Three-Dimension Spatial–Spectral Attention Transformer for Hyperspectral Image Denoising

Qiang Zhang, *Member, IEEE*, Yushuai Dong, Yaming Zheng, *Graduate Student Member, IEEE*, Haoyang Yu, *Member, IEEE*, Meiping Song, Lifu Zhang, *Senior Member, IEEE*, and Qiangqiang Yuan, *Member, IEEE*

***Abstract*— Hyperspectral image (HSI) denoising is a crucial step for its subsequent applications. In this article, we propose TDSAT, a 3-D spatial–spectral attention Transformer model designed to effectively remove noise in HSI processing while preserving essential spectral and spatial information. The primary objective of this model is to utilize the 3-D Transformer to explore the global spectral–spatial features in HSI, learn the relationships among different bands, and preserve high-quality spectral and spatial information for denoising. The proposed method consists of three main components: the multihead spectral attention (MHSA) module, the gated-dconv feedforward network (GDFN) module, and the spectral enhancement (SpeE) module. The MHSA module learns the relationships among different bands and emphasizes the local spatial information. The GDFN module explores more expressive and discriminative spectral features. The SpeE module enhances the perception of subtle differences between different spectrums. Moreover, unlike the previous Transformer denoising method that can only handle fixed bands, the proposed method combines 3-D convolution and spectral–spatial attention Transformer blocks, enabling the denoising of HSI with an arbitrary number of bands. Experimental results demonstrate that TDSAT outperforms compared methods. The code is available at https://github.com/Featherrain/TDSAT.**

***Index Terms*— 3-D, denoising, hyperspectral image (HSI), spatial–spectral self-attention, Transformer.**

Received 4 June 2024; revised 31 July 2024; accepted 9 September 2024. Date of publication 11 September 2024; date of current version 27 September 2024. This work was supported in part by the Open Fund of State Key Laboratory of Remote Sensing Science under Grant OFSLRSS202301, in part by the Fundamental Research Funds for the Central Universities under Grant 3132024262, in part by China Postdoctoral Science Foundation under Grant 2023M740460 and Grant 2022T150080, and in part by the National Natural Science Foundation of China under Grant 62401095 and Grant 42471380. *(Corresponding author: Haoyang Yu.)*

Qiang Zhang is with the Center of Hyperspectral Imaging in Remote Sensing (CHIRS), Information Science and Technology College, Dalian Maritime University, Dalian 116026, China, and also with the State Key Laboratory of Remote Sensing Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China (e-mail: qzhang95@dlmu.edu.cn).

Yushuai Dong, Yaming Zheng, Haoyang Yu, and Meiping Song are with CHIRS, Information Science and Technology College, Dalian Maritime University, Dalian 116026, China (e-mail: 15047694784@163.com; zym1505@dlmu.edu.cn; yuhy@dlmu.edu.cn; smping@163.com).

Lifu Zhang is with the State Key Laboratory of Remote Sensing Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China (e-mail: zhanglf@radi.ac.cn).

Qiangqiang Yuan is with the School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China (e-mail: yqiang86@gmail.com).

Digital Object Identifier 10.1109/TGRS.2024.3458174

I. Introduction

With the rapid development of remote sensing imaging technology and intelligent interpretation algorithms, remote sensing images have become indispensable data in various applications, such as land survey [1], [2], agriculture and forestry monitoring [3], and military warning [4]. HSI is widely applied in various applications due to its high spectral resolution. However, noise reduction remains a significant issue in this field. Despite the primary advantage of spectral resolution in HSI, noise contamination due to environmental factors and sensor limitations persists as a crucial challenge for interpreting HSI data [5], [6]. HSIs are usually affected by various types of noise [7], [8]. Noise may greatly degrade the quality of HSI, thereby affecting the performance of subsequent tasks such as unmixing [9], feature learning [10], classification [11], and target detection [12]. Therefore, denoising techniques aim to reduce noise in hyperspectral images (HSIs), thereby enhancing data quality and improving the accuracy and reliability of subsequent analysis and applications.

Up until now, numerous prior- and learning-based methods for HSI denoising have been proposed [13]. Prior-based methods usually constrain the solution space of the ill-posed problem by constructing certain priors. These priors include total variation [14], [15], sparsity [16], [17], low-rank [18], [19], [20], [21], [22], nonlocal similarity [23], and so on. Such methods regularize the model and achieve better HSI denoising effects. In addition, low-rank tensor representation-based methods leverage the structural correlation and low-rank priors [24], [25] to extract structural information from HSI for denoising tasks. However, prior-based methods mostly rely on manually designed prior knowledge and need to be solved by intricate iterative optimization.

In contrast, learning-based methods [7], [26], [27], [28], [29], [30] could effectively and flexibly address the HSI denoising problem by utilizing the power of deep neural network. In recent years, inspired by the advancements in natural image denoising, researchers have applied deep learning models to HSI denoising. For instance, promising results have been achieved by harnessing spectral–spatial correlation through a deep residual convolutional network [31]. Several studies have separately extracted the spatial and spectral features of HSI,

employing cascaded convolutional blocks or multibranch networks for parallel feature extraction [32], [33]. Nevertheless, the majority of methods handle HSI following the strategy for RGB image denoising, disregarding the inherent spectral neighboring dependence of HSI noise. Furthermore, hybrid networks based on 3-D convolution and 3-D quasi-recurrent pooling have been designed to simultaneously capture spatial and spectral features, showing the benefits of both convolutional and recurrent neural networks [34], [35], [36], [37]. By large-scale training and adaptive learning, these methods are capable of learning features and noise distributions within HSI data, whereas most of the works that utilize convolutional neural networks demonstrate inadequate performance in capturing long-range contextual details, which is also a drawback of convolution.

In recent years, Transformer has been widely applied to visual tasks due to its powerful ability to model long-range dependencies [38], [39]. For instance, Dosovitskiy et al. [40] first used a Transformer to segment images into small patches for image recognition. Liu et al. [41] proposed the Swin Transformer, which performs self-attention with moving windows in the spatial domain, effectively reducing computational complexity. In addition, Zamir et al. [42] utilized the Restormer model to learn multiscale local–global representations, without decomposing them into local windows.

Compared with RGB images, HSI typically contains more features due to the rich bands. Current Transformer-based denoising methods for HSIs mainly focus on employing 2-D Swin Transformer and spectral Transformer [43], [44]. However, employing a 2-D Transformer for HSI denoising alters the inherent 3-D structure of HSI, such as global spectral correlations or spatial self-similarity. The UNFOLD method [45] treats HSI denoising as a 3-D task by partitioning hyperspectral data into nonoverlapping cubes and flattening them into 1-D vectors, utilizing 3-D Transformers, 3-D CNN, and 3-D U-Net to synergize spatial and spectral information. To reduce computational burden, patch partitioning disrupts the holistic nature of 3-D hyperspectral data. Therefore, the effective exploration of global spatial–spectral similarity is still a challenge for Transformer-based HSI denoising. In addition, there are few researches on the integrated processing spectral domains via 3-D Transformer.

To tackle the above challenges, we propose a novel method for HSI denoising, named the 3-D spatial–spectral attention Transformer (TDSAT) model. The 3-D Transformer takes the HSI as a whole 3-D input, rather than as multichannel 2-D data, thereby better preserving the complete spectral information of each pixel. Unlike existing Transformer approaches, TDSAT models the HSI bands in the spectral domain rather than the spatial domain, which reduces computational complexity while facilitating the learning of spectral characteristics. This approach leverages the rich spectral features of HSI while also capturing its spatial information. Operating by an encoder–decoder framework, this model employs a 3-D spectral attention mechanism to capture the spectral correlation and nonlocal spatial similarity in HSI.

Overall, the contributions of this work could be summarized as follows.

