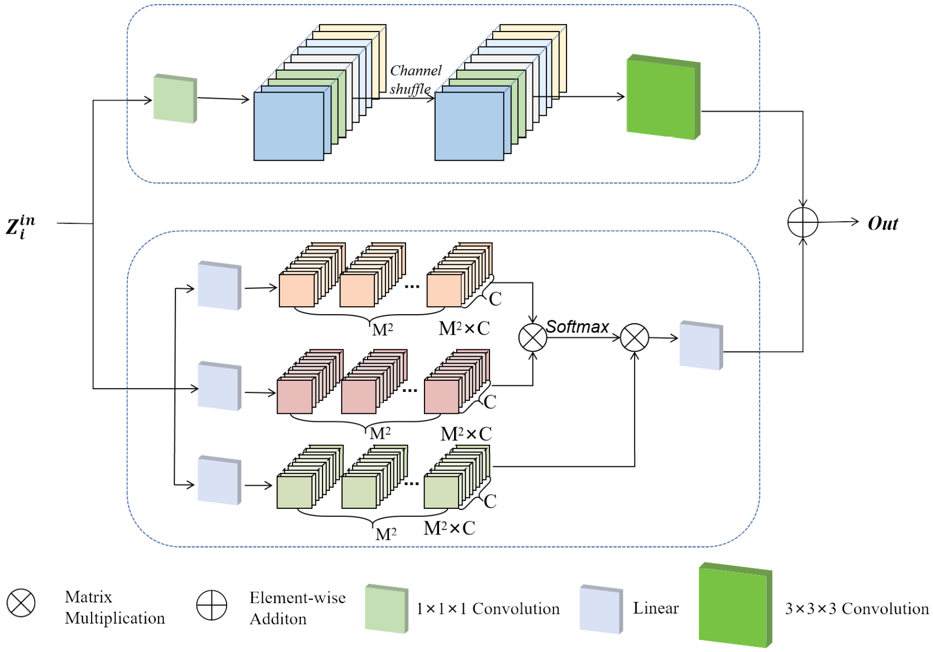


FIGURE 2. Dual-channel spatial feature fusion module DC-SFM.

SFM is designed in this paper, as shown in Figure 2. DC-SFM includes global processing unit and local processing unit. In DC-SFM, self-attention mechanism is used to build the global processing unit, aiming to obtain more abundant global features. The local processing unit focuses on mining the details of local regions. The two units work together to serve the comprehensive denoising of hyperspectral images.

In order to enhance information exchange and fusion between different channels, $1 \times 1 \times 1$ convolution is used to adjust the channel dimension for the local processing unit. This operation can improve the interaction between various channels, thereby enhancing the ability to integrate information. Next, channel shuffling is performed to better integrate channel information. Channel shuffling divides the input feature map into multiple independent subsets along the channel direction, and uses depthwise separable convolution within each subset to facilitate information shuffling between channels. This operation not only saves computational resources but also maintains the correlation between channels. Subsequently, the processed outputs of each subset are merged again along the channel dimension to form a new multi-channel feature map. Finally, $3 \times 3 \times 3$ convolution is used to obtain richer and more refined local representations. The output of the local processing unit can be expressed as

$$Out_1 = C_{3\times3\times3}\left(Cs\left(C_{1\times1\times1}(Z_i^{in})\right)\right) \tag{9}$$

where Z_i^{in} is the input feature, $C_{1\times1\times1}$ the $1 \times 1 \times 1$ convolution, Cs the channel shuffling, $C_{3\times3\times3}$ the $3\times3\times3$ convolution, and Out_1 the output of the local processing unit.

In the global processing unit, input Z_i^{in} is first linearly projected to obtain query Q_{spa}, key K_{spa}, and value V_{spa}, which can be expressed as:

$$Q_{spa} = W_q Z_i^{in}, \quad K_{spa} = W_k Z_i^{in}, \quad V_{spa} = W_v Z_i^{in} \tag{10}$$

where W_q, W_k, and W_v denote the weight of size $C \times C$.

Q_{spa} and K_{spa} perform point-wise multiplication to obtain a spatial attention map $M_{spa} \in R^{M^2 \times M^2}$. Then, the output of the global processing unit can be expressed as

$$M_{spa} = Soft\max\left(\frac{Q_{spa} \cdot K_{spa}}{\sqrt{d}} + B\right) \tag{11}$$

$$Out_2 = M_{spa} V_{spa} \tag{12}$$

where B is the relative deviation, d the dimension of Q_{spa}, and its value is C/N.

