Hybrid Convolutional and Attention Network for Hyperspectral Image Denoising

Shuai Hu, Feng Gao, *Member, IEEE*, Xiaowei Zhou, Junyu Dong, *Member, IEEE*,
and Qian Du, *Fellow, IEEE*

Abstract—Hyperspectral image (HSI) denoising is critical for the effective analysis and interpretation of hyperspectral data. However, simultaneously modeling global and local features is rarely explored to enhance HSI denoising. In this letter, we propose a hybrid convolution and attention network (HCANet), which leverages both the strengths of convolution neural networks (CNNs) and Transformers. To enhance the modeling of both global and local features, we have devised a convolution and attention fusion module aimed at capturing long-range dependencies and neighborhood spectral correlations. Furthermore, to improve multi-scale information aggregation, we design a multi-scale feed-forward network to enhance denoising performance by extracting features at different scales. Experimental results on mainstream HSI datasets demonstrate the rationality and effectiveness of the proposed HCANet. The proposed model is effective in removing various types of complex noise. Our codes are available at https://github.com/summitgao/HCANet.

Index Terms—Hyperspectral image, image denoising, Transformer, attention mechanism, deep learning.

I. Introduction

HYPERSPECTRAL imaging is a powerful technique that allows the acquisition of rich spectral information from an object or a scene. Compared with RGB data, hyperspectral image (HSI) captures fine-grained spectral information. Hence, HSIs have been extensively used in many practical applications, such as unmixing [1] and ground object classification [2]. However, HSI is often afflicted by inevitable mixed noise generated during the sensor imaging process, which is caused by insufficient exposure time and reflected energy. These noise may degrade the image quality and hinder the performance of subsequent analysis and interpretation. Eliminating these noise could improve the accuracy of ground object detection and classification. Therefore, HSI denoising is a critical and indispensable technique in the preprocessing stage of many remote sensing applications.

Motivated by the spatial and spectral properties of HSI, traditional HSI denoising methods exploit the optimization schemes with priors, such as low rankness [3], total variation [4], non-local similarities [5], and spatial-spectral correlation

[6]. Although these methods have achieved appreciable performance, they commonly depends on the degree of similarity between the handcrafted priors and the real-world noise model. In recent years, convolutional neural networks (CNNs) [7] have provided new ideas for HSI denoising, demonstrating notable performance advancements. Maffei et al. [8] proposed a CNN-based HSI denoising model by taking the noise-level map as input to train the network. Wang et al. [9] proposed a convolutional network based on united Octave and attention mechanism for HSI denoising. Pan et al. [10] presented a progressively multiscale information aggregation network to remove the noise in HSI. These CNN-based methods use convolution kernels for local feature modeling.

More recently, with the emergence of Vision Transformer (ViT) [11], Transformer-based methods have achieved significant success in various computer vision tasks. Existing Transformer-based image denoising methods have achieved great success through learning the global contextual information. However, if local features are considered and exploited effectively, the HSI denoising performance may improve further. So, it is important to take account of both local and global information by combining CNN and Transformers to enhance denoising performance.

It is commonly non-trivial to build an effective Transformer and CNN hybrid model for HSI denoising, due to the following two challenges: *1) The optimal hybrid architecture for local and global feature modeling still remains an open question.* Convolutional kernels capture local features, which means losing the long-distance information interaction. The combination of convolution and attention could offer a viable solution. *2) The single-scale feature aggregation of the feed-forward network (FFN) in Transformer is limited.* Some methods use depth-wise convolution to improve local feature aggregation in FFN. However, due to the larger number of channels in the hidden layer, single-scale token aggregation can hardly exploit rich channel representations.

To solve the aforementioned two challenges, we proposed a <u>H</u>ybrid <u>C</u>onvolution and <u>A</u>ttention <u>Net</u>work (HCANet) for HSI denoising, which simultaneously exploits both the global contextual information and local features, as illustrated in Fig. 1. Specifically, to enhance the modeling of both global and local features, we have devised a convolution and attention fusion module (CAFM) aimed at capturing long-range dependencies and neighborhood spectral correlations. Furthermore, to improve multi-scale information aggregation in FFN, we design a multi-scale feed-forward network (MSFN) to enhance denoising performance by extracting features at different scales.