1) The proposed model could effectively exploit the intrinsic characteristics of HSI in both spectral and spatial dimensions. By focusing on modeling in the spectral domain, it effectively captures the correlation between adjacent bands, as well as the dependence between distant bands for HSI denoising.
2) Addressing the inherent spectral correlations in HSI, the multihead spectral attention (MHSA) module learns the relationships between different bands. The gated-dconv feedforward neural network module excavates more expressive and discriminative spectral features. The spectral enhancement (SpeE) module enhances the perception of subtle differences between different spectrums. Simultaneously, joint 3-D convolution can effectively utilize spatial information.
3) In contrast to most Transformer-based methods that could only handle the fixed number of bands for HSI, TDSAT can directly denoise HSI with an arbitrary number of bands. In addition, the model is lightweight and efficient, exhibiting favorable performance and low computational costs.

The remainder of this article is organized as follows. Section II reviews existing methods related to HSI denoising. Section III provides a brief introduction to the degradation model and presents detailed descriptions of the proposed method. Section IV conducts an analysis of experimental results and ablation studies. Finally, Section V gives the conclusion of this article.

II. Related Works

Currently, most HSI denoising methods could be classified into prior-based methods and learning-based methods. Here, we provide an overview of the relevant work in this area.

A. Prior-Based Methods

Most prior-based methods utilize prior knowledge for HSI denoising. Priors such as total variation, sparse representation, low-rank matrix, and low-rank tensor are employed to model the HSI data. For instance, Maggioni et al. [46] proposed a block-matching 4-D (BM4D) model, which exploits local correlation within each HSI subcube and nonlocal correlation among different HSI subcubes. In addition, Peng et al. [47] introduced the tensor dictionary learning (TDL) method, which considers the nonlocal similarity in the spatial domain of HSI and global correlation across different bands. Furthermore, Zhang et al. [18] developed an efficient recovery method based on the low-rank matrix. To combine spatial nonlocal similarity with low-rank characteristics of the global spectrum, He et al. [23] proposed a unified HSI denoising model called NGmeet. Xue et al. [48] constructed a nonlocal low-rank regularized tensor decomposition model, effectively utilizing the global spectral correlation and nonlocal self-similarity for HSI denoising. Besides, Chen et al. [49] attempt to model the HSI noise using a non-i.i.d. mixture of Gaussians. However, the main drawback of these methods is that the introduced prior usually only reflects certain characteristics of HSI. Moreover, the optimization procedure of prior-based methods is often complex and time-consuming.

B. Learning-Based Methods

This category of methods addresses HSI denoising in an end-to-end manner, employing a data-driven strategy. Learning-based methods alleviate the need for intricate manual feature design and parameter tuning across various scenarios by training massive labeled data. For example, inspired by the success of the deep convolutional neural network for natural image denoising [50], Chang et al. [51] introduced a deep convolutional neural network for HSI denoising. Besides, Yuan et al. [31] proposed an HSI denoising convolutional neural network (HSID-CNN), which restores HSI using two parallel feature extraction branches. Subsequently, considering the directionality of spatial information and the diversity of spectral information, Zhang et al. [32] developed a spatial–spectral gradient network (SSGN) to remove mixed noise in HSI. Guan et al. [52] proposed a recursive convolutional neural network (DnRCNN) to explore the interband and intraband correlations for HSI destriping. These approaches demonstrate that extraction of spatial and spectral features is beneficial for HSI denoising. However, these approaches also exhibit a common drawback: neglecting the intrinsic spatial–spectral correlation within HSI. It still remains exploration potential and application prospects in fully leveraging this correlation for HSI denoising.

To address this issue, methodologies utilizing 3-D convolution and its variations have proven pivotal in capturing local spatial–spectral correlations for HSI denoising. For instance, Dong et al. [53] designed a 3-D U-Net denoising model, which decomposes 3-D convolution into 2-D spatial convolution and 1-D spectral convolution. Wei et al. [35] introduced a 3-D convolutional quasi-recurrent neural network (QRNN3D) that combines 3-D convolution for extracting spatial–spectral correlation features with quasi-recurrent networks. Pan et al. [36] separately utilized the 3-D spectral and spatial components to extract the inherent spectral and spatial characteristics of HSI noise. Lai and Fu [37] developed a mixed attention network (MAN), which simultaneously considers spectral correlation and the interaction between spatial–spectral features at different scales for HSI denoising. However, owing to the constrained long-sequence modeling capability of CNN, these methods frequently fall short in capturing the comprehensive global spatial–spectral correlation in HSI, resulting in suboptimal exploitation of 3-D HSI.

In recent years, the use of the Transformer model for hyperspectral information processing has demonstrated remarkable achievements. For instance, Pang et al. [54] proposed a 3-D quasi-recurrent and Transformer network (TRQ3DNet), which combines 3-D quasi-recurrent layers with Swin blocks to denoise HSI. Besides, SERT [43] leveraged the global low rankness and nonlocal similarity of spatial–spectral cubes, which effectively remove noise in HSI. Furthermore, SST [44] utilizes a spatial Transformer to model spatial correlation and a spectral Transformer to model spectral correlation. However, limited by hardware conditions and the square growth of the Transformer, these approaches still have the limitation of realizing global spatial–spectral correlation modeling and capturing both interband and intraband correlations.

As 3-D data, HSI exhibits strong correlations across all dimensions. The 3-D Transformer [55] can better capture long-range dependencies among HSI. Therefore, we propose a 3-D spatial–spectral attention Transformer (SAT) model for HSI denoising. By emphasizing modeling in the spectral domain, the proposed model effectively harnesses spectral correlations, enabling comprehensive modeling of global spatial–spectral correlations. This approach leverages the rich spectral features of HSI data while effectively preserving the spatial structural characteristics and spectral fidelity of HSI. Subsequently, a detailed description of this model is given in the following.

III. Method

In this section, we first provide a brief introduction to the noise degradation model of HSI. Subsequently, we present the overall architecture of the proposed model and the specific module designs. Finally, a particular explanation of the training and optimization of the model is provided.

A. HSI Noise Degradation Model

In this article, the noisy HSI is denoted as $\boldsymbol{Y} \in \mathbb{R}^{H\times W\times C}$, where H, W, and C represent the height, width, and spectral number of the HSI, respectively. Generally, the noise degradation model of HSI could be expressed as

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

where $\boldsymbol{X} \in \mathbb{R}^{H\times W\times C}$ represents the noise-free HSI. $\boldsymbol{N} \in \mathbb{R}^{H\times W\times C}$ stands for the interfering noise. In real-world scenarios, HSI may be contaminated by various types of noise, such as Gaussian noise, stripe noise, deadline noise, impulse noise, and mixed noise. The object of HSI denoising is to estimate the noise-free HSI from the noisy HSI.

B. Overall Model Architecture and Specific Module Design

The overall structure of the proposed TDSAT for HSI denoising is illustrated in Fig. 1. The feature extractor and reconstructor use innovative components, such as 3-D spectral–spatial attention blocks, to capture detailed image features. A bottleneck design compresses and expands feature maps, boosting feature expressiveness and reuse. By employing upsampling and downsampling operations, the network can acquire features from images at different scales, leveraging contextual information from diverse image regions while maintaining the same computational cost. Unlike traditional skip connections, TDSAT employs a learnable feature fusion skip connection to control the flow of low-level features from the encoder to the decoder. The proposed 3-D SAT module could better facilitate the extraction of spectral correlations and spatial similarities for 3-D HSI restoration. Different from most Transformer-based denoising methods that could only handle HSI with a fixed number of bands, TDSAT is capable of processing HSI with an arbitrary number of bands.