Therefore, the output of DC-SFM is expressed as

$$Out = Out_1 + Out_2 \tag{13}$$

The dual-channel design of DC-SFM essentially decomposes the spatial degradation process in hyperspectral images. One branch handles the degradation caused by the atmosphere and optical system, while the other branch handles the degradation caused by the sensor. This decomposable design more comprehensively models complex mixed degradation processes than single-mechanism approaches, such as pure CNNs or pure Transformers. Consequently, it effectively avoids issues like excessive smoothing or detail loss that arises from overly simplistic mechanisms.

2.3. Multi-Scale Spectral Feature Extraction with Self-Attention Module (MSFE)

The dual-channel spatial feature fusion module (DC-SFM) can capture rich spatial details of a hyperspectral image. However, due to the unique spectral information of a hyperspectral image, simple spatial features are not sufficient to fully reflect their intrinsic characteristics. Therefore, a multi-scale spectral feature extraction with self-attention module (MSFE) is proposed to

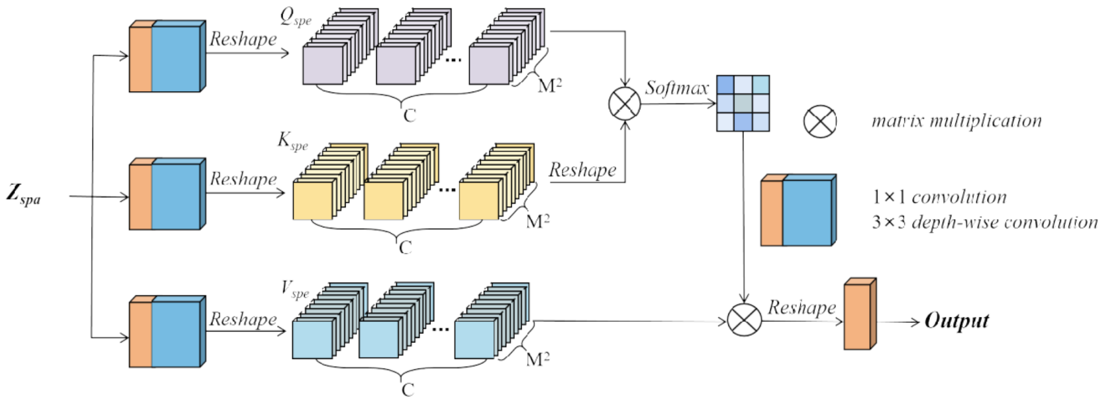


FIGURE 3. Multi-scale spectral feature extraction with self-attention module MSFE.

fully extract and utilize global correlation in the spectral domain. It focuses on analyzing the connection between different bands in hyperspectral image and analyzing spectral features from a multi-scale perspective. The MSFE can not only capture local spectral changes but also perceive a wider spectral context.

The structure of MSFE is shown in Figure 3. Firstly, the input feature Z_{spa} is processed through 1×1 convolution and 3×3 depth-wise convolution to obtain query Q_{spe}, key K_{spe}, and value V_{spe}, which can be expressed as

$$\begin{aligned} Q_{spe} &= W_1 W_2 Z_{spa}, \quad K_{spe} = W_1 W_2 Z_{spa}, \\ V_{spe} &= W_1 W_2 Z_{spa} \end{aligned} \tag{14}$$

where W_1 denotes the 1×1 convolution, and W_2 denotes the 3×3 depth-wise convolution. The W_1 and W_2 in Q_{spe}, K_{spe}, and V_{spe} are independent convolutional kernels, so the generated Q_{spe}, K_{spe}, and V_{spe} are numerically distinct due to the absence of weight sharing.

The similarity between query Q_{spe} and key K_{spe} is calculated to obtain the attention matrix $M_{spe} \in R^{C \times C}$, which reflects the interdependence between query and key in the entire feature space. Then, the attention matrix is multiplied by V_{spe} to generate a context-aware feature representation that could capture long-range dependency between features, thereby enhancing the model's understanding and expression of complex patterns. Compared to Z_{spa}, Z_{spe} has more spectral details and preserves key spatial information.

$$M_{spe} = Soft\max\left(\frac{Q_{spa} \cdot K_{spa}}{\varepsilon} + B\right) \tag{15}$$

$$Z_{spe} = V_{spe} \cdot M_{spe} \tag{16}$$

where ε is a learnable parameter that adjusts the point-wise multiplication of Q_{spe} and K_{spe}, and B is a learnable relative position encoding.

The MSFE module constructs a learning framework aligned with hyperspectral physics principles. The multiscale convolutions analyze spectral curves, while self-attention selects and fuses features. This design ensures that critical spectral features are preserved during denoising while noise and outliers are effectively suppressed, fundamentally guaranteeing spectral fidelity in the denoised images.

