Low-Rank Prompt-Guided Transformer for Hyperspectral Image Denoising

Xiaodong Tan, Mingwen Shao, *Member, IEEE*, Yuanjian Qiao, Tiyao Liu, and Xiangyong Cao

***Abstract*— Hyperspectral image (HSI) denoising is an essential preprocessing step for downstream applications. Although vision transformer (ViT)-based approaches show impressive denoising performance through self-similarity modeling, these methods still fail to exploit spatial and spectral correlations while ensuring flexibility and efficacy. To address this issue, we propose a hyperspectral denoising transformer using low-rank prompt (HyLoRa), simultaneously taking the spatial self-similarity and spectral low-rank property into account for HSI denoising. Specifically, to fully utilize intrinsic similarity in spatial domain, we perform cross-shaped window-based spatial self-attention for effectively modeling local and global similarity. Moreover, to exploit low-rank inductive bias, we integrate a low-rank prompt module into attention calculation for counting corrected low-dimensional vectors from a large collection of HSIs. This helps to better refine underlying noise-free structure representations. Compared to existing works, powerful capabilities for modeling spatial and spectral correlations can be built to correct low-rank representation in the feature space. Extensive experiments on both simulated and real remote sensing noise demonstrate that our HyLoRa consistently surpasses the state-of-the-art methods.**

***Index Terms*— Hyperspectral image (HSI) denoising, low-rank representation, prompt learning, transformer.**

I. Introduction

A HYPERSPECTRAL imaging sensor captures electromagnetic energy within its field of view using hundreds or even thousands of spectral channels, offering a finer spectral resolution compared to multispectral or RGB cameras [1]. This enhanced spectral resolution allows for precise identification of materials and objects within the captured imagery [2]. Consequently, hyperspectral imaging finds applications in diverse fields, including Earth remote sensing (for tasks like object classification [3], [4], [5], landcover change detection [6], and anomaly identification [7]), as well as in agriculture [8], and surveillance [9]. Nonetheless, the narrower spectral bands in hyperspectral cameras result in reduced photon intake, leading to images with a lower signal-to-noise ratio (SNR). This decrease in SNR diminishes the reliability of features measured or information extracted from HSIs. Therefore, HSIs denoising is a crucial preprocessing step before proceeding with further applications [10].

Similar to RGB images, self-similarity exists in the spatial domain of HSIs. This allows for the combination of similar pixels throughout the entire image to remove noise simultaneously [11]. Furthermore, because hyperspectral imaging systems capture images with a high spectral resolution, HSIs possess intrinsic correlations between bands in the spectral domain [12]. Therefore, devising denoising techniques for HSIs should take both spatial and spectral domains into account.

Traditional model-based HSI denoising methods rely on manually crafted priors to explore these spatial self-similarity and spectral correlation by iteratively solving optimization problems. These priors often include total variation [23], [24], non-local similarity [11], [16], [25], low-rank properties [26], [27], [28], [29], and sparsity [30] regularizations. Albeit these methods deliver commendable performance, their effectiveness heavily relies on how closely the manually crafted priors align with the actual noise model in real-world scenarios. Moreover, these methods can be difficult to optimize for modern hardware due to their intricate processing pipelines.

Recently, deep learning methods have been applied to HSI denoising, achieving impressive restoration quality [14], [15], [18], [19], [20], [31], [32]. However, the majority of these studies employ convolutional neural networks (CNNs) for feature analysis and rely on filter response to distinguish noise from signal within a confined window-shaped receptive area [33], [34]. These methods, thus, can exploit spectral correlation instead of global spatial similarity, resulting suboptimal denoising performance.

There are also many approaches that combine model-based and deep learning methods [19], [22], [35], [36], [37]. Most of these methods apply deep learning models to each iteration of the optimization algorithm. That is, the network maps the features of the iterative process, instead of original image. For instance, RCILD [37] uses a network to predict clean representative coefficient image (RCI) throughout the alternating optimization. This fashion eliminates the need to use other time-consuming iterative algorithms to solve the optimization problem. Yet, the networks used are mostly CNNs, which

Manuscript received 9 January 2024; revised 12 March 2024, 23 April 2024, and 4 June 2024; accepted 10 June 2024. Date of publication 14 June 2024; date of current version 27 June 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFA1000102; in part by the National Natural Science Foundation of China under Grant 62376285 and Grant 61673396; and in part by the Natural Science Foundation of Shandong Province, China, under Grant ZR2022MF260. *(Corresponding author: Mingwen Shao.)*

Xiaodong Tan, Mingwen Shao, Yuanjian Qiao, and Tiyao Liu are with Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao, Shandong 266580, China (e-mail: reyes.tan@foxmail.com; smw278@126.com; yjqiao@s.upc.edu.cn; Z22070050@s.upc.edu.cn).

Xiangyong Cao is with the School of Computer Science and Technology and the Ministry of Education Key Laboratory for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: caoxiangyong@xjtu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2024.3414956

suffer from the difficulty of modeling global spatial self-similarity.

As another line of deep learning networks, vision transformers (ViTs) have emerged as a promising approach, achieving competitive results in both high-level and low-level vision tasks [21], [38], [39], [40], [41]. They demonstrate a strong capability for capturing long-range dependencies within image regions. As their core component, self-attention interacts fully with the entire range of features in the image. However, this leads to a quadratic increase in the computational complexity of the network as the image size grows. Several studies have explored efficient spatial attention strategies. For instance, the Swin Transformer [42] divides a single feature map into multiple square windows of equal size and performs attention only inside the windows.

And CSWin Transformer [39] introduces a stripe window to expand the attention area. These approaches reduce computational complexity by sacrificing the ability to model global similarity. Nevertheless, these methods usually use inputs with small resolution to further minimize the complexity, which in turn limits the application to hyperspectral images (HSIs). Given the typically HSI with large resolution, it thus becomes challenging to efficiently model non-local spatial similarity beyond noisy pixels without incurring unnecessary computational burden. Therefore, it remains more challenging to find a transformer design that strikes the balance between computational burden and modeling self-similarity performance.

On the one hand, transformer is able to model spatial self-similarity well relying on a large receptive field, but on the other hand, the model has very little inductive bias regarding HSIs. Transformer therefore usually needs more training data to get better performance. The high spatial similarity and spectral correlation within HSIs suggest that the data exhibits a low-rank structure [43]. That is to say, a clean HSI can be recovered from a low-rank structure extracted from the corresponding degraded HSI. However, to utilize this property in either traditional or deep learning methods, it is necessary to go through some matrix factorization and tensor decomposition methods, which entails an additional computational burden [44], [45]. Hence, it is also worth investigating the way to introduce low-rank inductive bias in the transformer.

In the field of natural language processing (NLP), prompt learning techniques are used to supply contextual information to fine-tune models on a specific task [46], [47], [48]. The concept of prompting involves adding instructional text to the beginning of input data, enabling pre-trained language models (LMs) to better comprehend various downstream tasks. Certain studies have introduced the idea of treating these prompts as continuous tensors tailored to the task and optimizing them directly using gradient-based methods during fine-tuning, which is referred to as prompt tuning.

However, in the field of computer vision, the relationship between image pixels is not a language contextual relationship. Therefore, the form of applying this approach in tasks of high-level and low-level vision [49], [50], [51] has been changed. For instance, PromptIR [50] is trained on diverse degraded images to encode degradation-specific information into prompt. Then, at inference stage, the prompt can help restoration network adaptively repair impaired images. Since prompt can encode the features of different types of degradation, we believe that it can also encode low-rank priors. The prompt approach is hence conjectured to be a suitable way to introduce a low-rank induction bias.

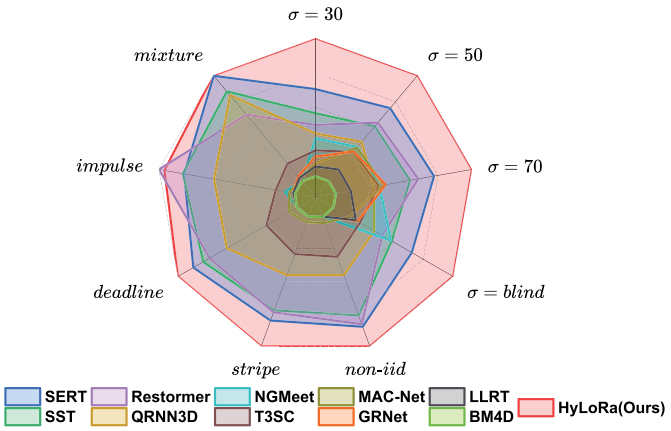

Fig. 1. Our HyLoRa achieves favorable PSNR on a broad set of denoising tasks on ICVL [13] dataset in comparison to other state-of-the-art methods (SERT [14], SST [15]) and other baselines (BM4D [16], LLRT [17], NGMeet [11], GRNet [18], MAC-Net [19], QRNN3D [20], Restormer [21], T3SC [22]).

As mentioned above, there are still huge challenges for hyperspectral denoising using the transformer architecture. In response to these issues, in this study, we propose a hyperspectral denoising transformer using low-rank prompt (HyLoRa), which combines the spatial self-similarity and the low-rank inductive bias of HSI. Unlike current works, the low-rank structure is not directly available outside the model, but is recovered from the feature extracted with the prompt correction. This ensures both the computational complexity of the network and the denoising performance of HSIs. The main contributions of this study can be summarized as follows.

1) We propose HyLoRa, a prompt-based hyperspectral denoising framework, which fully introduces the low-rank inductive bias of HSIs into the transformer with enhanced spatial similarity modeling.
2) Our prompted low-rank transformer block (PLRTB) utilizes cross-shaped window-based self-attention that can capture sufficient spatial similarity information. This block contains a low-rank prompt module, which can generate HSI's inductive bias information to guide attention calculation to effectively remove the corruption from the input image.
3) Our extensive experiments and network analysis showcase the benefits of HyLoRa, as it achieves SOTA performance across a range of noise degradation scenarios (as shown in Fig. 1).

II. Related Work

This section briefly introduces recent works related to the methodology presented in this study, i.e., ViT and prompt learning.

A. Vision Transformer

ViTs have been extensively used in various visual tasks due to their powerful ability to capture long range information

[42], [52], [53]. ViTs also have been exploited for HSIs processing, including classification [2], anomaly detection [54], image fusion [55], [56], and various restoration [57], [58]. The core of the transformer lies in inter-feature self-attention computation. When self-attention is performed in the spatial dimension, features in the full image are fully interacted. However, this leads to a quadratic increase in the computational complexity of the model as the image size increases, which is intolerable for HSI processing. To alleviate this problem, some studies have proposed window-based self-attention [39], [42], [59]. For instance, Swin Transformer [42] divides the image equally into multiple small patches in the spatial dimension and then performs self-attention inside the small patches. The efficiency of computation now depends on the size of patches rather than the entire image. However, this cripples the ability to acquire global information from long distances that exceed the patch size. To recover the global modeling capability, VSA [59] considers utilizing windows of diverse sizes to adaptively interact with features at different spatial distances. Nevertheless, feature maps with different sizes involved in self-attention prevent data parallelism, which further reduces the speed of model training and inference. To summarize, if self-attention is performed only inside the window, the model loses global information. Therefore, we can increase the interaction between different windows to ensure the full utilization of long-distance information to enhance global modeling capabilities.

Meanwhile, transformer applied to HSI suffers from inadequate and inefficient inductive bias regarding HSI. For example, Wang et al. [54] just treat the transformer as a nonlinear regressor without considering any hyperspectral inductive biases. Long et al. [58] merely use window self-attention to reduce computational complexity and leave the super-resolution to just a single upsampling layer. Meanwhile, although some works [15], [55] take into account spatial similarity and spectral correlation, the self-attention of both greatly increases the computational burden. Unlike these methods, we consider spatial similarity and spectral correlation through low-rank property. First, our cross-shaped window-based self-attention is considered to enhance the modeling of similarity in spatial space. Second, to reduce the computation complexity, we utilize the prompt-based method to refine noise-free low-rank structure inside our efficient window-based self-attention. Such an approach balances the model's computational complexity and denoising performance to some extent.