Given a degraded HSI $\boldsymbol{Y} \in \mathbb{R}^{H\times W\times C\times 1}$, TDSAT first obtains low-level feature embeddings through the feature extraction block. The embedded features $\boldsymbol{F_o} \in \mathbb{R}^{H\times W\times D\times C}$ are further processed by a three-layer channel-encoding Transformer

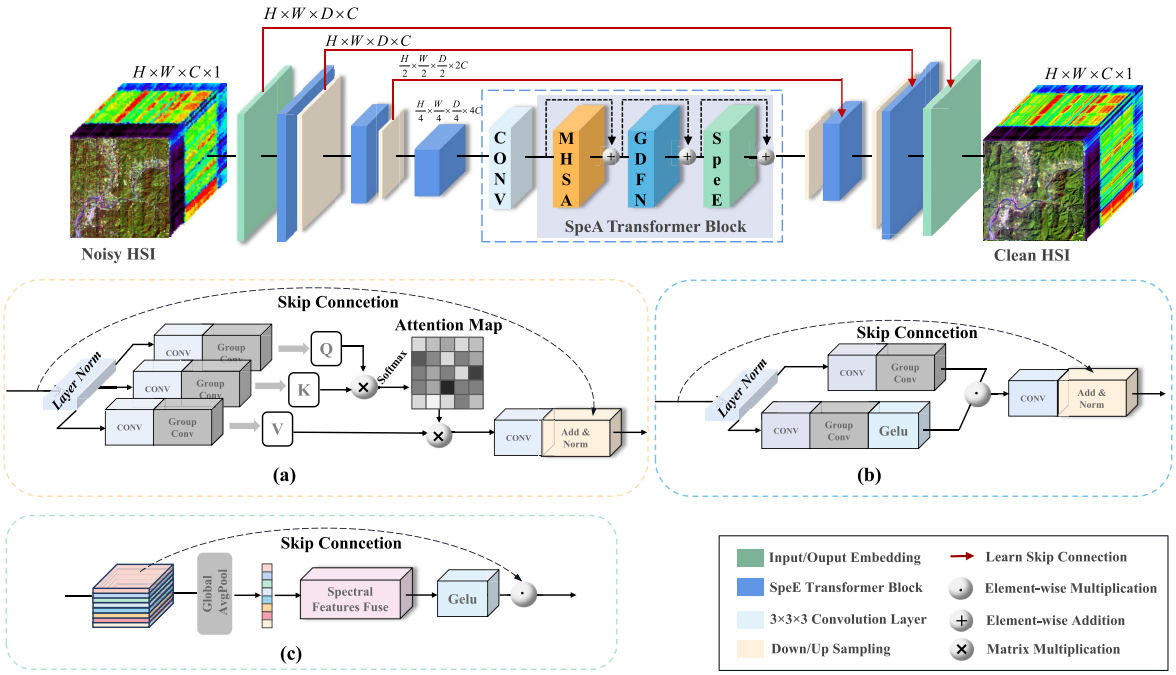


Fig. 1. Overall architecture of 3-D spatial–spectral attention Transformer (TDSAT). (a) MDTA. (b) GDFN. (c) SpeE.

block to obtain multiscale features. Here, H represents height, which denotes the vertical pixel count of the HSI. W stands for width, indicating the horizontal pixel count of HSI. C denotes the spectral dimension, representing the number of different bands in HSI. D stands for the number of feature maps, i.e., the number of feature maps at a particular layer in the network. Subsequently, symmetrical decoders are employed to obtain deep features $\boldsymbol{F_r}$ of the HSI. Finally, the deep features are reconstructed to obtain the noise-free image $\boldsymbol{X}$. We construct the entire network by stacking the SAT block to fully exploit interspectral and intraspectral correlations.

The TDSAT model utilizes an MHSA module and a gated-dconv feedforward network (GDFN) module. The MHSA module performs self-attention along the spectral dimension rather than the spatial dimension. This module implicitly encodes global contextual information by computing attention over the spectral domain, enabling global attention computation under linear computational complexity. In addition, considering the rich spectral information in HSI, TDSAT further develops an SpeE module to better leverage spectral information.

To preserve spatial features across layers, TDSAT further incorporates 3-D convolution operations before each SAT block. By combining 3-D convolutions and Transformer, TDSAT effectively captures local spatial information for HSI restoration. In contents, we provide the detailed explanations of MHSA, GDFN, SpeE and LSC modules.

1) Multihead Spectral Attention Module: First, given that HSI data comprise contiguous spectral bands, the interband correlations are crucial for image analysis and understanding. The MHSA operates by concurrently processing multiple attention heads to learn relationships between different bands. Each head focuses on capturing distinct aspects of spectral information. Second, the MHSA module learns interband relationships through attention weights, guiding the restoration of HSIs. This attention mechanism enables the model to dynamically focus on the most critical spectral features for the current task, thereby enhancing processing accuracy. As shown in Fig. 1(a), the MHSA module performs channel-level attention computation on the input feature map, using a multihead attention mechanism. First, the input feature map $\boldsymbol{F} \in \mathbb{R}^{H\times W\times D\times C}$ is normalized and processed through $1 \times 1 \times 1$ convolutions to encode channel context information. Then, the convolution operation at the channel level is applied to generate query matrices $\mathbf{Q} \in \mathbb{R}^{C\times HWD}$, key matrices $\mathbf{K} \in \mathbb{R}^{HWD\times C}$, and value matrices $\mathbf{V} \in \mathbb{R}^{HWD\times C}$. It uses depthwise separable convolutions to enhance the feature representation of local information. Subsequently, L_2 normalization is applied to the dot product between the query and key matrices, followed by the computation of attention weights $\mathbf{A} \in \mathbb{R}^{C\times C}$. Finally, the attention weights are multiplied by the value matrix and summed to obtain the weighted feature representation. Through this process, the proposed model could learn the relationships between specific bands and extract relevant information, achieving the goal of feature enhancement. The output of the MHSA module could be represented as

$$\hat{\boldsymbol{F}} = \boldsymbol{W}_p \,\text{Attention}\big(\hat{\boldsymbol{Q}}_c, \hat{\boldsymbol{K}}_c, \hat{\boldsymbol{V}}_c\big) + \boldsymbol{F} \tag{2}$$

$$\text{Attention}\big(\hat{\boldsymbol{Q}}_c, \hat{\boldsymbol{K}}_c, \hat{\boldsymbol{V}}_c\big) = \hat{\boldsymbol{V}}_c \cdot \text{Soft}\max\big(\hat{\boldsymbol{Q}}_c \cdot \hat{\boldsymbol{K}}_c/\alpha\big) \tag{3}$$

where α is a learnable scaling parameter to control the magnitude of the dot product output. $\boldsymbol{W}_p$ is a linear projection matrix.

2) GDFN Module: The GDFN module utilizes gating mechanisms to control information flow, incorporating spatial and spectral information adjustments after capturing local

spatial features through convolution and global spectral features through the MHSA. In standard Transformer networks, the multilayer perceptron (MLP) enhances the model's feature representation capability. To better utilize global spectral features, we propose replacing the typical MLP feedforward network with our introduced GDFN. As shown in Fig. 1(b), the GDFN module first processes the input feature map using convolutional operations and learns the relationships between different bands through channelwise grouped convolutions. Subsequently, it introduces a gating mechanism through the GELU activation function and elementwise multiplication to control the flow of information and selectively enhance features. Finally, convolutional operations are used to project the gated feature representation back to the original number of channels, outputting the final feature maps.

In the GDFN module, the gating mechanism could suppress features with less information. It enables the model to learn the more expressive and discriminative feature representations, thereby improving the model's reconstruction performance. The gating mechanism selectively allows useful information to be passed to the next layer of the network. The approach employing the gating mechanism in the GDFN module is denoted as

$$\hat{\boldsymbol{F}} = \boldsymbol{W}_{\boldsymbol{p}}^{o}\text{Gating}(\boldsymbol{F}) + \boldsymbol{F} \tag{4}$$

$$\text{Gating}(\boldsymbol{F}) = \Phi\left(W_c^1 W_p^1 \text{LN}(\boldsymbol{F})\right) \odot W_c^2 W_p^2 \text{LN}(\boldsymbol{F}) \tag{5}$$

where $\boldsymbol{F}$ and $\hat{\boldsymbol{F}}$ represent the input and output features, respectively. $\odot$ denotes elementwise multiplication. Φ is a nonlinear activation layer. LN stands for a normalization layer. W_p^o, W_p^1, and W_p^2 are linear projection matrices. W_c^1 and W_c^2 are two $3 \times 3 \times 3$ channelwise convolutions.