3. EXPERIMENTS AND ANALYSIS

3.1. Datasets

The ICVL dataset [34] and the Urban real dataset [35] are used to evaluate the proposed model. The ICVL dataset contains 201 hyperspectral images with a spectral range of 400–700 nm. 100 clean hyperspectral images are randomly selected and add different intensities of Gaussian noise as the training set. The size of the image in ICVL dataset is $1392 \times 1300 \times 31$. Each image is cropped, randomly flipped, and adjusted to form a $64 \times 64 \times 31$ training set. For the test set, 50 hyperspectral images were selected, and each hyperspectral image is cropped to a size of $512 \times 512 \times 31$ to achieve better visual effect.

3.2. Experimental Details

In the experiments, Adam optimization algorithm [36] is employed to fine tune our network architecture. Parameter initialization follows the Xavier initialization scheme, which can help accelerate the training process and avoid gradient vanishing or exploding problems. Setting the batch size to 8 means that during each iteration, the network will process 8 samples simultaneously to update weights, balancing memory usage and training efficiency. The training runs a total of 100 epochs. The learning rate adopts a phased approach. The initial learning rate is set to 1×10^{-4}, which provides sufficient flexibility in the early stage of training to quickly adapt to data pattern. After the 60th epoch, the learning rate decreases to 1×10^{-5}. This strategy helps refine the weight adjustment of the network, prevent overfitting, and ensure stable convergence of the model in later training. In order to quantify the deviation between the predicted result of the model and the actual hyperspectral image, mean square error (MSE) is used as the loss function, because MSE loss can effectively measure the average square difference between the predicted value and the true value, and can sensitively reflect the quality of denoising effect.

The proposed method is compared with traditional methods such as KSVD [14], BM4D [19], and deep learning based methods such as SERT [29], D2Net [25] and DIP-SLR [23]. Traditional methods run in MATLAB environment, while deep learning methods and the proposed method are implemented in PyTorch environment and run with GeForce RTX 3060.

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

σ	Index	Noisy	KSVD	BM4D	SERT	DIP-SLR	D2Net	Our
30	PSNR	18.58	29.65	36.41	43.35	42.13	42.51	**43.77**
	SSIM	0.121	0.830	0.917	0.991	0.947	0.962	**0.997**
	SAM	31.013	9.172	4.242	1.973	2.183	1.840	**1.731**
50	PSNR	13.15	28.23	35.35	40.93	38.51	40.83	**41.43**
	SSIM	0.042	0.797	0.891	0.984	0.952	0.974	**0.995**
	SAM	56.637	18.859	7.452	3.271	3.310	3.581	**2.161**
70	PSNR	11.27	25.89	32.05	39.44	39.31	39.64	**39.99**
	SSIM	0.035	0.891	0.844	0.987	0.947	0.950	**0.992**
	SAM	53.885	13.758	6.707	**3.107**	3.271	3.260	3.146
Blind	PSNR	16.83	29.79	38.20	42.24	40.73	42.01	**43.03**
	SSIM	0.103	0.814	0.853	0.958	0.943	0.990	**0.995**
	SAM	72.516	33.936	13.873	3.011	3.270	3.105	**2.833**

In this paper, spectral angular mapping (SAM), structural similarity (SSIM), and peak signal-to-noise ratio (PSNR) are selected as evaluation metrics to quantitatively evaluate the denoising performance. The larger the values of PSNR and SSIM are, the better the denoising effect is, while the smaller the value of SAM is, the better the spectral fidelity of the denoised image is.

3.3. Experiments on the ICVL Dataset

3.3.1. Results on the Hyperspectral Image with Gaussian Noise

In order to evaluate the performance of our method in dealing with Gaussian noise, the zero-mean additive Gaussian white noise is added on the original pure hyperspectral image. The noise intensity is controlled by its variance σ, and different values of σ represent different degrees of noise pollution. The experimental results on the ICVL dataset are summarized in Table 1. It is worth noting that the best performance results in Table 1 are marked in bold, and the second-best performance results are underlined. It can be seen from Table 1 that when the noise variance σ is 30 and 50, respectively, the noise reduction effect of the proposed method is the best. When the noise variance σ is 70, the values of PSNR and SSIM obtained by the proposed SSAT-HSI are still the highest. Even in the case of blind noise, all the metrics of the proposed SSAT-HSI are still the best.

In order to more intuitively show the noise reduction performance of the proposed SSAT-HSI, Figure 4 shows the hyperspectral image after denoising when the variance of Gaussian noise is 50. In order to better display the denoising effect, the local details of the hyperspectral image are enlarged. It can be seen that the denoised images obtained by KSVD and BM4D still contain noise, and the denoising effect is not satisfactory. The results of DIP-SLR and D2Net have artifacts near the eaves, and the proposed method in this paper has better edge detail preserving ability. The experimental results show that the proposed SSAT-HSI achieves favorable performance in low noise environments, demonstrating its robustness and excellent denoising capabilities.