Shuai Hu, Feng Gao, Xiaowei Zhou, and Junyu Dong are with the School of Computer Science and Technology, Ocean University of China, Qingdao 266100, China.

Qian Du is with the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS 39762 USA.

$H \times W \times B$ $H \times W \times C$ $\frac{H}{2} \times \frac{W}{2} \times 2C$ $\frac{H}{4} \times \frac{W}{4} \times 4C$

Conv

Observed HSI

CAMixing Block ×2

CAMixing Block ×3

CAMixing Block ×4

Clean HSI

Conv

CAMixing Block ×2

CAMixing Block ×3

$H \times W \times C$ $\frac{H}{2} \times \frac{W}{2} \times 2C$

⊕ Element-wise Summation
C Channel-wise Concat
⊗ Matrix Multiplication

(a) Framework of hybrid convolution and attention network (HCANet)
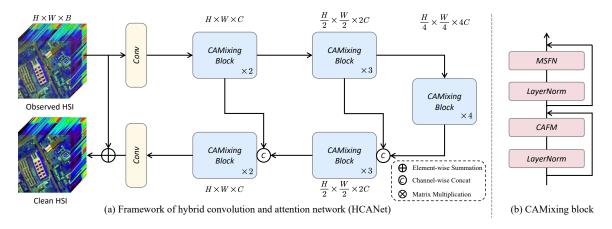
MSFN

LayerNorm

CAFM

LayerNorm

(b) CAMixing block

Fig. 1. Illustration of our proposed hybrid convolution and attention network (HCANet) for HSI denoising. (a) Framework of HCANet. (b) Inner structure of CAMixing block.

Three parallel dilated convolutions with different strides are used in MSFN. By conducting experiments on two real-world datasets, we validate that our proposed HCANet is superior to other state-of-the-art competitors.

The contributions of this letter can be summarized as follows:

- The promising yet challenging problem of global and local feature modeling for HSI denoising is explored. To the best of our knowledge, this is the first work to combine convolution and attention mechanism for the HSI denoising task.
- We propose multi-scale feed-forward network, which seamlessly extract feature at different scales, and effectively suppress the noise from multiple scales.
- Extensive experiments are conducted on two benchmark datasets, which demonstrates the rationality and effectiveness of the proposed HCANet. As a side contribution, we have released our codes to benefit other researchers.

II. METHODOLOGY

In this section, we present the Hybrid Convolution and Attention Network (HCANet) for hyperspectral image denoising. As shown in Fig. 1, the main structure of the model is a U-shaped network with several Convolution Attention Mixing (CAMixing) blocks. Each CAMixing block includes two parts: convolution-attention fusion module (CAFM) and multi-scale feed-forward network (MSFN). For HSI, 3D convolutions capture spatial and spectral features comprehensively, but increase parameters. To manage complexity, we use 2D convolutions for channel adjustment, effectively exploiting HSI features.

For a noisy hyperspectral image $\mathbf{I} \in \mathbb{R}^{H \times W \times B}$, where $H \times W$ denote the spatial resolution and B denotes the channel dimension, our HCANet first uses $3 \times 3 \times 3$ convolution to obtain low-level feature. Then, we use a U-shaped network with several CAMixing blocks and skip connections to obtain the noise residual map $\mathbf{I}_N \in \mathbb{R}^{H \times W \times B}$, which has the same shape as the input noisy image. The reconstructed clean HSI can be expressed as: $\hat{\mathbf{I}} = \mathbf{I} + \mathbf{I}_N$. Finally, the reconstruction loss with a global gradient regularizer is used to train the HCANet.
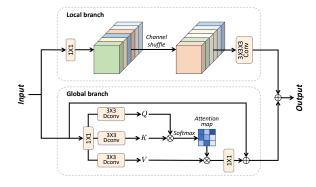

Fig. 2. Illustration of the proposed convolution and attention fusion module (CAFM). It consists of local and global branches. In the local branch, convolution and channel shuffling are employed for local feature extraction. In the global branch, the attention mechanism is used to model long-range feature dependencies.