B. Prompt Learning

Prompt originally refers to the prepending language instruction of the input text, allowing a pre-trained LM to better comprehend specific tasks [49]. By opting a suitable prompt, the LM can be generalized to hit new tasks directly without training. Some studies have introduced the idea of treating these prompts as continuous variables tailored to the task and optimizing them directly using gradient-based methods during fine-tuning, which is referred to prompt learning [46], [47], [48]. In the field of computer vision, to accommodate the notion that images have no context information, prompt is directly involved in the training of the network as optimizable variables [49]. This approach, presented in various applicable forms, has already produced fairly positive outcomes in both high-level and low-level vision tasks [49], [50], [51].

The prompt technique has also been applied to image restoration tasks [50], [60], [61]. And it is found that the optimized prompt gains some corresponding degradation adaptive information. In TransWeather [60], prompt as queries for different severe weather are optimized by doing cross-attention operations with the model's encoder. Moreover, AWRCP [61] obtains a high-quality prompt through the pre-training process and uses this prompt directly during the subsequent restoration tasks to enhance the performance. Similarly, PromptIR [50] finds that restoration of diverse degraded images can be accomplished by training the network and prompt end-to-end directly on the degraded images.

Considering this gain effect that prompt plays in image restoration, we decided to use this approach to introduce inductive bias in HSIs. Thus, in this study, we craft a low-rank prompt guiding transformer to achieve better HSI denoising performance. During the training stage, the low-rank prompt counts the noise-free low-rank structural correction information from many HSIs in a low-dimensional space as the network is optimized. This prompt is then used in the inference phase to guide self-attention to correct the noise-removed feature maps.

III. Method

We denote a clean HSI tensor by $\mathcal{H} \in \mathbb{R}^{B\times M\times N}$, where $M \times N$ and B represent the spatial size and spectral bands, respectively. In this study, the additive noise model is the sole focus of our attention, and therefore can be formulated as

$$\mathcal{X} = \mathcal{H} + \mathcal{N} \tag{1}$$

where $\mathcal{X} \in \mathbb{R}^{B\times M\times N}$ is the noised HSI obtained by adding diverse noise $\mathcal{N} \in \mathbb{R}^{B\times M\times N}$ to a clean $\mathcal{H}$. There are many types of $\mathcal{N}$ such as Gaussian noise, stripe noise, impulse noise, deadline noise, and their mixtures. We need to perfect a network to fit the mapping from $\mathcal{X}$ to $\mathcal{H}$.

A. Network Architecture

As shown in Fig. 2, HyLoRa consists of several prompted restore stages (PRSs) between convolution layers on both sides of the start and end. According to [62], using convolution early can improve the performance of transformer in vision tasks. Therefore, given a noisy HSI input $\mathcal{H}$, we first use a 3×3 convolution layer $\text{Conv}_{\text{start}}(\cdot)$ to map the B spectral space to C channels feature space $F_0 \in \mathbb{R}^{H\times W\times C}$ as

$$F_0 = \text{Conv}_{\text{start}}(\mathcal{H}). \tag{2}$$

Then, we extract restored feature $F_{\text{RF}} \in \mathbb{R}^{H\times W\times C}$ from F_0 as

$$F_{\text{RF}} = M_{\text{RF}}(F_0) \tag{3}$$

where $M_{\text{RF}}(\cdot)$ is the restore feature module and it contains K PRSs. Specifically, intermediate features $F_1, F_2, \ldots F_K$ are

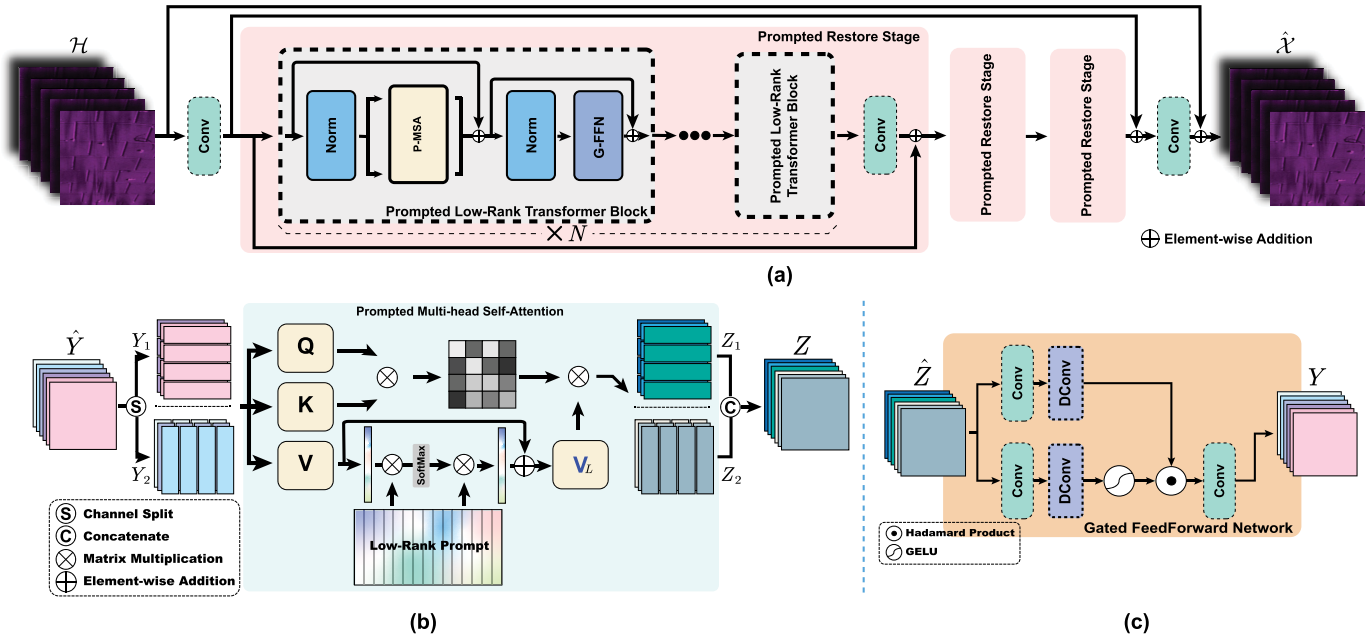

Fig. 2. Illustration of our HyLoRa. (a) Overall architecture of HyLoRa consists of several PRSs. Each PRS is composed of a sequence of PLRTBs. (b) P-MSA in PLRTB. (c) G-FFN.

derived stage by stage as

$$F_i = \mathrm{PRS}_i(F_{i-1}) \tag{4}$$

where $\mathrm{PRS}_i(\cdot)$ is the ith PRS. Meantime, we assign the last feature output F_K to F_{RF}. And during a single PRS, features are processed by a series of PLRTBs which we will explain their components in Section III-B. F_{RF} is then passed to another 3×3 convolution layer $\mathrm{Conv}_{\mathrm{end}}(\cdot)$ mapping from feature space to spectral space as

$$O = \mathrm{Conv}_{\mathrm{end}}(F_{\mathrm{RF}}) \tag{5}$$

where $O \in \mathbb{R}^{H\times W\times B}$ is the predicted residual [63] between a clean HSI and noisy. At last, we get the predicted clean HSI $\hat{\mathcal{X}}$ through

$$\hat{\mathcal{X}} = O + \mathcal{H}. \tag{6}$$

B. Prompted Multihead Self-Attention

We first make full use of spatial similarity, which means that some similar pixels can simultaneously be removed from the noise. Therefore, to remove the noise of individual pixels, it is necessary to obtain a similar signal from the entire image to complement the noisy signal. Consequently, we adopt a self-attention mechanism similar to CSWin [39] with local interaction inside the window in different directions and a global interaction at the end. The following outlines the specific process.

Let $\hat{Y} \in \mathbb{R}^{H\times W\times C}$ be the feature after the first norm layer in PLRTB. Then the whole self-attention process can be formalized as

$$Y_1, Y_2 = \mathrm{ChannelSplit}(\hat{Y}) \tag{7}$$

$$Z_1 = \text{HP-MSA}(Y_1),\ Z_2 = \text{VP-MSA}(Y_2) \tag{8}$$

$$Z = \mathrm{Concatenate}(Z_1, Z_2) \tag{9}$$

where HP-MSA and VP-MSA denote the horizontal and vertical prompted multihead self-attention (P-MSA), respectively. During channel split, we divide the inputs equally into two groups along the channel dimension, i.e.,

$$Y_1 = \hat{Y}[:, :, :C//2] \tag{10}$$

$$Y_2 = \hat{Y}[:, :, C//2:] \tag{11}$$

in which // is the exact division symbol, and [:] denotes the slice operator. Y_1, Y_2 are then calculated by P-MSA in horizontal and vertical directions, respectively. This process can be demonstrated in Fig. 3. The two groups of the input feature maps are divided into equal-sized windows horizontally and vertically. Within the window, the red pixel interacts with all other blue pixels, including itself, through self-attention. The outputs are then concatenated and fused to achieve inter-window interaction.

Here, we take the horizontal one as an instance to formalize. After channel splitting, Y_1's shape is $H \times W \times (C//2)$. Supposing the size of horizontal window as $[h, W]$ and $h \ll H$, Y_1 is partitioned into non-overlapping windows as $[Y_1^1, Y_1^2, \ldots, Y_1^N]$ where $Y_1^i \in \mathbb{R}^{h\times W\times (C//2)}$ and $N = (H/h)$.

Unlike SERT [14], we introduce inductive bias of low rank directly into the standard MSA as follows:

$$Q_1^i = Y_1^i W_1^q,\ K_1^i = Y_1^i W_1^k,\ V_1^i = Y_1^i W_1^v \tag{12}$$

$$V_L{}_1^i = \mathrm{SoftMax}\left(V_1^i W_R P\right) P^T W_R^{\dagger} + V_1^i \tag{13}$$

$$Z_1^i = \mathrm{SoftMax}\left(Q_1^i K_1^{i\,T}/\sqrt{d} + \mathrm{Pos}\right) V_L{}_1^i \tag{14}$$

where $W_1^q, W_1^k, W_1^v \in \mathbb{R}^{(C//2)\times(C//2)}$ are the linear transform matrices of query $Q_1^i \in \mathbb{R}^{h\times W\times(C//2)}$, key $K_1^i \in \mathbb{R}^{h\times W\times(C//2)}$, and value $V_1^i \in \mathbb{R}^{h\times W\times(C//2)}$. $W_R \in \mathbb{R}^{(C//2)\times R}$ is the projection mapping original value to low-dimensional vector. And $P \in \mathbb{R}^{R\times L}$ is our low-rank prompt. $W_R^{\dagger} \in \mathbb{R}^{R\times(C//2)}$ is the inverse projection mapping prompted vector to value space. Pos is the

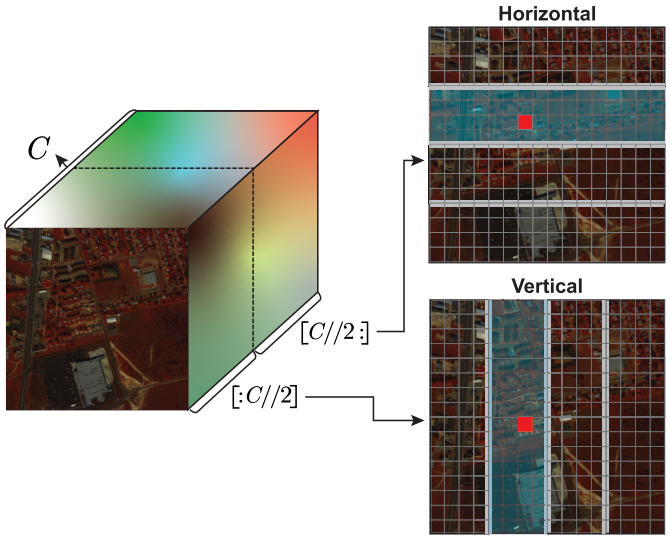

Fig. 3. Illustration of the cross-shape window-based self-attention. The input feature map is divided into two parts along the spectral direction for horizontal and vertical window attention, respectively. This creates a cross-form fusion of the entire image information.