3) SpeE Module: Different from natural images, HSI possesses a wider spectral range. In addition, neighboring bands in HSI exhibit similar spatial information. Moreover, due to the specific spectral reflectance properties, pixel values at the same spatial position across different bands are related to the material characteristics of objects. This highlights a significant characteristic of HSI, which is the interband correlation. MHSA and GDFN modules could be utilized to explore the correlation between different bands. These modules allow the proposed model to leverage features from more informative bands. However, if the individual features of each band do not have sufficient distinctiveness, these advantages may be diminished. Therefore, TDSAT employs the SpeE module to further enhance the features of each band for exploring the intraspectral correlation.

The specific structure of the SpeE module is illustrated in Fig. 1(c). First, the input features undergo a 3-D average pooling operation, which converts the feature maps of each channel into a scalar value. This value represents the average intensity of all features within this channel. Then, the spectral feature fusion (SFF) operation is applied to each channel's features, combining them with weighted sums to enhance the expressive power of these features. The normalized feature maps are elementwise multiplied with the input tensor to achieve feature weighting. Finally, the output tensor is returned as the module's result. The procedure could be formulated as follows:

$$\hat{\boldsymbol{F}} = (\text{GELU}(\text{SFF}(\text{ AveragePool }(\boldsymbol{F})))) \odot \boldsymbol{F} \tag{6}$$

$$\text{SFF}(\boldsymbol{F}) = \boldsymbol{F} \cdot \boldsymbol{W} \tag{7}$$

where $\boldsymbol{F}$ and $\hat{\boldsymbol{F}}$ represent the input and output feature maps, respectively. $\boldsymbol{W} \in \mathbb{R}^{C_i \times C_o}$ stands for the channel fusion weight. This module aims to enhance the ability to perceive subtle differences between different bands.

4) Learnable Skip Connection: Skip connections are a key component of the U-Net [53] architecture that distinguishes it from other network structures. It provides a direct and efficient way to recover the low-level information lost during the downsampling process in traditional encoder–decoder networks.

TDSAT employs learnable skip connections (LSCs) in place of traditional concat skip connections. It facilitates controlled feature transfer, consequently reducing the number of network parameters. LSC computes attention weights through two convolutional layers, explicitly weighting shallow features from the encoder and decoder. It enhances the more important spatial and spectral features for HSI denoising. First, the skip connection weights $\boldsymbol{W}_i$ for the ith layer are computed as follows:

$$\boldsymbol{W_i} = \sigma(\text{Conv}_{3\times3\times3}(\text{Tanh}(\text{Conv}_{1\times1\times1}([\boldsymbol{X}_i, \boldsymbol{Y}_i])))) \tag{8}$$

where $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$ correspond to the features from the ith layer of the encoder and decoder, respectively. The input features are concatenated along the channel dimension. Initially, the number of channels in the input is doubled using a convolutional layer. Subsequently, a nonlinear transformation is applied through the Tanh function. Following this, another convolutional layer is utilized to generate attention weights. These attention weights could learn feature relationships at different positions and scales. They generate the attention weights that adapt to different spatial dimensions. Finally, the sigmoid function is employed to constrain the attention weights between the range of 0–1

$$\boldsymbol{F_i} = (1 - \boldsymbol{W}_i) \odot \boldsymbol{X}_i + \boldsymbol{W}_i \odot \boldsymbol{Y_i} \tag{9}$$

where $\boldsymbol{F_i}$ represents the output of the ith layer skip connection in the decoder. The attention weights $\boldsymbol{W}_i$ computed by the convolutional layer are used to fuse the two features. $(1-\boldsymbol{W}_i)\odot\boldsymbol{X}_i$ denotes the feature attenuation operation for the encoder. $\boldsymbol{W}_i\odot\boldsymbol{Y}_i$ stands for the feature enhancement operation for the decoder

C. Model Training and Optimization

1) Loss Function: For image denoising, the L_1 and L_2 norms are commonly used [33] as the loss function. To strike a balance between noise removal and detail preservation, this work adopts the L_2 norm as the loss function for the model defined as

$$\text{Loss} = \frac{1}{N}\sum_{i=1}^{N}\left\|X_i - X_i^{\text{gt}}\right\|_2 \tag{10}$$

where N denotes the number of training patches. X_i and X_i^{gt} are patches from the denoised image and the original image, respectively.

2) Training Details: This work conducts model training and optimization via the ICVL dataset [56]. It comprises 201 HSIs with spatial dimensions of 1392 × 1300 and covers 31 bands ranging from 400 to 700 nm. During the training procedure, 100 HSIs are used for training, ten HSIs for validation, and 40 HSIs for testing. To enhance the training effectiveness, uniform stride cropping is employed to process the original HSI data into multiple overlapping cubes. Each cube has a spatial dimension of 64 × 64 while maintaining the same number of bands. In addition, a random rotation and scaling strategy is introduced to augment the sample size.

To train a more robust HSI denoising model, the training procedure is divided into three stages, progressively increasing the noise intensity and types. In the first training stage, encompassing the initial ten epochs, HSI polluted with Gaussian noise of intensity $\sigma = 50$ is used as the training set. This stage aims to enable the model to rapidly converge and effectively recovery HSI disturbed by Gaussian noise. During the second training stage, spanning from epoch 10 to epoch 20, the optimization continues using HSI contaminated with Gaussian noise, with noise intensities uniformly sampled from 30 to 70. The objective of this stage is to further improve the model's adaptability to varying intensities of Gaussian noise. In the third training stage, from epoch 20 to epoch 50, the model is optimized via random combinations of different noise types from Case 1 to Case 4 (details are provided in Section IV). The purpose of this stage is to enable the model to handle multiple types of noise, thereby enhancing its robustness in practical scenarios.

The model utilizes the ADAM [57] for parameter optimization, with a fixed learning rate set to 1e-4 and a batch size of 8. For the software, the proposed model is implemented using the PyTorch deep learning framework. For the hardware, the proposed model makes use of an NVIDIA RTX 4090 GPU, i9-12900K CPU, and 64-GB RAM.

IV. Experiments

In this section, we compare the results of different methods on simulated and real HSI denoising experiments. In addition, in the discussion part, we provide the results of ablation experiments and complexity analysis of the proposed model.

A. Experimental Settings

1) Test Dataset: We conducted simulated experiments using 40 ICVL HSIs, each cropped to a final dimension of 512 × 512 × 31, including Gaussian denoising at various levels of intensity ($\sigma = 30$, $\sigma = 50$, $\sigma = 70$, and $\sigma = 30-70$ no-reference noises), as well as complex noise denoising (from Case 1 to Case 5). Simulated experiments enable the evaluation of different methods' denoising effectiveness under various intensities and types of noise. In addition, we also carry out the real HSI denoising experiments on the Urban, EO-1 Hyperion, and Gaofen-5 Shanghai datasets. Simulated experiments enable the evaluation of different methods' denoising effectiveness under various intensities and types of noise.

2) Comparative Methods and Evaluation Metrics: To evaluate the denoising effectiveness of the proposed method, we compare it with existing deep learning-based methods, QRNN3D [35], UNFOLD [45], SQAD [36], MAN [37], and SERT [43], and the prior-based methods, LRTV [15], RCTV [58], and NGMeet [23]. All deep learning-based methods are trained and tested under the same conditions to ensure fairness. We employ three quantitative evaluation metrics: peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and spectral angle mapper (SAM) to assess the performance of different HSI denoising methods. The first two metrics are used to measure the quality of spatial information reconstruction. The last metric is utilized to quantify the degree of spectral distortion. In general, higher PSNR/SSIM values and lower SAM values indicate better denoising effects.