HSI denoising aims to reconstruct the corresponding clean HSI from a noisy HSI. Fusing global and local features in HSI is important to enhance the denoising task. Thus, the CAFM is designed in the CAMixing block. To further utilize the contextual cues in the feed-forward network for HSI denoising, the MSFN is designed in the CAMixing block. Next, we present the details of the CAFM and MSFN.

A. Convolution and Attention Fusion Module

Convolutional operations, limited by their local nature and restricted sensory field, are insufficient in modeling global features. In contrast, the Transformer, enabled by the attention mechanism, excels in extracting global features and capturing long-range dependencies. Convolution and attention are complementary to each other to model both global and local features. Inspired by this, we design a convolution and attention fusion module (CAFM), as shown in Fig. 2. We employ a self-attention mechanism in the global branch to capture broader hyperspectral data information, while a local branch focuses on extracting local features for comprehensive denoising.

The proposed CAFM consists of local and global branches. In the local branch, to enhance cross-channel interaction and

promote information integration, we first use 1×1 convolution to adjust the channel dimensions. Following this, a channel shuffling operation is performed to further mix and blend the channel information. Channel shuffle partitions the input tensor along the channel dimension into groups, wherein depth-wise separable convolution is employed within each group to induce channel shuffling. Subsequently, the resulting output tensors from each group are concatenated along the channel dimension to generate a novel output tensor. Subsequently, we utilize a $3 \times 3 \times 3$ convolution to extract features. The local branch can be formulated as:

$$\mathbf{F}_{\text{conv}} = W_{3 \times 3 \times 3}(\text{CS}(W_{1 \times 1}(\mathbf{Y}))), \quad (1)$$

where $\mathbf{F}_{\text{conv}}$ is the output of local branch, $W_{1 \times 1}$ denotes 1×1 convolution, $W_{3 \times 3 \times 3}$ denotes $3 \times 3 \times 3$ convolution, CS represents channel shuffle operation, and $\mathbf{Y}$ is the input feature.

In the global branch, we first generates query($\mathbf{Q}$), key($\mathbf{K}$) and value($\mathbf{V}$) via 1×1 convolution and 3×3 depth-wise convolution, yielding three tensors with the shape of $\hat{H} \times \hat{W} \times \hat{C}$. Next, $\mathbf{Q}$ is reshaped to $\hat{\mathbf{Q}} \in \mathbb{R}^{\hat{H}\hat{W} \times \hat{C}}$, and K is reshaped to $\hat{\mathbf{K}} \in \mathbb{R}^{\hat{C} \times \hat{H}\hat{W}}$. Then, we compute the attention map $\mathbf{A} \in \mathbb{R}^{\hat{C} \times \hat{C}}$ via the interaction of $\hat{\mathbf{Q}}$ and $\hat{\mathbf{K}}$. The computational burden is reduced instead of computing the huge regular attention map of size $\mathbb{R}^{\hat{H}\hat{W} \times \hat{H}\hat{W}}$. The output $\mathbf{F}_{\text{att}}$ of global branch is defined as:

$$\mathbf{F}_{\text{att}} = W_{1 \times 1}\text{Attention}\left(\hat{\mathbf{Q}}, \hat{\mathbf{K}}, \hat{\mathbf{V}}\right) + \mathbf{Y}, \quad (2)$$

$$\text{Attention}\left(\hat{\mathbf{Q}}, \hat{\mathbf{K}}, \hat{\mathbf{V}}\right) = \hat{\mathbf{V}}\text{Softmax}\left(\hat{\mathbf{K}}\hat{\mathbf{Q}}/\alpha\right), \quad (3)$$

where α is a learnable scaling parameter to control the magnitude of matrix multiplication of $\hat{\mathbf{K}}$ and $\hat{\mathbf{Q}}$ before applying the softmax function.