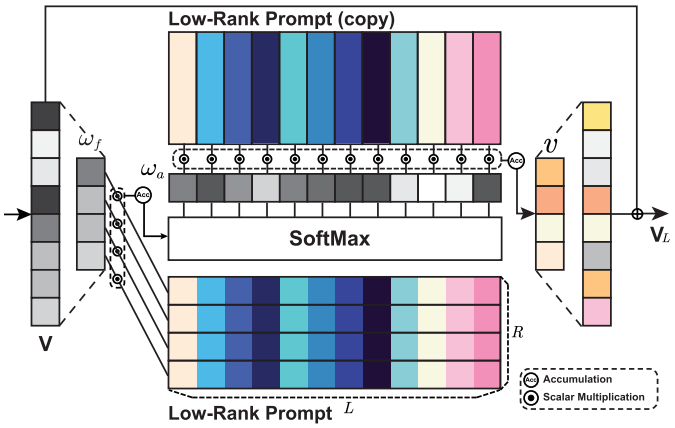

Fig. 4. Illustration of the low-rank prompt. The correction vector for the low-rank structure is obtained using input V through multiweighted summation with low-rank prompt. This facilitates the removal of noise in feature maps.

learnable position encoding. And d denotes the dimension of feature, which in our design is equal to $C//2$.

For a transformer for image restoration, both its intermediate features and the final clean image are obtained from linear combination of the values in self-attention. Due to spectral correlation, an HSI should exist in a low-rank space, which means that an image can be a linear combination of a set of vectors in this space. This happens to match the output form of the self-attention. So, we introduce low-rank prompt to all values when we get the output of self-attention for improving feature representation. Our prompt matrix constrains its rank by defining its shape (the rank of the matrix is less than or equal to any of the dimension sizes). Thus, with supervised training, prompt takes on a complementary role to values for attention in low-rank space.

The interaction between features and prompts in the network is defined by (13). And our intentions can be demonstrated in Fig. 4. To prevent confusion, we omit all the superscripts and subscripts of V mentioned above in this figure. First, each vector in V is mapped to a fusion weight ω_f in the low-dimensional space. This weight is then used by prompt P to obtain another assignment weight ω_a of length L through a weighted linear accumulation in the R direction. Subsequently, the prompt will be exploited again from the another L direction. At this point, the L vectors of the prompt will be mixed under assignment weight ω_a to obtain complementary information vector v. And then we map this vector back to the space that V lies in. The corrected V_L is eventually obtained by adding the original V to this complement. Then the specific formalization is as follows:

$$\omega_f = VW_R \tag{15}$$

$$\omega_a = \text{SoftMax}\left(\sum_{i=1}^{R} \omega_{f_i} P_i\right) \tag{16}$$

$$v = \sum_{j=1}^{L} \omega_{a_j} P_j^T \tag{17}$$

$$V_L = vW_R^{\dagger} \tag{18}$$

where P^T is a copy of transposed P. In order to simplify these weighted summations, we reduced these equations to (13) in matrix multiplication form. Intuitively, the low-rank prompt optimizes over the dataset through the above operations, capturing the predominant feature patterns of a large number of HSIs in the low-dimensional space. This aids in the recovery of a noise-free structure.

Lastly, the output of HP-WSA can be formalized as

$$\text{HP-MSA}(Y_1) = \text{Merge}\left(Z_1^1, Z_1^2, \ldots, Z_1^N\right). \tag{19}$$

Similar to the horizontal one, the vertical P-MSA partitions input along the vertical direction and gets output with the same shape of the horizontal one. The two groups, at last, are concatenated together as the whole P-MSA's output.

C. Gated Feedforward Network

The output of P-MSA serves as input to FFN, improving the representation. Consequently, it is important to devise a competent FFN that refines feature maps aiding to reconstruct a precise and clear HSI. In this study, we propose two essential alterations to FFN with the aim of enhancing representation learning: 1) the introduction of a gating mechanism and 2) the incorporation of point-wise and depth-wise convolutions. The architecture of our gated feedforward network (G-FFN) is shown in Fig. 2(c). The gating mechanism is defined as the Hadamard product of two up and down parallel branches. These two branches contain stacks of point-wise and depth-wise convolutions to extract features across the spectral and spatial dimensions, respectively. The output of the upper branch is then used as element-wise weights to dynamically regulate the information flow of each element in the output from the lower branch. As in [21], we use convolution layers to fuse the gate's output from different dimensions. Given the input $\hat{Z} \in \mathbb{R}^{H\times W\times C}$, the G-FFN can be formulated as follows:

$$\hat{Z}_1 = \text{Conv}_{\text{depth}}^1\left(\text{Conv}_{\text{point}}^1(\hat{Z})\right) \tag{20}$$

$$\hat{Z}_2 = \text{GELU}\left(\text{Conv}^2_{\text{depth}}(\text{Conv}^2_{\text{point}}(\hat{Z}))\right) \quad (21)$$

$$Y = \text{Conv}_{\text{point}}\left(\hat{Z}_1 \odot \hat{Z}_2\right) \quad (22)$$

where $\text{Conv}^i_{\text{point}}$ is a 1×1 convolution, $\text{Conv}^i_{\text{depth}}$ is a 3×3 depth-wise convolution, and $\odot$ represents the Hadamard product. Intuitively, G-FFN can get an attentional weight based on the current input, thus selectively inflating the focus of the extracted features.

IV. Experiment

In this section, we begin by assessing our approach through synthetic experiments, encompassing scenarios involving Gaussian noise and complex instances. Following that, a real remote sensing dataset is used to evaluate the practicality of our method in real-world scenarios. Finally, ablation experiments are performed for module analysis to confirm the effectiveness of the proposed model.

A. Experimental Settings

1) Simulated Noise Datasets: Simulated experiments are carried out using the ICVL [13] dataset, a well-established resource in the realm of simulated research. The ICVL dataset comprises 201 HSIs with dimensions of 1392×1300 and encompasses 31 spectral bands spanning from 400 to 700 nm. Following [20], 100 of these HSIs are allocated for training with corresponding augmentation, 5 are designated for validation, and 50 are reserved for testing. The training phase data is cropped to $64 \times 64 \times 31$, while the test phase data will be center cropped to $512 \times 512 \times 31$. The simulated noise types include independent identically distributed Gaussian of different intensities, non-independent identically distributed Gaussian, impulse, stripe, deadline, and mixture noises. For generalization experiment, we also directly use pre-trained models on ICVL dataset to test on CAVE [64] and Harvard [65] datasets, which have the same spectral parameters as ICVL.

In addition to conducting experiments with close-range HSIs mentioned above, we also apply all the competing methods to Houston 2018, a remote sensing image data. It measures 1202×4172 pixels and contains a total of 48 bands. We also use pre-trained models to transfer directly to this dataset for testing. Following [66], we select 46 bands that are relatively clean, and crop a 512×512 cube from the center. To better simulate the real noise within remote sensing images, we only introduce complex noise to the clean HSI.

2) Real Noise Datasets: In addition to simulated noise, we also apply all the competing methods to real noise in remote sensing HSIs, specifically Urban, Indian Pines, and Pavia University datasets. The Urban dataset comprises an image with dimensions of 307×307 and contains 210 spectral bands, ranging from 400 to 2500 nm. For Urban dataset, since there is no clean HSI available, we utilize the APEX dataset for the pre-training phase. During this phase, we introduce band-dependent noise levels ranging from 0 to 55 into all clean HSIs. The configuration aligns with the setting used in [14]. The dataset of Indian Pines consists of 145×145 pixels, capturing 224 distinct bands of spectral reflectance data across the spectrum from 400 to 2500 nm. In addition, the Pavia University dataset comprises 610×340 with 103 spectral bands spanning from 430 to 860 nm. However, the first three bands of the Pavia University dataset are characterized by a high level of noise degradation. The spatial resolution of this dataset is defined by 1.3 m/pixel. Then for these datasets, we adopted the same setup as in [18]. More specifically, for constructing training pairs, we select the first 31 bands and then add Gaussian noise with noise level $\sigma = 15$ to these bands. These training pairs are then used to fine-tune our network. For testing set, we divided all datasets into multiple sub-HSIs, each containing 31 channels.

3) Compared Methods: We conduct a comparative evaluation, considering various traditional model-based HSI denoising techniques, namely BM4D [16], LLRT [17], NGMeet [11], LRTV [26], LRMR [67], and LRTDTV [68]. Additionally, we compare our method with several state-of-the-art deep learning-based approaches, namely GRNet [18], MAC-Net [19], QRNN3D [20], Restormer [21], T3SC [22], SST [15], and SERT [14].

4) Hardware Setting and Criteria: The traditional methods are implemented in MATLAB using an AMD Ryzen 5700X, while our method and other deep learning networks are assessed using a single NVIDIA RTX 3090 GPU. To quantify the results, we employ peak SNR (PSNR), structural similarity index metric (SSIM), and spectral angle mapper (SAM) as the evaluation criteria.

5) Network and Training Setup: The number of output channels of $\text{Conv}_{\text{start}}(\cdot)$, i.e., C, is set to 96. And the number of input channels is determined by the dataset. Then there are three PRSs, each of which consists of six PLRTBs stacked with a 3×3 convolutional layer. In PLRTB, the width of rectangular windows is set to 16. And our low-rank prompt is set as a vector group with 128 vectors of six dimensions. For the G-FFN model, the initial point-wise convolution will map features to 256 dimensions, while the final point-wise convolution will map features to 96 dimensions.

B. Evaluation on Simulated Noise Datasets

1) Quantitative Performance: Table I reveals the denoising comparison regarding PSNR, SSIM, and SAM on ICVL testing set with corresponding level of Gaussian noise. The *blind* refers to the Gaussian noise with a random sigma ranging from 30 to 70. The top results are emphasized in bold. Among the traditional methods, NGMeet comes close to many of the best deep learning methods and even surpasses some. Our method outperforms all other methods and exceeds state-of-the-art by a minimum of 0.58 dB in PSNR. However, inside Table II, which is the complex noise simulation experiment, our method does not outperform all. Similarly, some of the most recent methods perform well in Gaussian noise, while not as well as some of the classical models of the past on complex noise. Still, our PSNR values are optimal in most cases of noise, which demonstrates the good generalization of our method. Meanwhile, thanks to the low-rank modeled directly inside the network using prompt, our model outperforms both implicit and explicit reliance on external low-rank modeling of SERT [14] and MAC-Net [19] under various noise settings.