B. Simulation Experiments

1) Gaussian Noise Removal on the ICVL Dataset: In this experiment, we apply Gaussian noise of various intensities (including 30, 50, and 70 and random intensity ranging from 30 to 70) to the ICVL dataset to evaluate the performance of different HSI denoising methods, as listed in Table I.

Table I presents the comparison of denoising results on the ICVL dataset at different Gaussian noise levels. Among the prior-based methods, NGMeet exhibits the best performance, outperforming the deep learning-based methods QRNN3D and SQAD. It could be observed that the proposed method achieves the best performance among all denoising methods, with a significant improvement over the comparison methods. In all the different Gaussian noise scenarios, TDSAT's PSNR is at least 0.43 dB higher than the other deep learning-based methods.

To further evaluate the denoising performance of TDSAT, Fig. 2 illustrates the denoising results of different methods at the Gaussian noise intensity of 50. To facilitate comparison, we select the local regions for magnified displaying. Among the prior-based methods, RCTV performs poorly, exhibiting significant noise residue. LRTV manages to remove Gaussian noise, whereas its spectral distortion issue is severe. NGMeet effectively eliminates noise through nonlocal self-similarity prior at the cost of oversmoothing in complex textures. Learning-based approaches harness their data-driven learning capability, consistently outperforming model-based methods. Compared with both prior- and deep learning-based methods, the proposed method achieves higher fidelity and superior denoising performance. The main reason lies in the fact that TDSAT takes the global spectral correlation and spatial similarity of HSI into account, thereby demonstrating better restoration capability.

2) Complex Noise Removal on the ICVL Dataset: In the experiments for complex noise removal, we simulate the complex noise scenarios by artificially applying various types of noise to the ICVL dataset. Five cases are designed as follows.

Case 1 (Non-i.i.d. Gaussian Noise): Gaussian noise with zero mean is added to each band, and the noise intensity is randomly set within the range of [10, 70].

TABLE I
AVERAGED DENOISING RESULTS OF DIFFERENT METHODS UNDER DIFFERENT GAUSSIAN NOISE LEVELS ON THE ICVL DATASET

Method	30			50			70			Blind		
	PSNR	SSIM	SAM	PSNR	SSIM	SAM	PSNR	SSIM	SAM	PSNR	SSIM	SAM
Noisy	18.59	0.1110	0.6840	14.15	0.0484	0.8685	11.22	0.0267	0.9941	16.54	0.1039	0.7687
LRTV	35.78	0.9188	0.1005	39.74	0.9251	0.2327	32.23	0.8522	0.1533	33.09	0.8623	0.2112
RCTV	36.50	0.9026	0.1323	33.76	0.6905	0.5159	30.86	0.6602	0.2498	33.88	0.8341	0.2733
NGMeet	39.41	0.9066	0.1201	38.81	0.9042	0.2475	37.44	0.9281	0.1201	38.90	0.9060	0.1383
QRNN3D	39.94	0.9498	0.0941	38.28	0.9309	0.0968	36.32	0.8920	0.1280	38.95	0.9356	0.1045
SQAD	39.91	0.9493	0.0920	37.56	0.9233	0.1111	34.69	0.8602	0.1796	38.47	0.9211	0.1352
UNFOLD	41.39	0.9601	0.0538	39.21	0.9394	0.0622	37.39	0.9712	0.0763	40.05	0.9470	0.0642
MAN	41.99	0.9675	0.0481	39.85	0.9505	0.0571	37.98	0.9311	0.0747	40.76	0.9567	0.0626
SERT	42.41	0.9703	0.0501	40.24	0.9540	0.0598	37.81	0.9317	0.0825	41.21	0.9602	0.0659
TDSAT	43.01	0.9727	0.0407	40.67	0.9564	0.0482	38.93	0.9400	0.0580	41.78	0.9637	0.0490

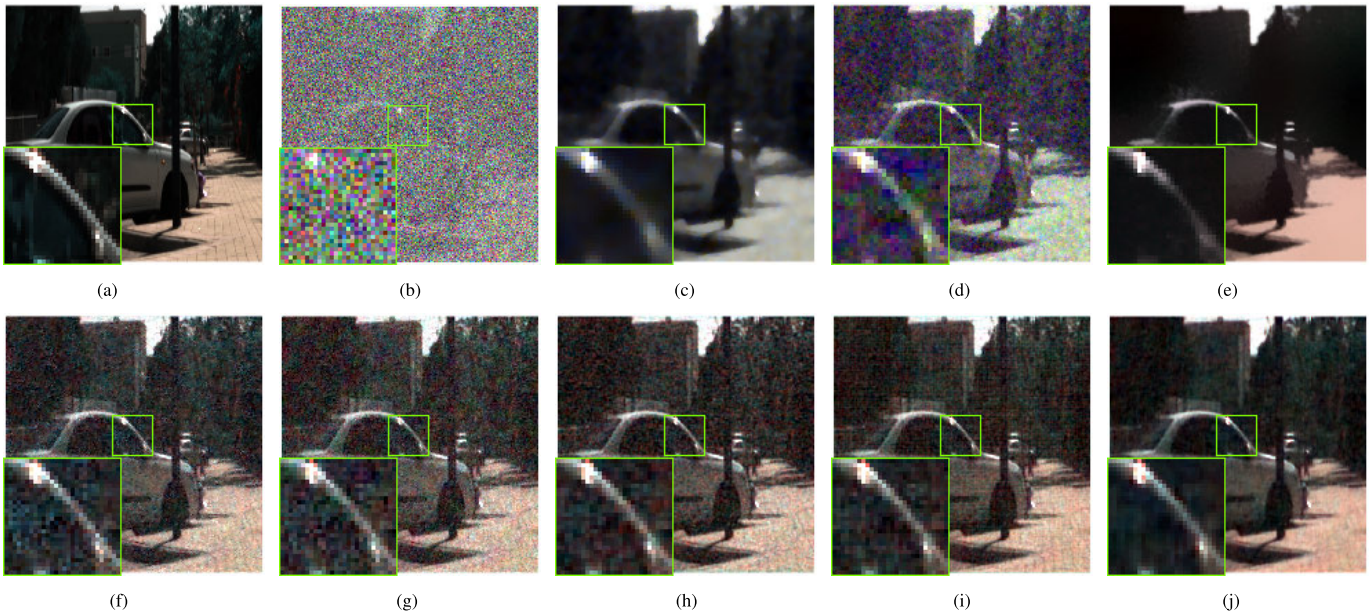


Fig. 2. Gaussian denoising results ($\sigma = 50$) for the bands (28, 18, and 15) on ICVL Labtest_0910-1504 data. (a) Ground truth. (b) Noisy. (c) LRTV. (d) RCTV. (e) NGMeet. (f) QRNN3D. (g) SQAD. (h) MAN. (i) SERT. (j) TDSAT.

Case 2 (Gaussian + Stripe Noise): Gaussian noise is applied to all the bands. In addition, 30% random bands are contaminated with stripe noise, with a simulated stripe noise ratio ranging of [5%, 15%].

Case 3 (Gaussian + Deadline Noise): Similar to Case 2, the only difference is that we apply dead line noise instead of stripe noise.

Case 4 (Gaussian + Impulse Noise): Gaussian noise is applied to all the bands. In addition, 30% random bands are contaminated with impulse noise, with an intensity ranging of [10%, 70%].

Case 5 (Mixture Noise): Gaussian noise is applied to all the bands. In addition, 30% random bands are contaminated with the noise types from Case 1 to Case 4 mentioned above.

Table II presents the quantitative metrics results of different methods under five complex noise cases on the ICVL dataset. One HSI is selected from each scenario, and the denoising results by different methods are shown in Fig. 3. To facilitate comparison, some details are magnified for better visualization.

In prior-based methods, NGMeet achieves better denoising results due to its advantage in tensor modeling and consideration of nonlocal low-rank prior. However, LRTV and RCTV significantly deteriorate when addressing complex mixture noise scenarios. Most prior-based methods assume that the noise distribution is the same across all bands, rendering them less effective in complex noise removal.