Finally, the output of the CAFM module calculation is computed as:

$$\mathbf{F}_{\text{out}} = \mathbf{F}_{\text{att}} + \mathbf{F}_{\text{conv}}. \quad (4)$$

B. Multi-Scale Feed-Forward Network

The original FFN in ViT is composed of two linear layers for single-scale feature aggregation. However, the information contained in single-scale feature aggregation of FFN is limited. To enhance the non-linear feature transformation, we propose a multi-scale feed-forward network (MSFN). After each CAMixing block, the outputs of CAFM are fed into the MSFN to aggregate multi-scale features and enhance the non-linear information transformation. Previous studies have revealed the efficacy of incorporating multi-scale information in image denoising tasks [10].

Details of the MSFN are illustrated in Fig. 3. Two 1×1 convolutions are used to expand the feature channels with expanding ratio $\gamma = 2$. The input features are handled in two parallel paths, and gating mechanism is introduced to enhance the non-linear transformation via element-wise product of features from both paths. In the lower path, depth-wise convolution is used for feature extraction. In the upper
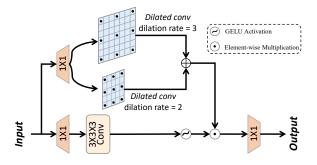

Fig. 3. Illustration of the multi-scale feed-forward network (MSFN).

path, multi-scale dilated convolutions are employed for multi-scale feature extraction. Two 3×3 dilated convolutions with dilation rates of 2 and 3 are used.

Given an input tensor $\mathbf{X} \in \mathbb{R}^{\hat{H} \times \hat{W} \times \hat{C}}$, MSFN is formulated as:

$$\begin{aligned} \text{Gating}(\mathbf{X}) = &\, \phi(W_{3 \times 3 \times 3}W_{1 \times 1}(\mathbf{X})) \\ &\odot (W_{3 \times 3}^{2}W_{1 \times 1}(\mathbf{X}) + W_{3 \times 3}^{3}W_{1 \times 1}(\mathbf{X})), \end{aligned} \quad (5)$$

$$\mathbf{X}_{out} = W_{1 \times 1}\text{Gating}(\mathbf{X}), \quad (6)$$

where $\odot$ denotes element-wise multiplication, ϕ represents the GELU non-linearity, $W_{3 \times 3}^{2}$ denotes 3×3 dilated convolution with dilation rate of 2 and $W_{3 \times 3}^{3}$ denotes 3×3 dilated convolution with dilation rate of 3. MSFN offers a distinct role compared to CAFM, focusing on enriching features with contextual information.

C. Loss Function

We use the L_1 loss function as our optimization objective for training the network, and the reconstruction loss is:

$$\mathcal{L}_{\text{rec}} = \|\hat{\mathbf{I}} - \mathbf{I}\|_1, \quad (7)$$

$\hat{\mathbf{I}}$ represents the estimated noise-free Hyperspectral Image (HSI), while $\mathbf{I}$ denotes the noisy HSI.

To enhance the quality of denoising and curtail redundancy, we incorporate a global gradient regularizer to constrain $\hat{\mathbf{I}}$,

$$\begin{aligned} \mathcal{L}_{grad} = &\, \|\nabla_h\hat{\mathbf{I}} - \nabla_h\mathbf{I}\|_2^2 + \|\nabla_v\hat{\mathbf{I}} - \nabla_v\mathbf{I}\|_2^2 + \\ &\, \|\nabla_s\hat{\mathbf{I}} - \nabla_s\mathbf{I}\|_2^2 \end{aligned} \quad (8)$$

where ∇_h, ∇_v, and ∇_s respectively represent the gradient operator applied along the horizontal, vertical, and spectral axes. Finally, the total loss function is as follows:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda\mathcal{L}_{grad}, \quad (9)$$

where λ is the weight parameter governing $\mathcal{L}_{grad}$. Based on empirical evidence, we set λ as 0.01 to maintain a balance between the various loss terms.

TABLE I
QUANTITATIVE EVALUATION OF THE PROPOSED HCANET AND OTHER COMPARING METHODS IN HSI IMAGE DENOISING TASK WITH FOUR NOISE MAGNITUDES ($\sigma = 30, 50, 70$ AND $blind$).