TABLE I

SIMULATED GAUSSIAN DENOISING RESULTS WITH DIFFERENT SIGMA ON ICVL. THE *Blind* REFERS TO THE GAUSSIAN NOISE WITH A RANDOM SIGMA RANGING FROM 30 TO 70. OUR METHOD CAN OUTPERFORM ALL OTHER METHODS IN TERMS OF PSNR, SSIM, AND SAM. THE BEST IS BOLD AND THE SECOND BEST IS UNDERLINED

Method	30			50			70			blind		
	PSNR	SSIM	SAM	PSNR	SSIM	SAM	PSNR	SSIM	SAM	PSNR	SSIM	SAM
Noisy	18.59	0.552	0.807	14.15	0.348	0.991	11.23	0.230	1.105	17.24	0.478	0.859
BM4D	37.84	0.920	0.135	34.98	0.866	0.178	33.11	0.815	0.214	36.27	0.852	0.180
LLRT	41.12	0.967	0.056	38.99	0.945	0.075	37.36	0.930	0.087	40.97	0.956	0.064
NGMeet	42.44	0.973	0.050	40.26	0.964	0.059	38.66	0.950	0.067	42.23	0.971	0.053
GRNet	41.74	0.973	0.059	40.07	0.960	0.064	38.85	0.949	0.071	41.07	0.967	0.060
MAC-Net	41.37	0.965	0.058	40.23	0.957	0.069	38.80	0.944	0.082	40.61	0.958	0.066
QRNN3D	42.58	0.972	0.056	40.45	0.957	0.069	38.48	0.937	0.090	41.74	0.965	0.063
Restormer	42.80	0.974	0.062	41.03	0.963	0.062	39.62	0.952	0.069	41.99	0.969	0.064
T3SC	42.02	0.969	0.075	40.05	0.955	0.084	38.63	0.942	0.093	41.18	0.962	0.081
SST	43.06	0.976	0.049	40.93	0.964	0.057	39.45	0.952	0.064	42.25	0.971	0.053
SERT	43.50	0.977	0.047	41.38	0.965	0.055	39.91	0.954	0.062	42.71	0.972	0.050
HyLoRa(Ours)	**44.19**	**0.978**	**0.041**	**42.01**	**0.966**	**0.048**	**40.49**	**0.956**	**0.054**	**43.43**	**0.973**	**0.043**

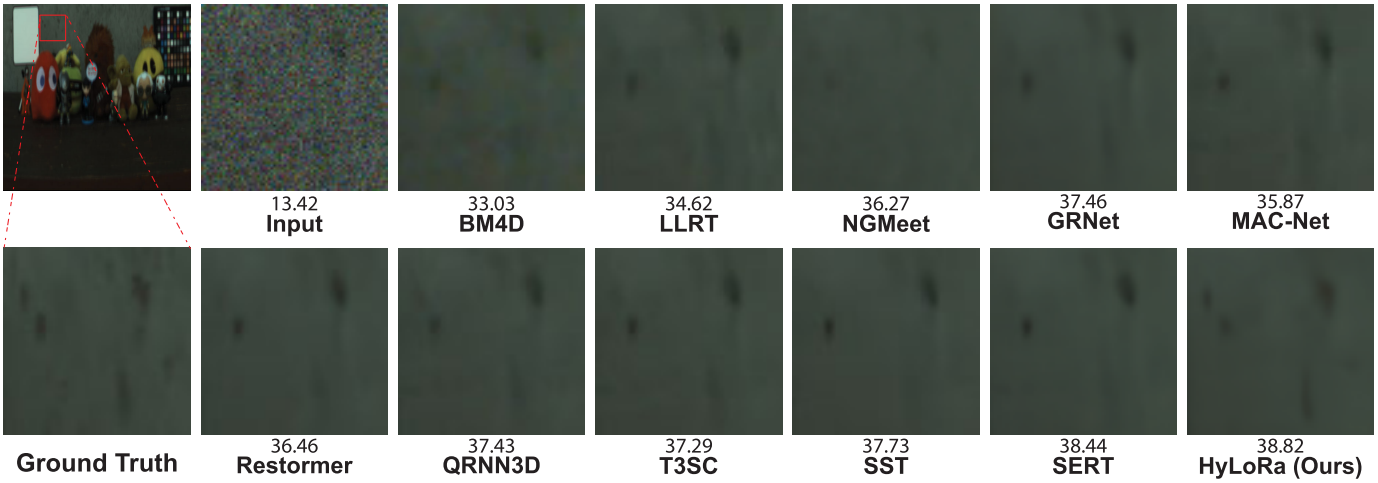

Fig. 5. Visual comparison for denoising simulated Gaussian noise with sigma = 70 on objects_0924-1641 HSI. The number below the image indicates the corresponding PSNR. Zoom in for a better view of the difference. For simple Gaussian noise, most methods can achieve considerable visual results and our method achieves the best PSNR.

TABLE II

SIMULATED COMPLEX DENOISING RESULTS ON ICVL. *non-iid* DENOTES NON I.I.D GAUSSIAN NOISE. *Stripe*, *Deadline*, *Impulse* DENOTE THE COMBINATION OF *non-iid* AND CORRESPONDING NOISE. *Mixture* DENOTES THE COMBINATION OF ALL THE MENTIONED NOISE TYPES. OUR METHOD CAN SURPASS THE IMAGE QUALITY METRICS OF ALL OTHER METHODS IN MOST CASES. THE BEST IS BOLD AND THE SECOND BEST IS UNDERLINED

Method	non-iid			stripe			deadline			impulse			mixture		
	PSNR	SSIM	SAM	PSNR	SSIM	SAM	PSNR	SSIM	SAM	PSNR	SSIM	SAM	PSNR	SSIM	SAM
Noisy	18.25	0.168	0.898	17.80	0.159	0.910	17.61	0.155	0.917	14.80	0.114	0.926	14.08	0.099	0.944
LRMR	32.79	0.719	0.178	32.65	0.710	0.185	31.74	0.698	0.263	29.64	0.623	0.309	28.80	0.633	0.321
LRTV	33.75	0.920	0.069	33.43	0.899	0.072	32.29	0.900	0.116	31.57	0.871	0.242	30.52	0.881	0.267
LRTDTV	36.37	0.948	0.061	36.16	0.917	0.082	33.97	0.896	0.101	35.30	0.908	0.092	32.71	0.885	0.110
GRNet	34.20	0.894	0.107	33.91	**0.982**	**0.040**	32.97	0.881	0.111	31.82	0.822	0.166	31.08	0.817	0.163
MAC-Net	38.58	0.957	0.091	38.18	0.981	0.043	36.41	0.951	0.109	33.28	0.889	0.226	30.04	0.840	0.306
QRNN3D	42.22	0.973	0.057	41.93	0.970	0.068	41.74	0.971	0.060	39.40	0.943	0.106	38.50	0.938	0.108
Restormer	43.63	**0.983**	0.041	43.04	0.952	0.092	42.46	0.980	0.044	**41.16**	**0.971**	**0.058**	37.61	0.951	**0.070**
T3SC	41.45	0.972	0.062	41.03	0.889	0.110	39.35	0.964	0.101	35.57	0.927	0.201	33.78	0.915	0.230
SST	43.42	0.982	0.041	42.99	0.980	0.043	42.70	0.980	0.044	40.48	0.965	0.067	38.66	0.953	0.071
SERT	43.69	0.982	0.041	43.24	0.981	0.044	43.05	0.980	0.043	40.44	0.962	0.074	39.25	0.954	0.075
HyLoRa(Ours)	**44.11**	**0.983**	**0.039**	**43.80**	0.971	0.059	**43.55**	**0.982**	**0.041**	40.57	0.961	0.080	**39.27**	**0.957**	0.073

2) Qualitative Performance: For Gaussian noise denoising, we pick image Labtest_0910-1513 to visualize our comparison in Fig. 5. As we can see, some of the latest deep learning methods can achieve great results. And our method is the best among these. Also, Restormer [21], which only considers RGB image restoration, gives good visual results and substantially outperforms MAC-Net [19], which is explicitly low-rank modeled, and SERT [14], which is implicitly low-rank modeled. As shown in Fig. 6, the spectral reflectance of a pixel (175, 15) in the image was sampled. Our method produces the curve that most closely resembles the ground truth.

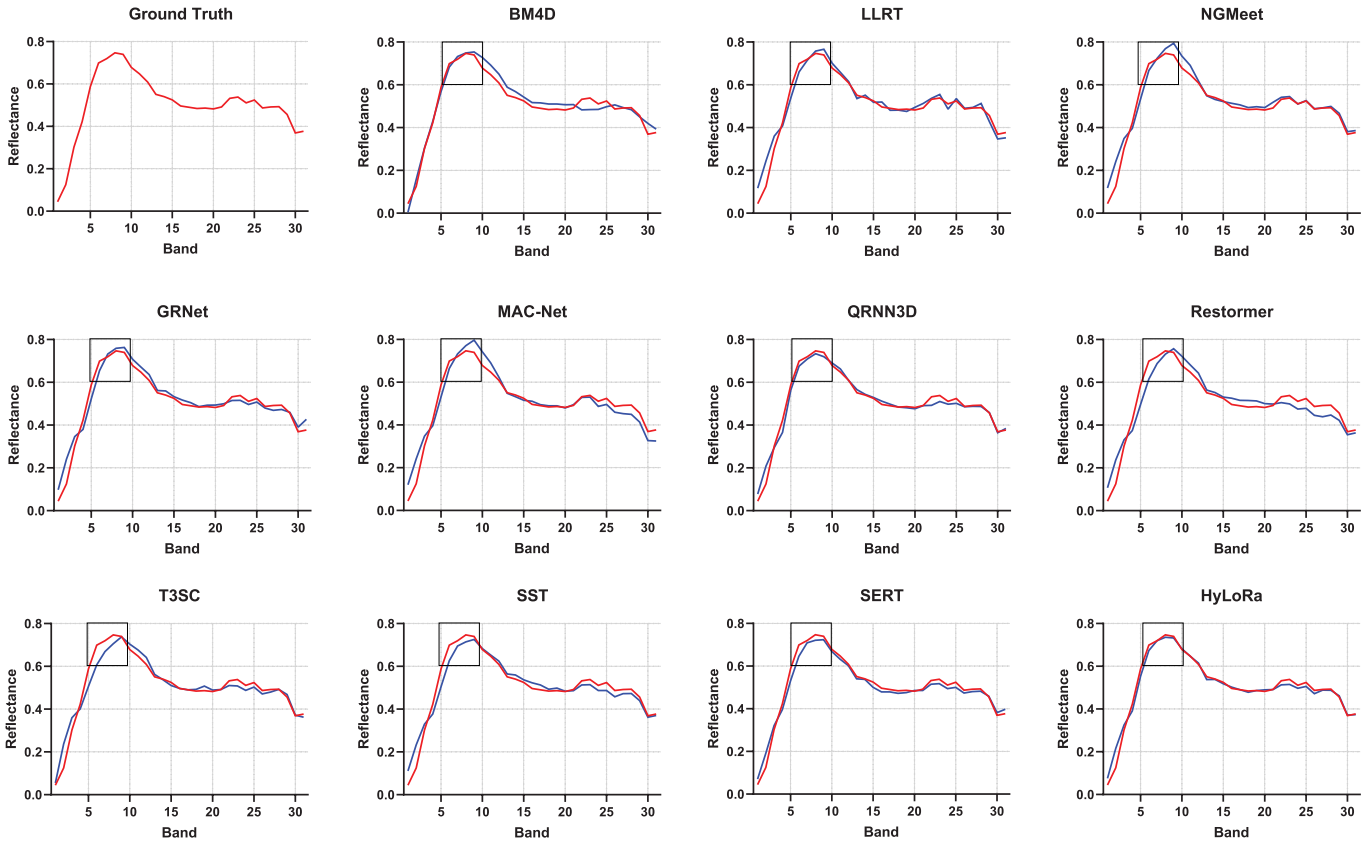

Fig. 6. Reflectance of Gaussian denoising results at pixel (200, 100) in objects_0924-1641 HSI. Our method produces the curve that most closely resembles the ground truth.