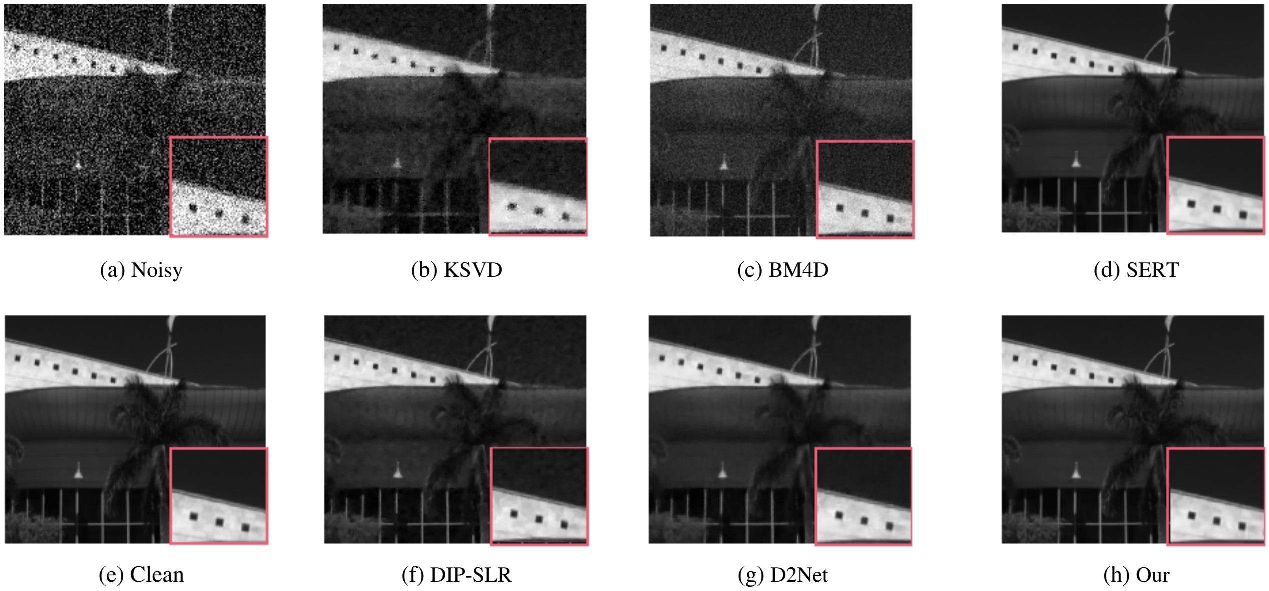
3.3.2. Results on the Hyperspectral Image with Complex Noise

To further validate the performance of the proposed SSAT-HSI, the denoising performance under various complex noises is analyzed. The complex noise mainly includes non-independent and identically distributed Gaussian noise, deadline noise, stripe noise, and their mixture noise. The experimental results on the ICVL dataset under various complex noises are presented in Table 2. From Table 2, it can be seen that the proposed method achieves the best noise reduction effect when dealing with deadline noise, stripe noise, and mixture noise. For non-i.i.d. Gaussian noise, the proposed SSAT-HSI yields the highest PSNR and SSIM values, with SAM being its second-strongest metric.

To illustrate the advantage of the proposed method over other methods, Figure 5 displays the mixture noise removal results of hyperspectral images obtained by different methods on the ICVL dataset. From Figure 5, it can be seen that the denoising results of KSVD and BM4D methods have artifacts and partial loss of structural information. The denoising results of SERT, DIP-SLR, and D2Net have some blurry details, while the result of our method is relatively clear. This is mainly because the MSFE module extracts spectral features from different receptive fields, and the subsequent self-attention mechanism dynamically assigns weights to different bands based on these multi-scale features. It automatically identifies and suppresses “low signal-to-noise ratio” bands severely contaminated by noise while enhancing “high signal-to-noise ratio” bands. This adaptive feature selection capability enables the model to intelligently prioritize during denoising. It effectively reduces noise while maximally preserving the shape and characteristics of the original spectral curve. Moreover, the strength of SSAT-HSI does not stem from a single module, but rather from the collaborative operation of DC-SFM and MSFE within cascaded SSDB modules. The residual connection and skip connection design ensures effective flow of shallow-level de-

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

Case	Index	Noisy	KSVD	BM4D	SERT	DIP-SLR	D2Net	Our
Case 1	PSNR	17.86	26.78	34.73	43.53	42.03	42.13	**44.21**
	SSIM	0.175	0.838	0.881	0.994	0.970	0.989	**0.996**
	SAM	36.211	18.564	4.526	**2.074**	3.192	3.108	2.430
Case 2	PSNR	17.30	26.91	25.95	39.66	35.49	38.95	**42.13**
	SSIM	0.159	0.809	0.869	0.985	0.970	0.972	**0.997**
	SAM	47.384	8.361	6.986	2.539	4.103	3.174	**2.421**
Case 3	PSNR	18.86	26.93	29.06	41.93	38.62	40.54	**42.16**
	SSIM	0.214	0.816	0.845	0.948	0.961	0.973	**0.979**
	SAM	56.924	18.965	17.532	2.462	3.101	2.516	**2.155**
Case 4	PSNR	13.99	28.38	32.44	40.78	36.15	38.97	**41.25**
	SSIM	0.108	0.863	0.841	0.984	0.926	0.959	**0.993**
	SAM	48.04	47.039	27.158	2.836	2.639	2.851	**2.461**