Case	Index	Noisy	LRMR	BM4D	LRTV	Restormer	MAFNet	Ours
30	PSNR	17.26	31.81	39.71	36.92	42.19	43.95	45.51
	SSIM	0.108	0.678	0.973	0.93	0.971	0.974	0.981
	SAM	0.698	0.182	0.079	0.089	0.05	0.031	0.028
50	PSNR	15.43	30.04	37.22	35.99	41.51	42.13	44.54
	SSIM	0.049	0.634	0.856	0.901	0.952	0.967	0.978
	SAM	0.885	0.222	0.166	0.121	0.047	0.034	0.033
70	PSNR	11.45	25.89	34.09	33.88	39.31	41.05	43.27
	SSIM	0.03	0.565	0.781	0.858	0.947	0.951	0.968
	SAM	1.01	0.275	0.189	0.155	0.041	0.036	0.037
blind	PSNR	14.85	30.07	37.1	37.23	40.73	42.21	43.97
	SSIM	0.056	0.648	0.867	0.924	0.95	0.962	0.978
	SAM	0.857	0.149	0.086	0.114	0.043	0.032	0.034

TABLE II
QUANTITATIVE EVALUATION OF THE PROPOSED HCANET AND OTHER COMPARING METHODS IN HSI IMAGE DENOISING TASK WITH FIVE COMPLEX NOISE CASES.

Case	Index	Noisy	LRMR	BM4D	LRTV	Restormer	MAFNet	Ours
Case1	PSNR	17.79	31.91	35.62	36.89	42.03	43.48	44.11
	SSIM	0.158	0.706	0.884	0.896	0.970	0.972	0.979
	SAM	0.799	0.215	0.146	0.12	0.044	0.037	0.033
Case2	PSNR	17.41	31.07	33.94	35.64	41.13	42.43	43.58
	SSIM	0.193	0.697	0.829	0.882	0.965	0.967	0.978
	SAM	0.808	0.238	0.153	0.182	0.05	0.039	0.036
Case3	PSNR	17.42	30.04	33.77	33.82	40.16	41.74	41.79
	SSIM	0.15	0.74	0.866	0.871	0.961	0.958	0.973
	SAM	0.882	0.242	0.179	0.147	0.052	0.035	0.041
Case4	PSNR	15.12	29.34	32.61	33.25	36.89	37.71	39.99
	SSIM	0.126	0.703	0.887	0.816	0.932	0.945	0.964
	SAM	0.891	0.286	0.193	0.174	0.083	0.077	0.056
Case5	PSNR	14.15	26.9	31.02	30.91	36.15	37.27	40.57
	SSIM	0.107	0.601	0.711	0.743	0.926	0.978	0.963
	SAM	0.912	0.305	0.22	0.198	0.079	0.067	0.052

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. Experiment Setup

Benchmark datasets. To verify the denoising performance of our model on hyperspectral images, we trained our model on ICVL dataset and evaluated the trained model on Pavia dataset. In the ICVL dataset, a total of 31 spectral bands were utilized to collect images with a resolution of 1392×1300. To facilitate the training process, the data were randomly cropped and transformed into cube data with a shape of $128 \times 128 \times 31$. We augmented the training dataset with rotation and scaling techniques to improve the model's robustness, which resulted in a dataset with $20,000$ new samples. To test the denoising effect of our model on real remote sensing images, we conducted experiments on Pavia dataset.

Noise setting. During the testing phase, we conducted experiments under two settings to demonstrate the effectiveness and generalizability of our proposed model. For the first setting, we tested our model with varying magnitudes of Gaussian noise from $\sigma = 30$ to $\sigma = 70$ and blind Gaussian noise (random noise magnitude). For the second setting, we evaluated its robustness against common complex noise types found in hyperspectral data obtained from real spaceborne sensors, such as Gaussian, impulse, and deadline noise. We defined five types of complex noise set:

1) *Case 1: Gaussian noise of varying magnitudes in all spectral channels, with a randomly chosen standard deviation from 30 to 70.*
2) *Case 2: Gaussian + Stripe noise.* In addition to Gaussian noise, we randomly add strip noise to spectral bands by polluting $5\% \sim 15\%$ of columns with strips.
3) *Case 3: Gaussian + Deadline noise.* On the basis of Case 1, we added the deadline noise in the third of the spectral bands. 5% to 15% of columns are conflicted with deadlines in each band.
4) *Case 4: Gaussian + Impulse noise.* On the basis of Case 1. Approximately one third of the spectral bands were chosen at random to increase the intensity of impulse noise by a range of 30% to 70%.
5) *Case 5: Mixture noise.* Similar to the above cases, each spectral band is affected by *non-i.i.d Gaussian noise* in Case 1. Additionally, each band is randomly affected by a combination of three other types of noise.

Baseline methods and Implementation Details. We compared HCANet with five state-of-the-art methods including model-driven methods and deep learning-based methods. For model-driven methods, we consider BM4D [12], as well as low-rank methods such as LRMR [13] and LRTV [14]. In terms of deep learning-based methods, we choose the well-known methods, Restormer [15] and MAFNet [10] for comparison. We used three evaluation metrics including peak signal-to-noise ratio (PSNR), structure similarity (SSIM), and spectral angle mapper (SAM) to quantify the denoising performance. Larger PSNR and SSIM values indicate better denoising results, while smaller SAM values indicate better denoising performance. We trained the models with an initial learning rate of 10^{-4} and the learning rate decreased gradually over training epochs. The HCANet was optimized by the Adam optimizer and it was trained for 100 epochs on Gaussian noise and 150 epochs on complex noise. We conducted all the experiments under PyTorch framework on a machine with an NVIDIA GTX 2080Ti GPU, Intel Xeon E5 CPU and 32GB RAM.

B. Experimental Analysis

A comprehensive quantitative comparison of our proposed HCANet and other benchmarking methods is shown in Table I and II. In the two tables, values highlighted by Red indicate the best while values highlighted by the Blue indicate the second best performance.

Table I showcases the results of all methods in different intensities of Gaussian noise. From this table, it is easy to find that our HCANet outperforms all other methods in terms of all metrics when the data contains single-type noise. Additionally, Table II presents the results with complex noise. We can see that our HCANet achieves the best performance in all noise cases. In addition to quantitative analysis, we also conducted qualitative comparisons, as illustrated in Figs. 4 and 5. False-color denoising results were obtained from three bands (17, 20, 30). It is clear that traditional denoising methods struggle with complex noise. While MAFNet removes most noise, it tends to oversmooth, losing some image details. Restormer performs poorly on complex noise. In contrast, HCANet effectively

TABLE III
ABLATION STUDIES ON ICVL DATASETS UNDER NOISE LEVEL
$\sigma = 30$, EPOCH = 50

Base model	Local branch	3D Conv	MSFN	PSNR ↑	SSIM ↑	SAM ↓
✓				39.94	0.965	0.045
✓	✓			40.58	0.967	0.039
✓	✓	✓		42.23	0.975	0.037
✓	✓	✓	✓	**42.68**	**0.976**	**0.035**

removes most noise while preserving local details, successfully restoring original image features.

To demonstrate the effectiveness of each component in our HCANet, we conducted ablation study on ICVL dataset. As shown in III, the HCANet (with all components) performs the best compared with all other variants. Furthermore, we find the performance constantly becomes better as we add local branch, 3D convolution, and MSFN upon the base model. It indicates the necessity of each component in our proposed HCANet.
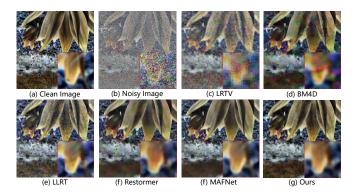

Fig. 4. Gaussian noise removal results under noise level $\sigma = 50$ on ICVL dataset with bands (17,20,30).
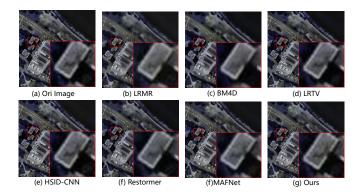

Fig. 5. Real noise removal results on Pavia dataset with bands (17, 20, 30).