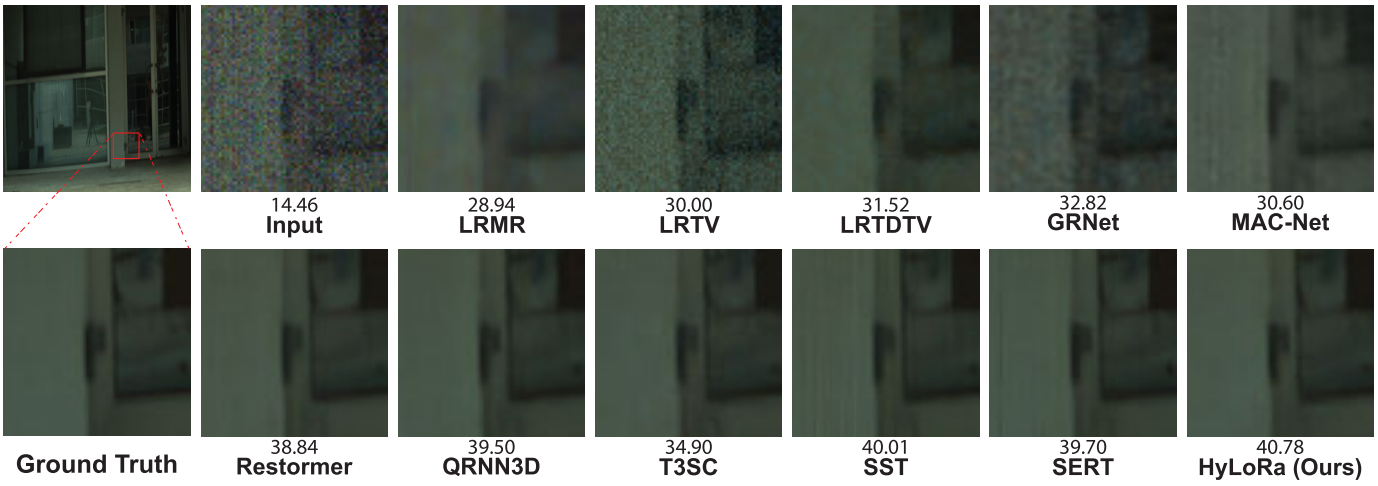

Fig. 7. Visual comparison for denoising simulated mixture noise on objects_0924-1652 HSI. The number below the image indicates the corresponding PSNR. Zoom in for a better view of the difference. For simple Gaussian noise, most methods can achieve considerable visual results and our method achieves the best PSNR.

Similarly, under other complex noise settings, we take image nachal_0823-1145 with a large number of flat regions for testing, which implies that the image itself definitely has low-rank properties. With Fig. 7, we can see that MAC-Net [19], which explicitly models low-rankness, can no longer get satisfactory results, while SERT [14] and ours, which rely on implicit modeling of the data, can still remove most of the noise. As shown in Fig. 8, the spectral reflectance of a pixel (127, 127) in the image was sampled. Our method produces the curve that most closely resembles the ground truth.

In order to show that our method is not particularly applicable to a particular band but brings about an overall effect, we utilize Master5000K_2900K and nachal_0823-1145 to compute band-by-band image quality metrics for Gaussian noise and complex noise, respectively. And through Fig. 9, we can see that our proposed is able to get considerable denoising performance on all bands.

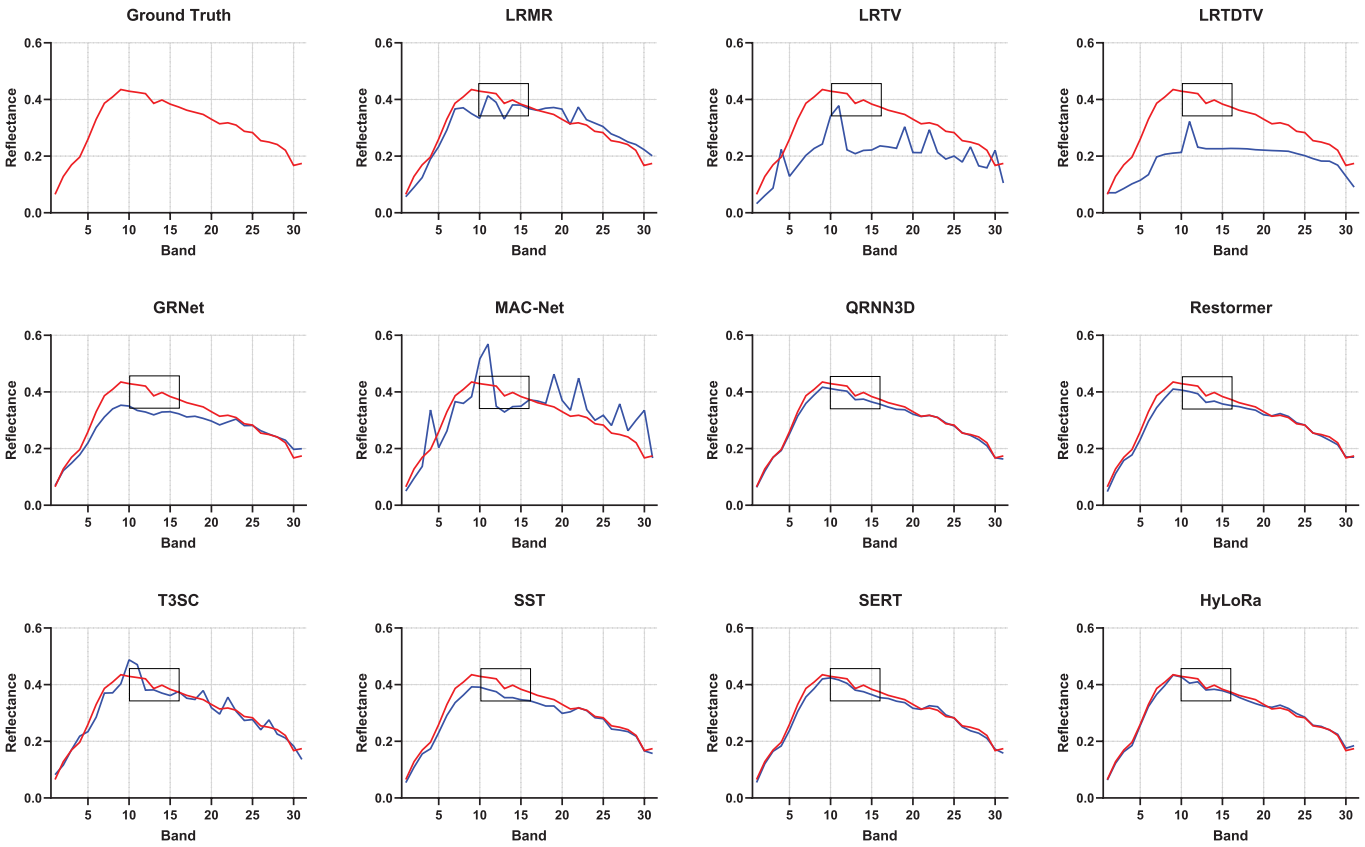

Fig. 8. Reflectance of complex mixture denoising results at pixel (250, 125) in objects_0924-1651 HSI. Our method produces the curve that most closely resembles the ground truth.

3) Generalization Performance: Another important measure of the performance of deep learning-based methods is generalization, which means that a good algorithm should perform well on data outside the training set. Here, we use two datasets, CAVE and Harvard, to conduct Gaussian noisy experiments. The results are presented in Tables III and IV. Due to the different camera parameters used in each dataset, the performance of our method is slightly degraded compared to other methods. Especially in the Harvard dataset, some of the metrics do not exceed those of the other methods. However, for most metrics, our approach remains effective even after transfer and outperforms other methods. For complex noise, we use Houston 2018. And the result is presented in Table V. We can see that all deep learning methods have a relatively large performance drop. We believe the reason for this is that ICVL's spectral channels are from 400 to 700 nm at 10 nm increments while Houston contains 48 bands spanning from 380 to 1050 nm, which introduces a significant bias. So, model-based approaches with zero-shot inference still maintain a relatively excellent result in this case. Especially for noise-heavy impulse and mixture settings, LRTDTV achieves optimal PSNR. However, among the deep learning methods, our approach achieves the best performance in the vast majority of cases and exhibits excellent adaptability and reliability in dealing with a wide range of complex noises.

C. Evaluation on Real Noise Datasets

We provide the denoising results of Urban HSI in Fig. 10. As we can see, images denoised by traditional methods usually exhibit blurred. Most deep learning-based methods fail to remove noise because of the noisy diversity over a large range of spectral space. However, some methods with low-rank modeling, like MAC-Net, SERT, and ours, process certain features in the more intrinsic low-rank space thereby effectively removing the noise. And our method can achieve a visually superior result.

Additionally, Fig. 11 shows the recovered HSIs of the Pavia University by the competing methods. It can be observed that LRTDTV can obtain better denoising performance among the model-based methods. And our method can achieve a better and comparable visualization among the deep learning-based methods.

We also provide the denoising results of Indian Pines in Fig. 12. As we can see, all methods provide some noise suppression, but some exhibit blurring. LRMR and LRTDTV show very good results in detail recovery and are slightly superior to ours in some regions of the image. However, our method still achieves a good visualization among the deep learning methods.

V. Discussion

A. Model Complexity

In the presented Table VI, our method emerges as a compelling choice among various deep learning methods. Although HyLoRa does not have the lowest parameter quantity, it remains within a reasonable range compared to other high-performing methods. This suggests a favorable balance

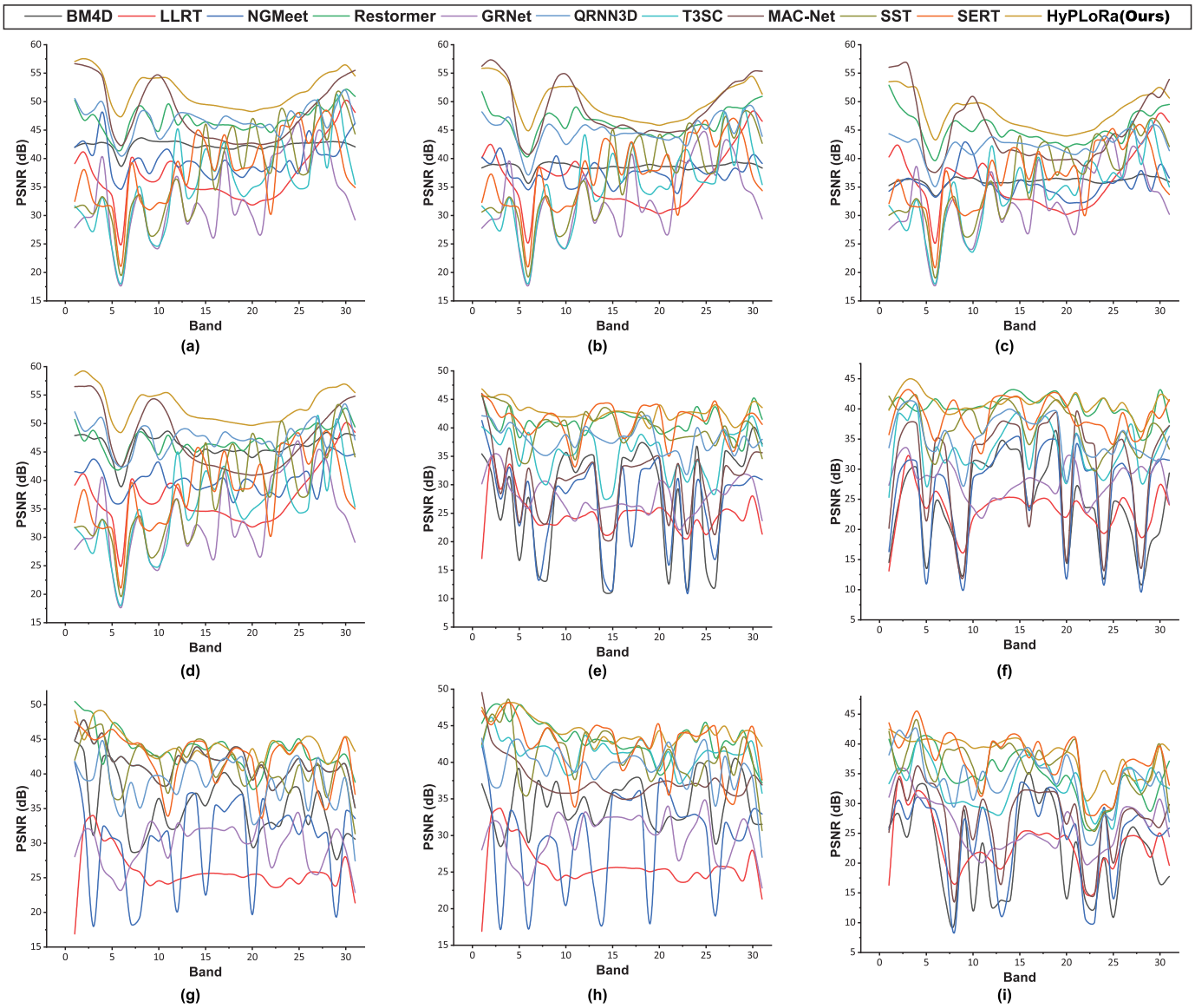

Fig. 9. Bandwise PSNR values of denoised image. (a)–(d) Denoising performance of Gaussian noise with sigma of 30, 50, 70, and blind, respectively. (e)–(i) Noise denoising performance of deadline, impulse, non-iid, stripe, and mixture, respectively. Our method achieves the highest PSNR in most bands.