(a) Noisy (b) KSVD (c) BM4D (d) SERT

(e) Clean (f) DIP-SLR (g) D2Net (h) Our

FIGURE 4. The denoising results on the ICVL dataset under Gaussian noise with variance $\sigma = 50$. (a) Noisy. (b) KSVD. (c) BM4D. (d) SERT. (e) Clean. (f) DIP-SLR. (g) D2Net. (h) Our.

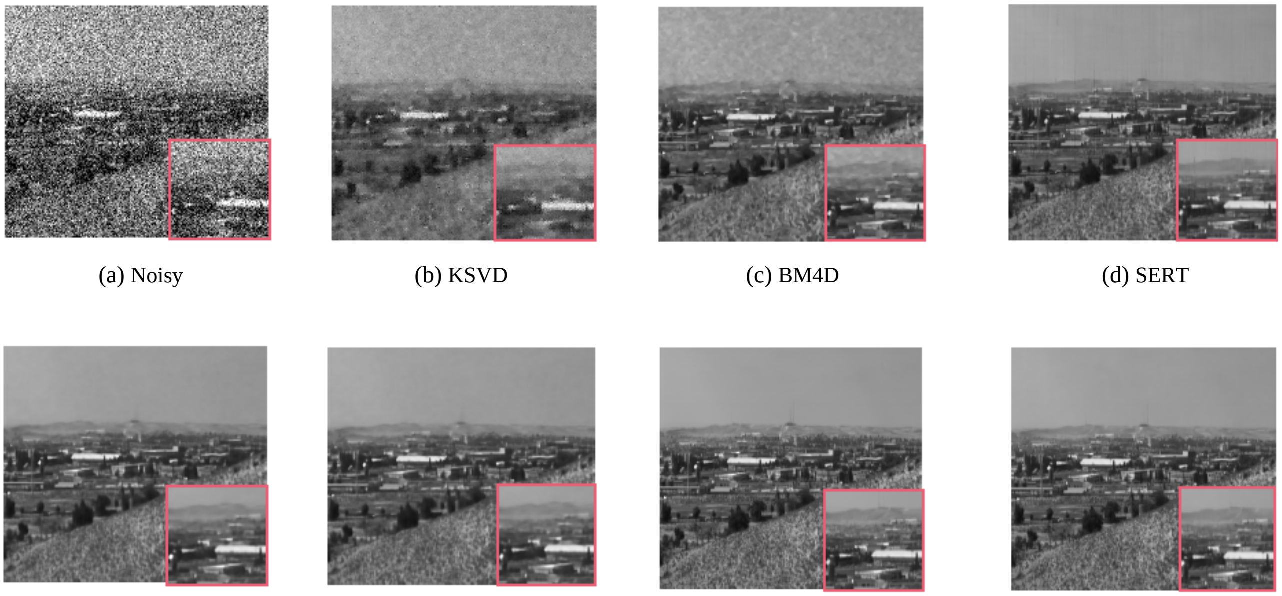
tailed information and deep-level semantic information, preventing information decay in deep networks. Through iterative optimization across multiple SSDB layers, spatial and spectral features undergo deep fusion at different levels, ultimately achieving strong robustness against complex noise — particularly mixed noise, as demonstrated in Case 4.

Furthermore, in order to further quantify the performance of the proposed method, the model complexity is analyzed here. The number of parameters in the proposed SSAT-HSI is 2.97 M, and the giga floating-point operations per second (GFLOPs) is 2671. Compared to the suboptimal model SERT with a parameter count of 1.91M and GFLOPs of 1018.9 [29], our model has a slightly higher complexity, but it achieves better denoising performance than SERT on both Gaussian noise and complex noise.