On the other hand, deep learning-based methods achieve fine denoising performance due to their powerful learning capability, as shown in Fig. 3. Nevertheless, these methods still suffer from residual noise or loss of spatial details

TABLE II
AVERAGED DENOISING RESULTS OF DIFFERENT METHODS UNDER FIVE COMPLEX NOISE CASES ON THE ICVL DATASET

Method	Non-iid Gaussian			Stripe			Deadline			Impulse			Mixture		
	PSNR	SSIM	SAM	PSNR	SSIM	SAM	PSNR	SSIM	SAM	PSNR	SSIM	SAM	PSNR	SSIM	SAM
Noisy	17.79	0.1616	0.7915	17.75	0.1598	0.7838	17.68	0.1624	0.7942	15.11	0.1229	0.8415	14.24	0.1022	0.8549
LRTV	33.14	0.8895	0.0753	23.99	0.6482	0.1912	31.21	0.8007	0.1333	36.59	0.8960	0.1884	24.11	0.6822	0.1546
RCTV	34.78	0.8609	0.0755	36.21	0.9180	0.1150	33.56	0.8380	0.1148	34.59	0.8075	0.1293	30.26	0.7691	0.1545
NGMeet	35.33	0.8880	0.0697	37.04	0.9180	0.1310	34.41	0.8503	0.1058	35.26	0.8599	0.2067	31.16	0.8856	0.1344
QRNN3D	40.56	0.9612	0.0732	38.88	0.9358	0.0910	40.28	0.9602	0.0745	40.39	0.9603	0.0740	37.87	0.9301	0.0920
UNFOLD	41.51	0.9647	0.0441	40.33	0.9638	0.0461	39.24	0.9634	0.0454	39.46	0.9431	0.0752	38.60	0.9390	0.0728
SQAD	40.04	0.9543	0.0723	38.01	0.9257	0.0980	39.75	0.9533	0.0742	39.83	0.9538	0.0732	37.20	0.9213	0.0989
MAN	43.18	0.9768	0.0363	41.30	0.9601	0.0633	43.05	0.9765	0.0374	43.02	0.9766	0.0377	40.66	0.9580	0.0651
SERT	43.23	0.9767	0.0409	40.27	0.9535	0.0709	42.82	0.9763	0.0428	43.00	0.9764	0.0422	38.60	0.9472	0.0757
TDSAT	44.05	0.9801	0.0333	42.28	0.9674	0.0582	43.92	0.9798	0.0339	43.94	0.9800	0.0346	41.46	0.9655	0.0564

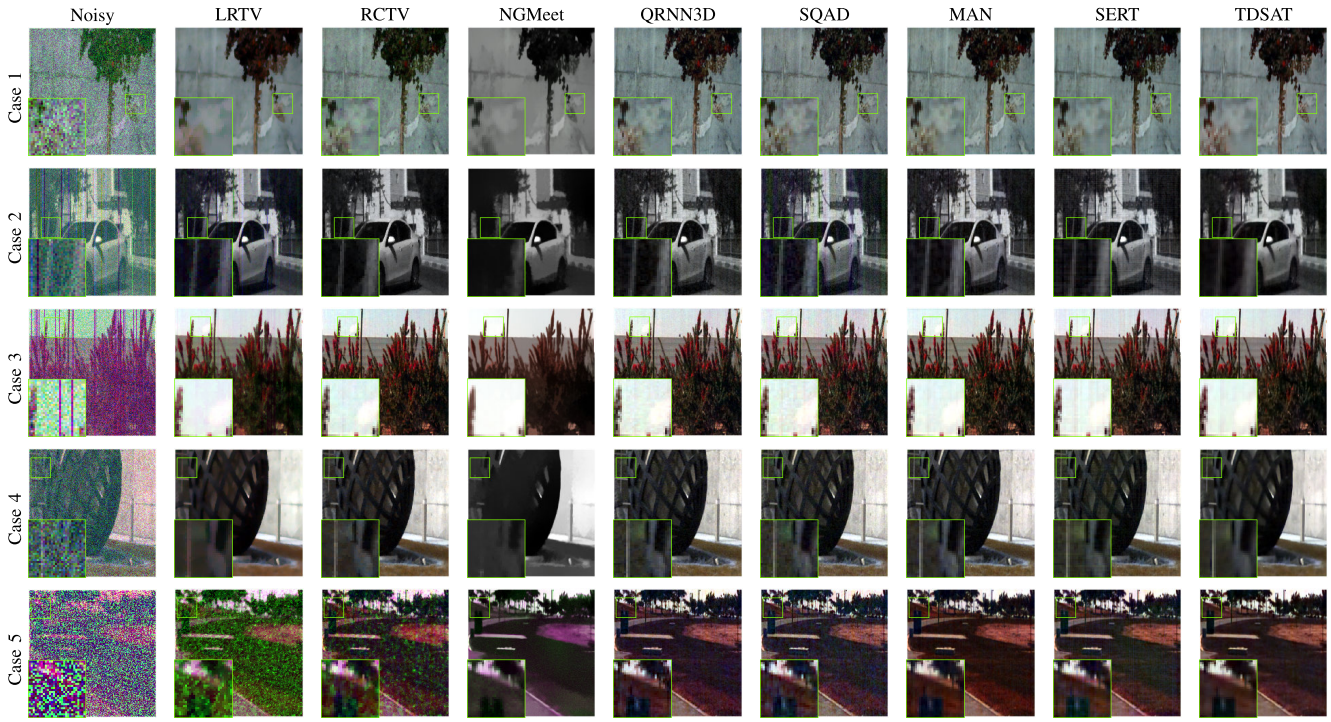


Fig. 3. Visual comparison of the denoised image for all the five cases of the ICVL dataset.

in the denoised results. Besides, they usually exhibit poor performance in removing stripe noise.

In comparison, the proposed TDSAT not only effectively removes complex noise but also preserves spatial details. It indicates that the proposed method has better capability in exploring spatial and spectral correlations.

C. Real Experiments

In this section, we carry out real HSI denoising experiments using the Urban, GF-5, and EO-1 noisy HSI data. The pretrained ICVL model is employed for processing these real datasets. However, it is noteworthy that SERT faces challenges when dealing with HSI data with varying numbers of spectral bands. To alleviate this issue, we partitioned the real HSI into multiple subimages, each containing 31 bands for testing SERT.

1) Urban: The spatial size of the Urban HSI data is 307 × 307, with 188 spectral bands. It covers a spectral range from 400 to 2500 nm. Some bands in this HSI are affected by atmospheric radiation interference, leading to mixed noise contamination including deadline, stripe, and Gaussian noise. Removing mixed noise is a highly challenging task. Fig. 4 provides the real experimental results of six contrastive algorithms and the proposed model on the Urban HSI data. It could be observed that TDSAT effectively removes mixed noise