TABLE III

GENERALIZATION GAUSSIAN DENOISING RESULTS WITH DIFFERENT SIGMA ON CAVE. THE *Blind* REFERS TO THE GAUSSIAN NOISE WITH A RANDOM SIGMA RANGING FROM 30 TO 70. THE BEST IS BOLD AND THE SECOND BEST IS UNDERLINED

Method	30			50			70			blind		
	PSNR	SSIM	SAM	PSNR	SSIM	SAM	PSNR	SSIM	SAM	PSNR	SSIM	SAM
Noisy	20.47	0.173	0.774	16.35	0.088	0.867	13.67	0.054	0.919	20.44	0.193	0.775
GRNet	34.24	0.910	0.304	33.49	0.894	0.325	32.84	0.879	0.346	34.07	0.907	0.310
MAC-Net	34.92	0.904	0.384	34.25	0.890	0.392	33.55	0.872	0.402	34.38	0.881	0.412
QRNN3D	37.50	0.938	0.257	36.01	0.915	0.300	34.66	0.887	0.347	37.21	0.934	0.264
Restormer	35.15	0.926	0.250	33.93	0.907	0.293	33.01	0.888	0.336	34.98	0.923	0.259
T3SC	36.33	0.919	0.239	35.34	0.903	0.252	34.52	0.889	**0.262**	36.15	0.917	0.241
SST	36.84	0.946	0.243	35.45	0.926	0.284	34.37	0.908	0.314	36.62	0.944	0.245
SERT	37.75	0.949	0.225	36.11	0.926	0.279	34.89	0.907	0.314	37.48	**0.951**	0.223
HyLoRa (Ours)	**38.04**	**0.953**	**0.196**	**36.47**	**0.934**	**0.240**	**35.23**	**0.916**	0.276	**37.78**	0.946	**0.197**

in resource utilization. And HyLoRa demonstrates relatively lower computational complexity with 658.08 GFLOPS, implying efficient resource utilization in HSI denoising. Finally, HyLoRa exhibits a short inference time of 0.6 s, indicating that it can generate images at a high speed, making it suitable for real-time or latency-sensitive applications. Therefore, HyLoRa stands out in terms of image quality, model complexity, computational efficiency, and inference speed, making it a noteworthy and competitive deep learning method.

TABLE IV

GENERALIZATION GAUSSIAN DENOISING RESULTS WITH DIFFERENT SIGMA ON HARVARD. THE *Blind* REFERS TO THE GAUSSIAN NOISE WITH A RANDOM SIGMA RANGING FROM 30 TO 70. THE BEST IS BOLD AND THE SECOND BEST IS UNDERLINED

Method	30			50			70			blind		
	PSNR	SSIM	SAM	PSNR	SSIM	SAM	PSNR	SSIM	SAM	PSNR	SSIM	SAM
Noisy	18.59	0.552	0.807	14.15	0.348	0.991	11.23	0.230	1.105	17.24	0.478	0.859
GRNet	40.38	0.951	0.089	39.30	0.935	0.095	38.49	0.922	0.099	40.17	0.950	0.089
MAC-Net	41.82	0.959	0.074	40.31	0.943	0.082	39.26	0.929	0.089	41.62	0.957	0.074
QRNN3D	40.88	0.952	0.083	39.36	0.933	0.094	37.79	0.910	0.110	40.56	0.949	0.085
Restormer	41.27	0.955	0.078	39.83	0.939	0.087	38.75	0.924	0.096	40.95	0.951	0.082
T3SC	41.79	0.954	0.084	40.30	0.935	0.091	39.12	0.919	0.098	41.47	0.950	0.085
SST	41.79	**0.960**	0.073	40.26	0.943	0.083	39.21	0.928	0.091	41.53	**0.958**	0.074
SERT	42.02	**0.960**	0.071	40.56	**0.944**	0.079	39.50	**0.930**	0.086	41.78	**0.958**	0.071
HyLoRa (Ours)	**42.62**	0.957	**0.064**	**40.89**	0.938	**0.075**	**39.67**	0.922	**0.085**	**42.21**	0.953	**0.065**

TABLE V

GENERALIZATION COMPLEX DENOISING RESULTS WITH DIFFERENT SIGMA ON HOUSTON 2018. THE BEST IS BOLD AND THE SECOND BEST IS UNDERLINED

Method	non-iid			stipe			deadline			impulse			mixture		
	PSNR	SSIM	SAM	PSNR	SSIM	SAM	PSNR	SSIM	SAM	PSNR	SSIM	SAM	PSNR	SSIM	SAM
Noisy	14.41	0.122	0.849	14.36	0.121	0.852	14.35	0.117	0.865	10.94	0.063	0.868	10.87	0.060	0.887
LRMR	27.02	0.616	0.260	26.94	0.613	0.262	26.60	0.604	0.271	22.82	0.460	0.279	22.49	0.444	0.294
LRTV	26.51	0.737	0.109	26.50	0.736	0.110	26.11	0.730	0.113	23.05	0.658	0.151	23.14	0.653	0.153
LRTDTV	33.33	0.895	0.101	33.12	0.887	0.105	32.47	0.878	0.112	**32.19**	0.866	**0.111**	**31.03**	0.849	0.126
GRNet	28.52	0.837	0.209	28.45	0.835	0.210	27.84	0.824	0.210	25.39	0.654	0.234	24.74	0.625	0.238
MAC-Net	**37.27**	**0.955**	**0.071**	**36.43**	0.946	**0.075**	32.55	0.917	**0.082**	21.96	0.456	0.149	21.60	0.432	0.156
QRNN3D	34.57	0.939	0.118	34.39	0.936	0.119	33.08	0.920	0.122	30.30	0.625	0.142	29.49	0.582	0.151
Restormer	32.86	0.936	0.140	32.49	0.928	0.142	30.90	0.897	0.146	21.44	0.447	0.217	20.97	0.420	0.223
T3SC	34.74	0.932	0.102	34.59	0.930	0.103	33.73	0.922	0.106	25.11	0.643	0.171	24.80	0.610	0.188
SST	33.89	0.945	0.129	33.67	0.942	0.130	32.55	0.927	0.133	26.03	0.686	0.151	25.32	0.654	0.157
SERT	33.87	0.942	0.119	33.66	0.939	0.121	32.29	0.921	0.124	28.03	0.773	0.138	27.06	0.733	0.146
HyPLoRa	34.96	0.951	0.091	34.61	**0.947**	0.093	**34.00**	**0.940**	0.096	31.08	**0.887**	0.115	30.29	**0.868**	**0.121**

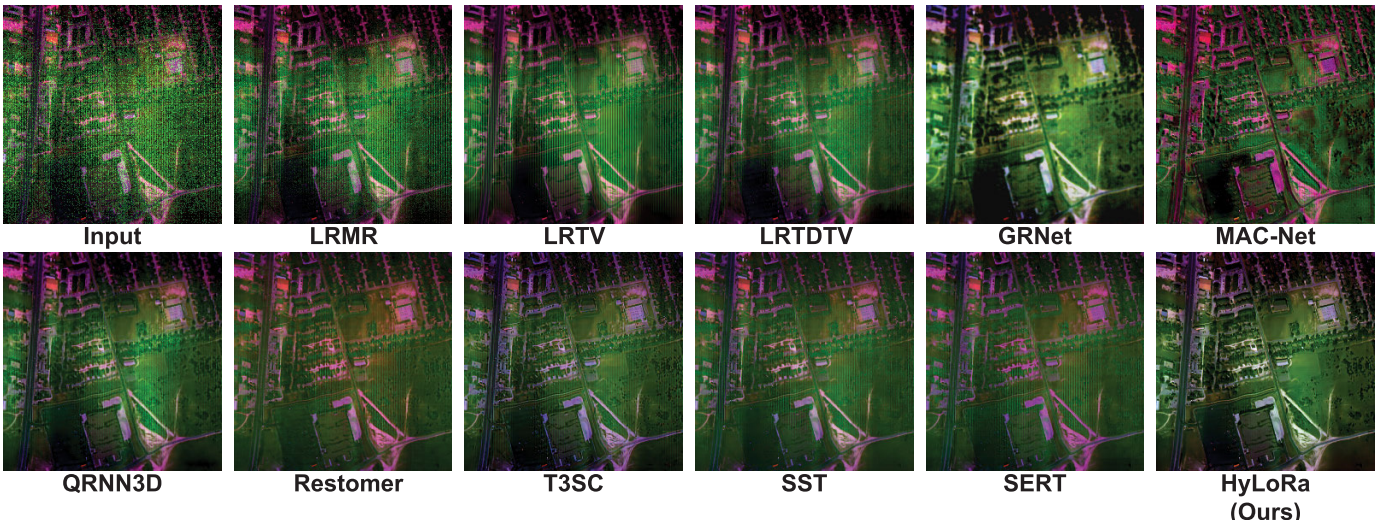

Fig. 10. Denoising results on real-world Urban dataset. The false-color images were generated by combining bands 109, 144, and 208. The proposed HyLoRa can provide a superior visual result.