3.4. Experiments on the Urban Dataset

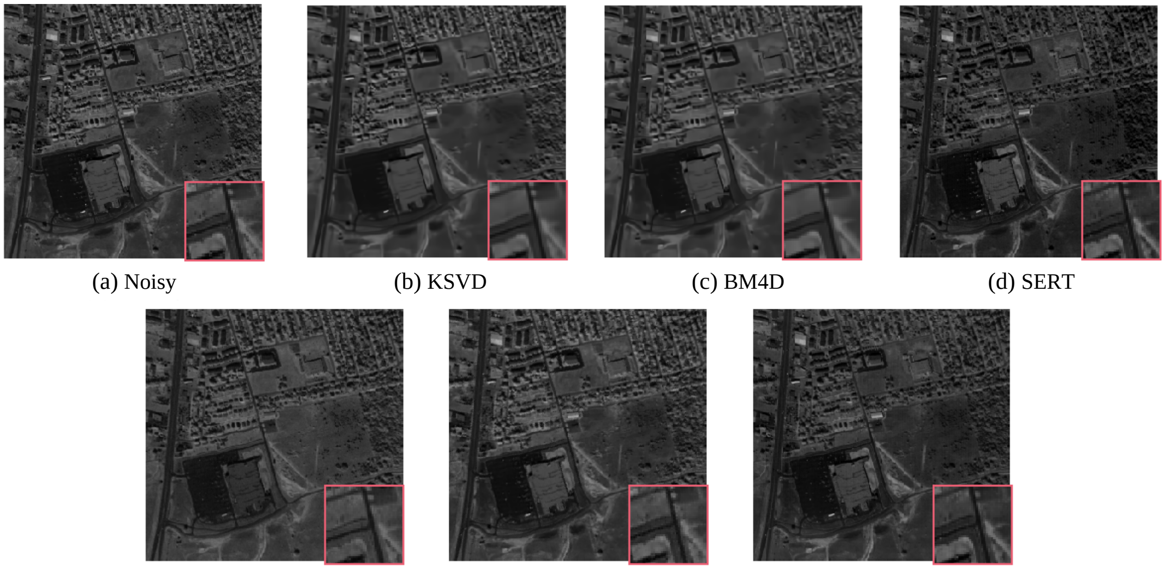
To further validate the effectiveness of the method proposed in this paper, experiments were conducted on the real dataset Urban. The spectral range of the Urban dataset is 0.4–2.5 µm, and the image size of each band is 307 × 307, consisting of a total of 210 bands. This dataset is affected by mixed noises, especially stripe noise.

Due to the lack of training samples that match the Urban dataset in practical scenarios, the model trained on the ICVL dataset is utilized for evaluation. Figure 6 presents the comparison of denoising results on real dataset Urban, using pseudo color images from the 60th band of data. From these results, it can be clearly observed that our method effectively suppresses complex noise in the image, and the resulting denoised image is not only clear, but also successfully preserves the subtle textures and features in the original image. Figure 7 gives the


(a) Noisy (b) KSVD (c) BM4D (d) SERT

(e) Clean (f) DIP-SLR (g) D2Net (h) Our

FIGURE 5. The mixture noise removal results of hyperspectral images obtained by different methods on the ICVL dataset. (a) Noisy. (b) KSVD. (c) BM4D. (d) SERT. (e) Clean. (f) DIP-SLR. (g) D2Net. (h) Our.


(a) Noisy (b) KSVD (c) BM4D (d) SERT

(e) DIP-SLR (f) D2Net (g) Our

FIGURE 6. Denoising results for the real dataset Urban. (a) Noisy. (b) KSVD. (c) BM4D. (d) SERT. (e) DIP-SLR. (f) D2Net. (g) Our.

curves of PSNR and SSIM on the Urban dataset. In Figure 7, through intuitive comparison, it can be clearly observed that the method proposed in this paper demonstrates significant superiority in two key image quality indicators PSNR and SSIM. These results not only reflect the ability of SSAT-HSI in restoring image details and structural integrity, but also confirm the effectiveness of the proposed method.

3.5. Ablation Experiment

To verify the effectiveness of the proposed modules, we conducted ablation experiments on the ICVL dataset and tested the performance of each module under Gaussian noise variance $\sigma = 50$. The results are shown in Table 3. In the experimental analysis, we focused on exploring the functions and performance of each key sub-component in the spatial spectral feature extraction module, especially dual channel spatial feature fusion module DC-SFM and multi-scale spectral feature extraction with self-attention module MSFE. These sub-components play a crucial role in the feature extraction process. Through their collaborative work, they can effectively capture and integrate spatial and spectral information in hyperspectral images, thereby improving the overall performance of the model.