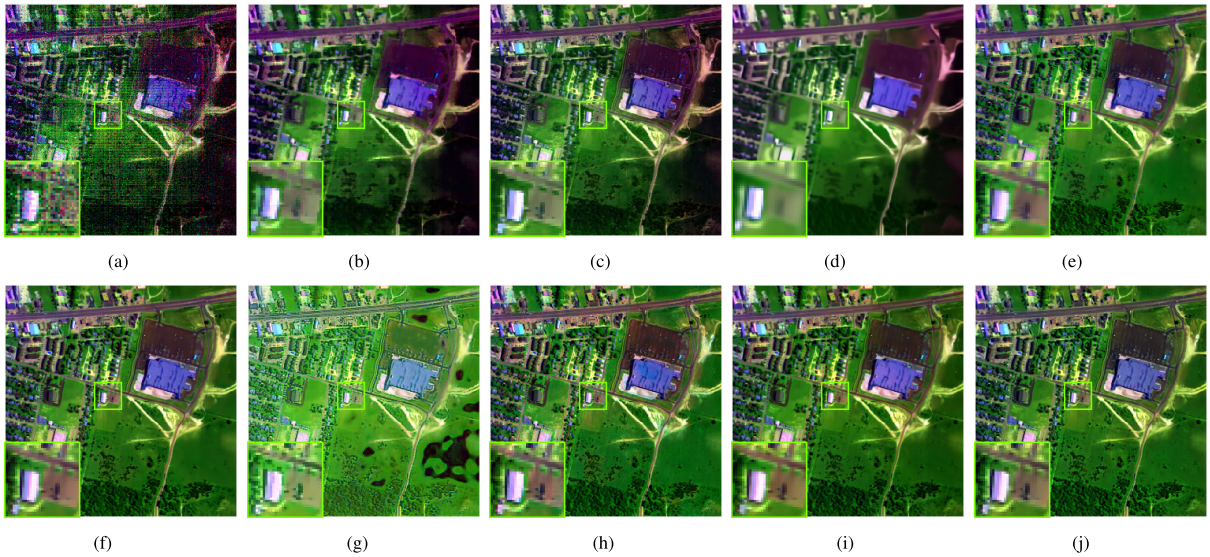


Fig. 4. Real denoised results on bands (188, 104, and 1) of the Urban HSI data. (a) Noisy. (b) LRTV. (c) RCTV. (d) NGMeet. (e) QRNN3D. (f) UNFOLD. (g) SQAD. (h) MAN. (i) SERT. (j) TDSAT.

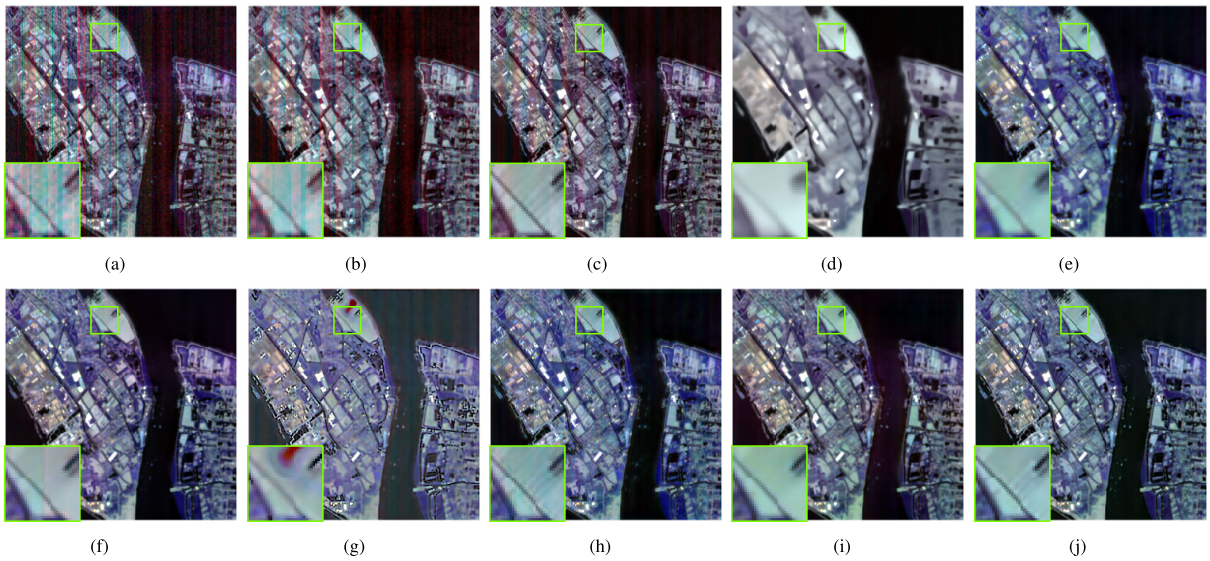


Fig. 5. Real denoised results on bands (152, 96, and 43) of the Gaofen-5 Shanghai dataset. (a) Noisy. (b) LRTV. (c) RCTV. (d) NGMeet. (e) QRNN3D. (f) UNFOLD. (g) SQAD. (h) MAN. (i) SERT. (j) TDSAT.

while preserving spatial details. NGMeet, due to the utilization of nonlocal similarity prior, results in oversmooth denoising results. QRNN3D, UNFOLD, and SQAD could remove the mixed noise. While QRNN3D sacrifices some texture detail information in the denoised results. SQAD exhibits noticeable spectral distortions and residual artifacts, possibly due to its strategy of separately processing spatial and spectral information. This strategy may affect the interband correlation. In comparison to the aforementioned methods, MAN, SERT, and the proposed method achieve superior denoising results.

2) Gaofen-5 Shanghai Dataset: The GF-5 satellite is the first full-spectrum hyperspectral satellite that can obtain comprehensive observations of the atmosphere and land in the world. The dataset captured by the GF-5 satellite in Shanghai contains Gaussian noise and stripe noise along with dense deadline and has been resized to dimensions of $300 \times 300 \times 155$. The pseudocolor denoising results are shown in Fig. 5. RCTV simultaneously characterizes the low-rank and local smooth properties, achieving excellent preservation of texture details. It could be observed that the experimental results of prior-based methods either result in excessive smoothness (NGMeet) or incomplete noise removal (LRTV and RCTV). Moreover, deep learning-based methods such as QRNN3D, UNFOLD, SQAD, and SERT suffer from loss of spatial details. In contrast, the denoising results of TDSAT are superior to other contrastive algorithms, demonstrating the effectiveness of the proposed approach.

3) EO-1 Hyperion Dataset: We utilize an EO-1 subimage with a spatial size of 304×304 and consisting of 166 spectral bands removed water absorption bands as the experimental data. This HSI scene is more complex than the former real Urban HSI. This dataset is severely contaminated by Gaussian noise, stripe noise, and deadline, leading to a degradation of HSI quality. The pseudocolor visualization and experimental

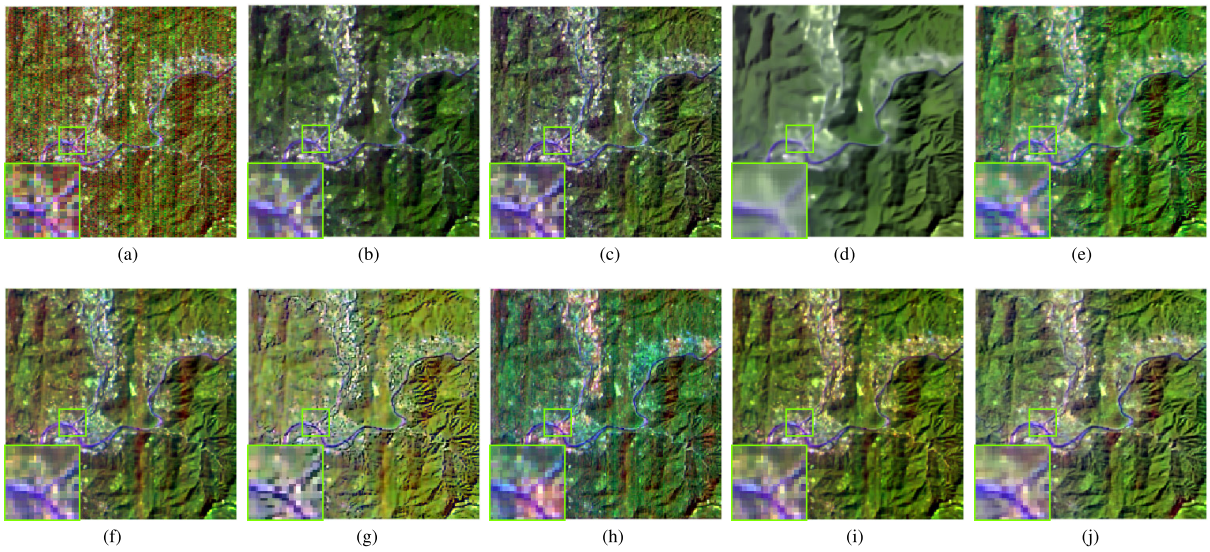


Fig. 6. Real denoised results on bands (166, 97, and 7) of the EO-1 Hyperion dataset. (a) Noisy. (b) LRTV. (c) RCTV. (d) NGMeet. (e) QRNN3D. (f) UNFOLD. (g) SQAD. (h) MAN. (i) SERT. (j) TDSAT.