TABLE VI

COMPARISONS OF PSNR, PARAMS, FLOPS, AND INFERENCE TIME OF DIFFERENT DEEP LEARNING METHODS AS THE INPUT SIZE IS $512 \times 512 \times 31$. MAC-NET AND T3SC ARE UNABLE TO DETERMINE SPECIFIC GFLOPS BECAUSE THEY REQUIRE MATRIX DECOMPOSITION

Metrics	GRNet	MAC-Net	QRNN3D	Restormer	T3SC	SST	SERT	HyLoRa
PSNR(db)	41.07	40.61	41.74	41.99	41.18	42.25	42.71	43.43
Params(M)	44.39	0.43	0.83	26.16	0.83	4.1	1.91	3.15
GFLOPS	314.69	–	2513.73	574.48	–	1081.8	1018.93	658.08
Times(s)	0.36	3.63	0.68	0.31	1.12	1.4	0.72	0.6

B. Cross-Shaped Window-Based Self-Attention

Compared to the pure window-based method, which might only consider information within a fixed rectangular region, the cross-shaped window offers a more flexible and adaptable approach. The larger receptive fields and double directions allow the model to selectively focus on features that might align better with the self-similarity in an HSI. Despite its enhanced modeling capabilities, the cross-shaped window

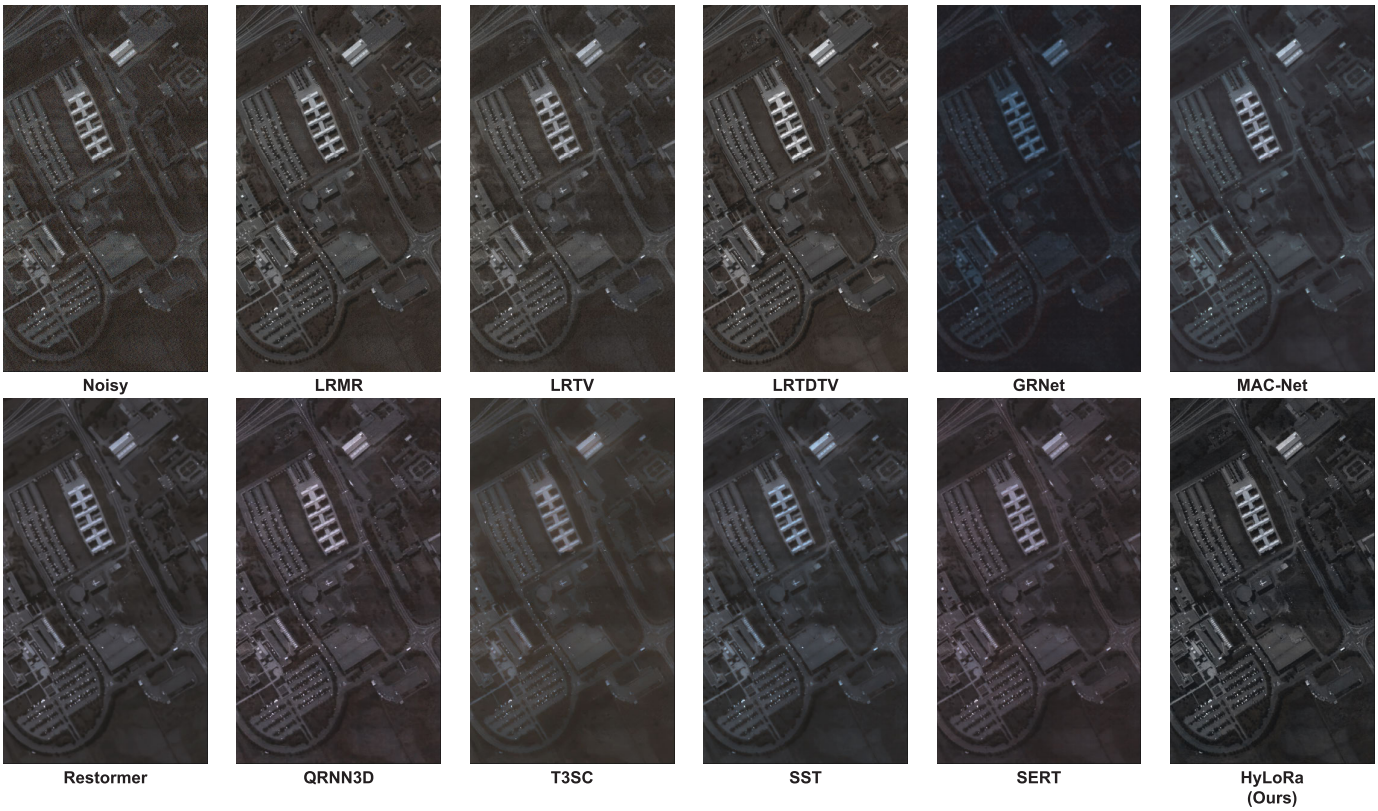

Fig. 11. Denoising results on real-world Pavia University dataset. The false-color images were generated by combining bands 1, 2, and 3. The proposed HyLoRa can provide a superior visual result.

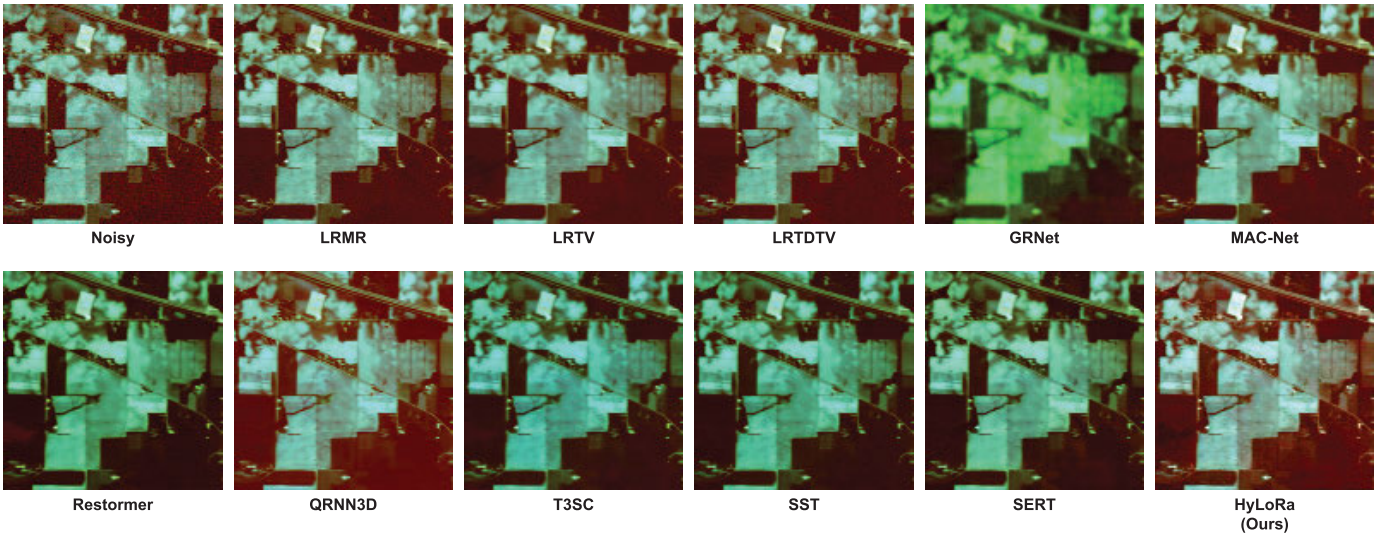

Fig. 12. Denoising results on real-world Indian Pines dataset. The false-color images were generated by combining bands 1, 20, and 30. The proposed HyLoRa can provide a superior visual result.

manages to maintain computational efficiency. This is because it narrows down the attention mechanism to a specific region, avoiding the need to consider the entire input space. By incorporating a window-like constraint, it reduces the number of computations required compared to methods that operate without such constraints. This makes it more scalable and applicable to larger datasets where computational efficiency is crucial.

From Table VII, it is evident that the attention mechanism using the cross-shaped window outperforms cases where only horizontal or only vertical window attention is employed. This suggests that considering both horizontal and vertical directions simultaneously, as facilitated by the cross-shaped window, leads to superior performance in capturing relevant dependencies and similarities within the data. In addition, we compare parallel horizontal and vertical attention, which

TABLE VII

ABLATION ANALYSIS FOR VARIOUS WINDOW SHAPES OF SELF-ATTENTION MECHANISMS ON ICVL DATASET UNDER RANDOM GAUSSIAN NOISE

Method	30	50	70	blind
Vertical	43.93	41.74	40.2	43.17
Horizon	43.91	41.7	40.17	43.16
Cross-Shaped	44.19	**42.01**	**40.49**	**43.43**
Parallel Cross-Shaped	**44.20**	42.00	40.47	**43.43**

TABLE VIII

ABLATION ANALYSIS OF THE PARAMETER $\boldsymbol{R}$ ON ICVL DATASET UNDER RANDOM GAUSSIAN NOISE

$\boldsymbol{R}$	30	50	70	blind
w/o	44.14	41.95	40.43	43.38
12	44.15	41.96	40.43	43.39
6	**44.19**	**42.01**	**40.49**	**43.43**
3	44.11	42.01	40.48	43.36

TABLE IX

ABLATION ANALYSIS OF THE FEEDFORWARD NETWORK ON ICVL DATASET UNDER RANDOM GAUSSIAN NOISE

FFN	GFLOPs	30	50	70	blind
MLP	96.6	44.11	41.88	40.40	43.30
G-FFN	**20.5**	**44.19**	**42.01**	**40.49**	**43.43**

means that we perform window attention on both the original feature map and its copy in different directions, and then fuse the output. Compared with splitting the feature map into two parts and conduct horizontal and vertical attention at the same time, this parallel approach does not bring significant performance improvement, but increases computational complexity.

C. Low-Rank Prompt

Table VIII presents the results of an ablation analysis conducted on our experiment, focusing on a parameter $\boldsymbol{R}$ of low-rank prompt module which represents the number of rows in this module. It is believed that parameter $\boldsymbol{R}$ represents the assumed dimensionality of the low-dimensional subspace where the low-rank structure of an HSI is presumed to exist. The performance metrics, measured under random Gaussian noise on the ICVL dataset, are reported for different values of $\boldsymbol{R}$. Specifically, the highest performance is achieved when $\boldsymbol{R}$ is set to 6, which is consistent with the regularity stated in Zhuang et al. [44] that data in ICVL can generally be expressed in a 6-D space. This finding suggests that the choice of $\boldsymbol{R}$ has a discernible impact on the effectiveness of the proposed method. Moreover, this ablation analysis demonstrates that our module contributes to correcting the feature map in low-rank space.

D. Gated Feedforward Network

Table IX shows the performance difference between a regular MLP and our proposed G-FFN. Obviously, G-FFN can help network better remove noise with less computation. Compared to the original MLP, G-FFN models the local self-similarity again through depth-wise convolution. Subsequently, point-wise convolution performs parallel fusion from channel direction similar to linear layer fusion for pixel by pixel. At last, the gating mechanism dynamically adjusts the flow of information in the feature map so that the network can adapt to different levels of noise degradation.

VI. CONCLUSION

This study proposes a HyLoRa designed for HSI denoising. The included low-rank prompt encourages the absorption of generalization capabilities from model-based approaches, facilitating a more accurate representation of spatial–spectral correlations in feature space. In addition, the network's ability to leverage both local and nonlocal spatial dependencies is enhanced by incorporating a cross-shape window-based self-attention. This results in a robust representation capability for modeling similarity. Our HyLoRa outperforms both model-based and deep learning-based methods in extensive experiments on both close-range and remote sensing HSIs. This is supported by improved subjective visual effects and objective quantitative metrics. Furthermore, ablation analysis confirms our intention to design the modules and their respective roles. For future work, we will explore ways to integrate other inductive bias (e.g., sparsity) into the network for more effective denoising.

REFERENCES

[1] J. Peng et al., "Low-rank and sparse representation for hyperspectral image processing: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 1, pp. 10–43, Mar. 2022.

[2] H. Zhou, X. Zhang, C. Zhang, and Q. Ma, "Vision transformer with contrastive learning for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[3] W. Li, Q. Liu, S. Fan, C. Xu, and H. Bai, "Dual-stream GNN fusion network for hyperspectral classification," *Appl. Intell.*, vol. 53, no. 22, pp. 26542–26567, Nov. 2023.

[4] Y. Li et al., "Joint spectral–spatial hyperspectral image classification based on hierarchical subspace switch ensemble learning algorithm," *Appl. Intell.*, vol. 48, no. 11, pp. 4128–4148, Nov. 2018.

[5] H. Zhou, X. Zhang, C. Zhang, Q. Ma, and Y. Jiang, "Dictionary cache transformer for hyperspectral image classification," *Appl. Intell.*, vol. 53, no. 22, pp. 26725–26749, Nov. 2023.