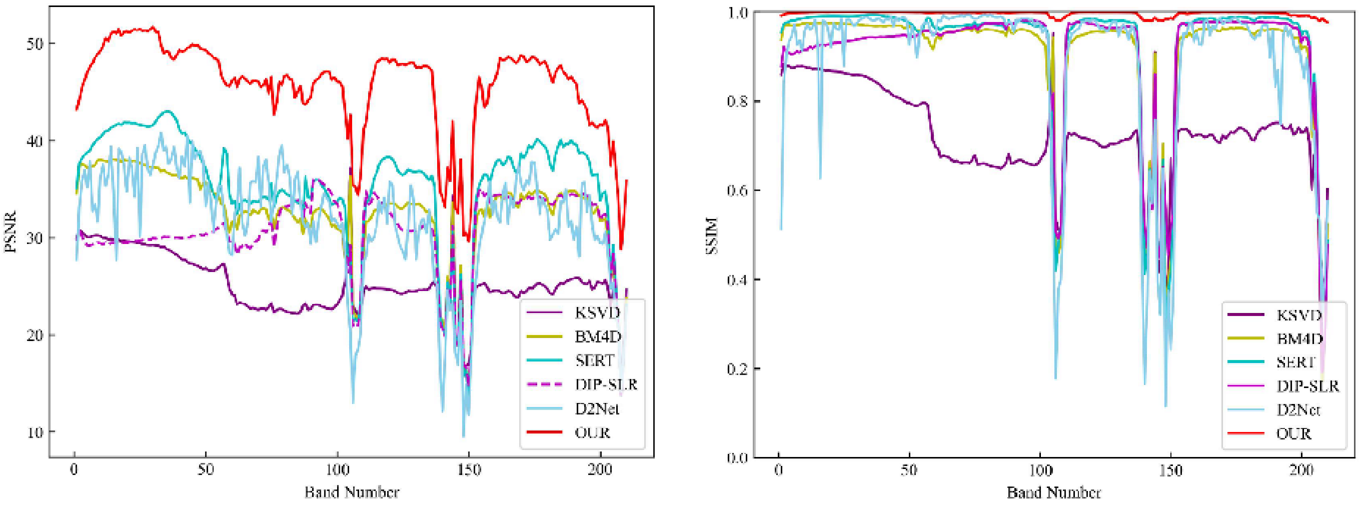


FIGURE 7. The curves of PSNR and SSIM on the Urban dataset.

From Table 3, it can be seen that removing the dual channel spatial feature fusion module DC-SFM leads to a decrease in PSNR of 0.64 dB, while removing the multi-scale spectral feature extraction with the self-attention module MSFE reduces PSNR of 0.92 dB. Therefore, it can be seen that the fusion of spatial and spectral information improves the denoising effect without significant increase in computational cost, fully verifying its effectiveness.

TABLE 3. Ablation experiments for each sub-module.

Baseline	DC-SFM	MSFE	PSNR (dB)	Params (M)
√	×	×	39.45	2.01
√	×	√	39.79	2.55
√	√	×	39.91	2.42
√	√	√	41.43	2.97

4. CONCLUSION

This study constructed a hyperspectral image denoising model based on Transformer spatial spectral attention network. This model mainly consists of a dual channel spatial feature fusion module and a multi-scale spectral feature extraction self-attention module. The dual-channel spatial feature fusion module, through convolution and attention mechanisms, collaborates with local processing units and global processing units to obtain rich local detail features while capturing global features. The multi-scale spectral feature extraction self-attention module analyzes spectral features from a multi-scale perspective, enhances the mining of global correlations in the spectral domain, and improves the model's understanding and processing capabilities of spectral features. Experiments were conducted on the ICVL dataset and Urban real dataset to compare various denoising methods, demonstrating that the method proposed in this paper has good denoising performance.

ACKNOWLEDGEMENT

This work was supported by the 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] Qin, X., Y. Zhang, and Y. Dong, "Domain alignment dynamic spectral and spatial feature fusion for hyperspectral change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 18, 557–568, 2025.

[2] Song, S.-Z., Y.-Y. Liu, Z.-Y. Zhou, X. Teng, J.-H. Li, J.-L. Liu, and X. Gao, "Identification of sorghum breed by hyperspectral image technology," *Spectroscopy and Spectral Analysis*, Vol. 44, No. 5, 1392–1397, 2024.

[3] Lu, B., P. D. Dao, J. Liu, Y. He, and J. Shang, "Recent advances of hyperspectral imaging technology and applications in agriculture," *Remote Sensing*, Vol. 12, No. 16, 2659, 2020.

[4] Zhang, J., Q. Zhang, J. Wang, Y. Wang, and Q. Li, "A dual branch based stitching method for whole slide hyperspectral pathological imaging," *Displays*, Vol. 89, 103090, 2025.

[5] Wu, G., M. A. A. Al-Qaness, D. Al-Alimi, A. Dahou, M. A. Elaziz, and A. A. Ewees, "Hyperspectral image classification using graph convolutional network: A comprehensive review," *Expert Systems with Applications*, Vol. 257, 125106, 2024.

[6] Zhao, X.-L., H. Zhang, T.-X. Jiang, M. K. Ng, and X.-J. Zhang, "Fast algorithm with theoretical guarantees for constrained low-tubal-rank tensor recovery in hyperspectral images denoising," *Neurocomputing*, Vol. 413, 397–409, 2020.