TABLE IV
COMPARATIVE ANALYSIS OF COMPUTATIONAL COMPLEXITY OF
MODELS UNDER BLIND GAUSSIAN NOISE, EPOCH = 50

Model	Params	GFLOPS	PSNR
MemNet	1.98M	32.41	35.60
HSID-CNN	0.63M	40.80	36.72
MAFNet	20.17M	20.15	40.51
Restormer	14.88M	9.55	40.27
Ours	4.75M	11.5	41.21

We also compared the computational complexity of the models, and the results are shown in Table IV. HCANet achieves optimal denoising performance while maintaining a relatively moderate number of parameters and computational complexity.

IV. CONCLUSION

In this letter, we propose HCANet, a novel network for HSI denoising. In particular, we proposed the convolution and attention fusion module, CAFM, to fuse both global and local features. Furthermore, we propose the multi-scale feed-forward network, MSFN to extract features from multiple scales and enhances the denoising performance. Experimental results on challenging HSI datasets demonstrate the effectiveness of our proposed model in comparison to the state-of-the-art HSI denoising methods. Our model achieves remarkable denoising performance in terms of both quantitative metrics and visual quality of the reconstructed images.

REFERENCES

[1] L. Qi, Z. Chen, F. Gao, J. Dong, X. Gao, and Q. Du, "Multiview spatial–spectral two-stream network for hyperspectral image unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023.

[2] J. Lin, F. Gao, X. Shi, J. Dong, and Q. Du, "SS-MAE: Spatial–spectral masked autoencoder for multisource remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, 2023.

[3] Y.-Q. Zhao and J. Yang, "Hyperspectral image denoising via sparse representation and low-rank constraint," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 296–308, 2015.

[4] Q. Yuan, L. Zhang, and H. Shen, "Hyperspectral image denoising employing a spectral–spatial adaptive total variation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 10, pp. 3660–3677, 2012.

[5] W. He, Q. Yao, C. Li, N. Yokoya, and Q. Zhao, "Non-local meets global: An integrated paradigm for hyperspectral denoising," in *IEEE CVPR*, 2019, pp. 6861–6870.

[6] Y. Peng, D. Meng, Z. Xu, C. Gao, Y. Yang, and B. Zhang, "Decomposable nonlocal tensor dictionary learning for multispectral image denoising," in *IEEE CVPR*, 2014, pp. 2949–2956.

[7] W. Xie and Y. Li, "Hyperspectral imagery denoising by deep learning with trainable nonlinearity function," *IEEE Geosci Remote Sens. Lett.*, vol. 14, no. 11, pp. 1963–1967, 2017.

[8] A. Maffei, J. M. Haut, M. E. Paoletti, J. Plaza, L. Bruzzone, and A. Plaza, "A single model CNN for hyperspectral image denoising," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2516–2529, 2020.

[9] S. Wang, L. Li, X. Li, J. Zhang, X. Zhao, X. Su, and F. Chen, "A denoising network based on frequency-spectral- spatial-feature for hyperspectral image," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 16, pp. 6693–6710, 2023.

[10] H. Pan, F. Gao, J. Dong, and Q. Du, "Multiscale adaptive fusion network for hyperspectral image denoising," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 16, pp. 3045–3059, 2023.

[11] A. Dosovitskiy, 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 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021, pp. 1–21.

[12] M. Maggioni, V. Katkovnik, K. Egiazarian, and A. Foi, "Nonlocal transform-domain filter for volumetric data denoising and reconstruction," *IEEE Tran. Image Process.*, vol. 22, no. 1, pp. 119–133, 2013.

[13] H. Zhang, W. He, L. Zhang, H. Shen, and Q. Yuan, "Hyperspectral image restoration using low-rank matrix recovery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 4729–4743, 2014.

[14] W. He, H. Zhang, L. Zhang, and H. Shen, "Total-variation-regularized low-rank matrix factorization for hyperspectral image restoration," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 178–188, 2016.

[15] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *IEEE CVPR*, 2022.