[6] C. Shi, Z. Zhang, W. Zhang, C. Zhang, and Q. Xu, "Learning multiscale temporal–spatial–spectral features via a multipath convolutional LSTM neural network for change detection with hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5529816.

[7] M. Wang, D. Hong, B. Zhang, L. Ren, J. Yao, and J. Chanussot, "Learning double subspace representation for joint hyperspectral anomaly detection and noise removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5507517.

[8] J. M. Jurado, A. López, L. Pádua, and J. J. Sousa, "Remote sensing image fusion on 3D scenarios: A review of applications for agriculture and forestry," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, Aug. 2022, Art. no. 102856.

[9] V. I. Pasha and D. B. Megherbi, "A deep learning approach for hyperspectral image classification with additive noise for remote sensing and airborne surveillance," in *Proc. IEEE 9th Int. Conf. Comput. Intell. Virtual Environ. Meas. Syst. Appl. (CIVEMSA)*, Jun. 2022, pp. 1–6.

[10] H. Fu et al., "A novel band selection and spatial noise reduction method for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5535713.

[11] W. He, Q. Yao, C. Li, N. Yokoya, and Q. Zhao, "Non-local meets global: An integrated paradigm for hyperspectral denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6861–6870.

[12] L. Zhuang and J. M. Bioucas-Dias, "Fast hyperspectral image denoising and inpainting based on low-rank and sparse representations," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 3, pp. 730–742, Mar. 2018.

[13] B. Arad and O. Ben-Shahar, "Sparse recovery of hyperspectral signal from natural RGB images," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 19–34.

[14] M. Li, J. Liu, Y. Fu, Y. Zhang, and D. Dou, "Spectral enhanced rectangle transformer for hyperspectral image denoising," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 5805–5814.

[15] M. Li, Y. Fu, and Y. Zhang, "Spatial–spectral transformer for hyperspectral image denoising," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 1, pp. 1368–1376.

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

[17] Y. Chang, L. Yan, and S. Zhong, "Hyper-Laplacian regularized unidirectional low-rank tensor recovery for multispectral image denoising," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5901–5909.

[18] X. Cao, X. Fu, C. Xu, and D. Meng, "Deep spatial–spectral global reasoning network for hyperspectral image denoising," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5504714.

[19] F. Xiong, J. Zhou, Q. Zhao, J. Lu, and Y. Qian, "MAC-Net: Model-aided nonlocal neural network for hyperspectral image denoising," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5519414.

[20] K. Wei, Y. Fu, and H. Huang, "3-D quasi-recurrent neural network for hyperspectral image denoising," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 363–375, Jan. 2021.

[21] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5718–5729.

[22] T. Bodrito, A. Zouaoui, J. Chanussot, and J. Mairal, "A trainable spectral–spatial sparse coding model for hyperspectral image restoration," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 5430–5442. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/2b515e2bdd63b7f034269ad747c93a42-Paper.pdf

[23] C. Jiang, H. Zhang, L. Zhang, H. Shen, and Q. Yuan, "Hyperspectral image denoising with a combined spatial and spectral weighted hyperspectral total variation model," *Can. J. Remote Sens.*, vol. 42, no. 1, pp. 53–72, Jan. 2016.

[24] J. Peng et al., "Fast noise removal in hyperspectral images via representative coefficient total variation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5546017.

[25] Z. Zha et al., "Nonlocal structured sparsity regularization modeling for hyperspectral image denoising," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5510316.

[26] 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, Jan. 2016.

[27] C. Cao, J. Yu, C. Zhou, K. Hu, F. Xiao, and X. Gao, "Hyperspectral image denoising via subspace-based nonlocal low-rank and sparse factorization," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 3, pp. 973–988, Mar. 2019.

[28] Y. Su, H. Zhu, K.-C. Wong, Y. Chang, and X. Li, "Hyperspectral image denoising via weighted multidirectional low-rank tensor recovery," *IEEE Trans. Cybern.*, vol. 53, no. 5, pp. 2753–2766, May 2023.

[29] Y. Chang, L. Yan, X.-L. Zhao, H. Fang, Z. Zhang, and S. Zhong, "Weighted low-rank tensor recovery for hyperspectral image restoration," *IEEE Trans. Cybern.*, vol. 50, no. 11, pp. 4558–4572, Nov. 2020.

[30] Y. Chen, T.-Z. Huang, W. He, X.-L. Zhao, H. Zhang, and J. Zeng, "Hyperspectral image denoising using factor group sparsity-regularized nonconvex low-rank approximation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5515916.

[31] Y. Yuan, H. Ma, and G. Liu, "Partial-DNet: A novel blind denoising model with noise intensity estimation for HSI," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5505913.

[32] F. Xiong, J. Zhou, J. Zhou, J. Lu, and Y. Qian, "Multitask sparse representation model-inspired network for hyperspectral image denoising," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5518515.

[33] Z. Wang, 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 Trans. Geosci. Remote Sens.*, vol. 60, 2022.

[34] H. V. Nguyen, M. O. Ulfarsson, and J. R. Sveinsson, "Sure based convolutional neural networks for hyperspectral image denoising," in *Proc. IGARSS*, Sep. 2020, pp. 1784–1787.

[35] F. Xiong, J. Zhou, S. Tao, J. Lu, J. Zhou, and Y. Qian, "SMDS-Net: Model guided spectral–spatial network for hyperspectral image denoising," *IEEE Trans. Image Process.*, vol. 31, pp. 5469–5483, 2022.

[36] Q. Zhang, Q. Yuan, M. Song, H. Yu, and L. Zhang, "Cooperated spectral low-rankness prior and deep spatial prior for HSI unsupervised denoising," *IEEE Trans. Image Process.*, vol. 31, pp. 6356–6368, 2022.

[37] J. Peng et al., "Learnable representative coefficient image denoiser for hyperspectral image," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5506516.

[38] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, Jun. 2021, pp. 1833–1844.

[39] X. Dong et al., "CSWin transformer: A general vision transformer backbone with cross-shaped windows," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 12114–12124.

[40] Z. Li, H. Chen, J. Wu, J. Li, and N. Jing, "SegMind: Semisupervised remote sensing image semantic segmentation with masked image modeling and contrastive learning method," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4408917.

[41] Z. Yang, M. Xu, S. Liu, H. Sheng, and J. Wan, "UST-Net: A U-shaped transformer network using shifted windows for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5528815.

[42] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[43] F. Zhu, Y. Wang, B. Fan, S. Xiang, G. Meng, and C. Pan, "Spectral unmixing via data-guided sparsity," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5412–5427, Dec. 2014.

[44] L. Zhuang, M. K. Ng, L. Gao, J. Michalski, and Z. Wang, "Eigen-image2Eigenimage (E2E): A self-supervised deep learning network for hyperspectral image denoising," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jul. 2023.

[45] W.-H. Wu, T.-Z. Huang, X.-L. Zhao, J.-L. Wang, and Y.-B. Zheng, "Hyperspectral image denoising via tensor low-rank prior and unsupervised deep spatial–spectral prior," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5545514.

[46] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 3045–3059.

[47] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 4582–4597.

[48] X. Liu et al., "P-Tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, Dublin, Ireland, May 2022, pp. 61–68.

[49] M. Jia et al., "Visual prompt tuning," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 709–727.

[50] V. Potlapalli, S. W. Zamir, S. Khan, and F. S. Khan, "PromptIR: Prompting for all-in-one blind image restoration," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 71275–71293.

[51] J. Ma, T. Cheng, G. Wang, Q. Zhang, X. Wang, and L. Zhang, "ProRes: Exploring degradation-aware visual prompt for universal image restoration," 2023, *arXiv:2306.13653*.

[52] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 548–558.

[53] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[54] J. Wang, Y. Liu, and L. Li, "Background augmentation with transformer-based autoencoder for hyperspectral anomaly detection," in *Intelligence Science IV*. Cham, Switzerland: Springer, 2022, pp. 302–309.

[55] T. You, C. Wu, Y. Bai, D. Wang, H. Ge, and Y. Li, "HMF-Former: Spatio-spectral transformer for hyperspectral and multispectral image fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[56] X. Wang, X. Wang, R. Song, X. Zhao, and K. Zhao, "MCT-Net: Multi-hierarchical cross transformer for hyperspectral and multispectral image fusion," *Knowl.-Based Syst.*, vol. 264, Mar. 2023, Art. no. 110362.

[57] D. Yu, Q. Li, X. Wang, Z. Zhang, Y. Qian, and C. Xu, "DSTrans: Dual-stream transformer for hyperspectral image restoration," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 3728–3738.

[58] Y. Long, X. Wang, M. Xu, S. Zhang, S. Jiang, and S. Jia, "Dual self-attention Swin transformer for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5512012.

[59] Q. Zhang, Y. Xu, J. Zhang, and D. Tao, "VSA: Learning varied-size window attention in vision transformers," in *Proc. ECCV*, 2022, pp. 466–483.
[60] J. M. Jose Valanarasu, R. Yasarla, and V. M. Patel, "TransWeather: Transformer-based restoration of images degraded by adverse weather conditions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 2343–2353.
[61] T. Ye et al., "Adverse weather removal with codebook priors," in *Proc. Int. Conf. Comput. Vis.*, Oct. 2023, pp. 12653–12664.
[62] T. Xiao, P. Dollar, M. Singh, E. Mintun, T. Darrell, and R. Girshick, "Early convolutions help transformers see better," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 30392–30400. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/ff1418e8cc993fe8abcfe3ce2003e5c5-Paper.pdf
[63] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.
[64] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar, "Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2241–2253, Sep. 2010.
[65] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 193–200.
[66] G. Fu, F. Xiong, J. Lu, J. Zhou, J. Zhou, and Y. Qian, "Hyperspectral image denoising via spatial–spectral recurrent transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5511214.
[67] 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, Aug. 2014.
[68] Y. Wang, J. Peng, Q. Zhao, Y. Leung, X.-L. Zhao, and D. Meng, "Hyperspectral image restoration via total variation regularized low-rank tensor decomposition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 4, pp. 1227–1243, Apr. 2018.

Xiaodong Tan received the B.Eng. degree in computer science and technology from Shandong University of Science and Technology, Qingdao, China, in 2022. He is currently pursuing the M.Sc. degree in computer science and technology with Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao.

His research interests include hyperspectral image processing and machine learning.

Mingwen Shao (Member, IEEE) received the Ph.D. degree in applied mathematics from Xi'an Jiaotong University, Xi'an, China, in 2005, and the Ph.D. degree in control science and engineering from Tsinghua University, Beijing, China, in 2008.

He is currently a Professor and a Ph.D. Supervisor with China University of Petroleum (East China), Qingdao, China. His research interests include rough sets, data mining, machine learning, and computer vision.

Yuanjian Qiao received the M.S. degree in electrical engineering and automation from the Qilu University of Technology (Shandong Academy of Sciences), Jinan, China, in 2021. He is currently pursuing the Ph.D. degree in advanced scientific and engineering computing with the School of Computer Science and Technology, China University of Petroleum (East China), Qingdao, China, under the supervision of Prof. Mingwen Shao.

His research interests include image restoration and deep learning.

Tiyao Liu received the B.S. degree in network engineering from Zaozhuang College, Zaozhuang, China, in 2018. He is currently pursuing the M.Eng. degree in computer science and technology with Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao, China.

His research interests include image processing, pattern recognition, and bioinformatics.

Xiangyong Cao received the B.Sc. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 2012 and 2018, respectively.

From 2016 to 2017, he was a Visiting Scholar with Columbia University, New York, NY, USA. He is currently an Associate Professor with the School of Computer Science and Technology, Xi'an Jiaotong University. His research interests include statistical modeling and image processing.