[7] Zhang, A., F. Liu, and R. Du, "Probability-weighted tensor robust PCA with CP decomposition for hyperspectral image restoration," *Signal Processing*, Vol. 209, 109051, 2023.

[8] Zhang, Q., Y. Dong, Q. Yuan, M. Song, and H. Yu, "Combined deep priors with low-rank tensor factorization for hyperspectral image restoration," *IEEE Geoscience and Remote Sensing Letters*, Vol. 20, 1–5, 2023.

[9] 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.

[10] Zha, Z., B. Wen, X. Yuan, J. Zhang, J. Zhou, Y. Lu, and C. Zhu, "Nonlocal structured sparsity regularization modeling for hyperspectral image denoising," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 61, 1–16, 2023.

[11] Wang, P., L. Wang, H. Leung, and G. Zhang, "Super-resolution mapping based on spatial-spectral correlation for spectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 59, No. 3, 2256–2268, 2020.

[12] Zhou, Y., Y. Chen, J. Zeng, W. He, and M. Huang, "Unidirectional spatial and spectral smoothed tensor ring decomposition for hyperspectral image denoising and destriping," *IEEE Geoscience and Remote Sensing Letters*, Vol. 21, 1–5, 2024.

[13] Zhao, S., X. Zhu, D. Liu, F. Xu, Y. Wang, L. Lin, X. Chen, and Q. Yuan, "A hyperspectral image denoising method based on land cover spectral autocorrelation," *International Journal of Applied Earth Observation and Geoinformation*, Vol. 123, 103481, 2023.

[14] 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.

[15] Dabov, K., A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Transactions on Image Processing*, Vol. 16, No. 8, 2080–2095, 2007.

[16] Chang, Y., L. Yan, and S. Zhong, "Hyper-Laplacian regularized unidirectional low-rank tensor recovery for multispectral image denoising," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5901–5909, Honolulu, HI, USA, 2017.

[17] Peng, Y., D. Meng, Z. Xu, C. Gao, Y. Yang, and B. Zhang, "Decomposable nonlocal tensor dictionary learning for multispectral image denoising," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2949–2956, Columbus, OH, USA, 2014.

[18] 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, 2022.

[19] 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.

[20] Meng, P., Z. Xu, X. Wang, W. Yin, and H. Liu, "A novel method for solving the inverse spectral problem with incomplete data," *Journal of Computational and Applied Mathematics*, Vol. 463, 116525, 2025.

[21] Jiang, Y., H. Liu, T. Ni, and K. Zhang, "Inverse problems for nonlinear progressive waves," *Calculus of Variations and Partial Differential Equations*, Vol. 64, No. 4, 116, 2025.

[22] Yin, W., Z. Shen, P. Meng, and H. Liu, "An online interactive physics-informed adversarial network for solving mean field games," *Engineering Analysis with Boundary Elements*, Vol. 169, 106002, 2024.

[23] 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.

[24] 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, 2019.

[25] 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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8084–8093, Long Beach, CA, USA, 2019.

[26] Wei, K., Y. Fu, and H. Huang, "3-D quasi-recurrent neural network for hyperspectral image denoising," *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 32, No. 1, 363–375, 2021.

[27] Maffei, A., J. M. Haut, M. E. Paoletti, J. Plaza, L. Bruzzone, and A. Plaza, "A single model CNN for hyperspectral image denoising," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 58, No. 4, 2516–2529, 2020.

[28] Wang, W., E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 568–578, Montreal, QC, Canada, 2021.

[29] Li, M., J. Liu, Y. Fu, Y. Zhang, and D. Dou, "Spectral enhanced rectangle transformer for hyperspectral image denoising," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5805–5814, Vancouver, BC, Canada, 2023.

[30] Zhang, Q., Y. Dong, Y. Zheng, H. Yu, M. Song, L. Zhang, and Q. Yuan, "Three-dimension spatial-spectral attention transformer for hyperspectral image denoising," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 62, 1–13, 2024.

[31] Wang, Z., X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general U-shaped transformer for image restoration," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17 683–17 693, New Orleans, LA, USA, 2022.

[32] Pang, L., W. Gu, and X. Cao, "TRQ3DNet: A 3D quasi-recurrent and transformer based network for hyperspectral image denoising," *Remote Sensing*, Vol. 14, No. 18, 4598, 2022.

[33] Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," *arXiv:2010.11929*, 2020.

[34] Arad, B. and O. Ben-Shahar, "Sparse recovery of hyperspectral signal from natural RGB images," in *Computer Vision-ECCV 2016: 14th European Conference*, 19–34, Amsterdam, The Netherlands, 2016.

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

[36] Kingma, D. P. and J. Ba, "Adam: A method for stochastic optimization," in *The 3rd International Conference for Learning Representations*, San Diego, USA, 2015.