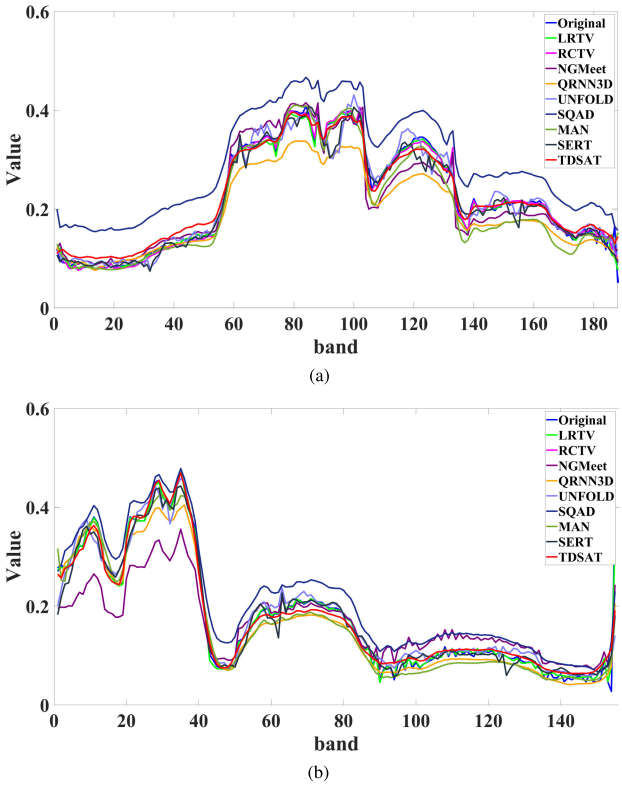


Fig. 7. Spectral curves corresponding to the real HSI denoising. (a) Spectral curves of Urban at pixel (1, 75). (b) Spectral curves of GF-5 at pixel (10, 10).

results of the EO-1 HSI data are shown in Fig. 6. Compared with the contrastive algorithms, the proposed method outperforms in terms of mixed noise removal and texture detail restoration. In summary, on the three real HSI datasets, the proposed method consistently achieves excellent denoising results, demonstrating the robustness and reliability of TDSAT.

4) Spectral Analysis: Compared to prior-based approaches, deep learning methods often suffer from spectral distortion issues. The spectral curves on the real HSI datasets are depicted in Fig. 7. It is evident that the proposed method could more accurately approximate the spectral curves of the original noise data compared with other deep learning-based HSI denoising methods. Furthermore, TDSAT exhibits smoother visual results on the real HSI datasets, which strongly indicates the effectiveness of exploring interspectral correlations in the proposed method.

TABLE III
DIFFERENT EMBEDDING DIMENSION COMPARISONS ON ICVL DATASET UNDER GAUSSIAN NOISE

dim	Params(M)	PSNR(dB)	SSIM	SAM	Time(s)
8	0.29	36.97	0.9220	0.0841	0.0210
12	0.63	38.40	0.9345	0.0721	0.0223
16	**1.09**	**40.21**	**0.9415**	**0.0665**	**0.0225**
20	1.69	40.28	0.9546	0.0583	0.0258

D. Discussion

1) Comparison of the Hidden Dim: The selection of the hidden channel dimension of the feature map is highly important for feature extraction. We analyzed the influence of the different embedding dimensions on the denoising, presented in Table III. By increasing the number of hidden embedding dimensions, the model gains enhanced capacity to capture and represent complex features within the data. Therefore, considering the balance between efficiency and performance optimization in the model, we opted for a hidden embedding size of 16. This decision enhances model performance without significantly raising computational complexity.

2) Ablation Analysis: To validate the effectiveness of the modules in TDSAT, we discuss the impact of each module on denoising performance. All modules undergo the same training process and are evaluated on the same noise dataset, as listed in Table IV. It is observed that the LSC, replacing the traditional concat skip connection, enhances the model's multiscale feature learning capability and significantly reduces the model parameters while retaining performance with minimal decline. The fusion of the 3-D convolution with the SAT module

TABLE IV
ABLATION STUDY OF THE PROPOSED MODULES IN TDSAT

LSC	Fusion	SpeE	Params (M)	GFLOPs	PSNR (dB)	SAM
×	×	×	0.83	1701.5	36.54	0.0962
✓	×	×	0.67	1493.5	36.49	0.0977
✓	✓	×	0.96	1828.6	37.76	0.0792
✓	✓	✓	1.09	1830.5	40.32	0.0561

TABLE V
COMPLEXITY COMPARISON OF DIFFERENT MODELS ON THE ICVL DATASET WITH DIMENSIONS OF 512 × 512 × 31

Metric	QRNN3D	SQAD	MAN	SERT	TDSAT
PSNR (dB)	38.95	38.47	40.76	41.21	**41.78**
Params (M)	0.86	**0.31**	0.82	1.91	1.09
GFLOPs	1256.8	825.8	962.0	478.8	**1830.5**
Time (s)	2.591	0.1120	0.1634	0.0709	**0.0208**

strengthens the exploration of spatial and spectral information, leading to a 1.27-dB improvement in PSNR performance. The impact of the SpeE module on denoising performance is most significant, highlighting the importance of utilizing spectral correlations in TDSAT.

3) Model Complexity Analysis: In Table V, we compare the denoising performance, parameter count, GFLOPs, and average runtime of different learning-based HSI denoising methods on the ICVL dataset. By introducing attention mechanisms in the spectral dimension, TDSAT achieves optimal performance with a reasonable parameter number when processing HSI data. Its advantage lies in its ability to effectively reduce computational complexity, thereby enhancing the model's efficiency and performance. As shown in Table IV, the proposed method's runtime is significantly less than that of the other deep learning methods. Besides, TDSAT achieves higher efficiency and shorter computational time, with the optimal HSI denoising results.

V. CONCLUSION

In this work, we propose a 3-D spectral–spatial attention Transformer (TDSAT) model for HSI denoising. By emphasizing spectral domain modeling, the TDSAT model effectively leverages global spatial–spectral correlations. It simultaneously explores the spectral correlations of HSI and nonlocal spatial similarity. The MHSA and feedforward neural network module could learn and capture interspectral features. The SpeE module enhances the spectral correlations. The combination of 3-D convolution with SAT further extracts joint spatial–spectral information for HSI denoising. Compared with existing HSI denoising methods, the proposed TDSAT demonstrates superior performance in both qualitative and quantitative experimental results. However, the "black-box" characteristic of the deep learning model may limit the interpretability in certain application scenarios. In future work, we will continue to explore more frameworks to enhance the model's interpretability and transparency.

REFERENCES

[1] J. Transon, R. d'Andrimont, A. Maugnard, and P. Defourny, "Survey of hyperspectral earth observation applications from space in the Sentinel-2 context," *Remote Sens.*, vol. 10, no. 2, p. 157, 2018.

[2] C. Weber et al., "Hyperspectral imagery for environmental urban planning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 1628–1631.

[3] M. S. Mohd Asaari et al., "Close-range hyperspectral image analysis for the early detection of stress responses in individual plants in a high-throughput phenotyping platform," *ISPRS J. Photogramm. Remote Sens.*, vol. 138, pp. 121–138, Apr. 2018.

[4] M. Shimoni, R. Haelterman, and C. Perneel, "Hypersectral imaging for military and security applications: Combining myriad processing and sensing techniques," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 101–117, Jun. 2019.

[5] B. Rasti, P. Scheunders, P. Ghamisi, G. Licciardi, and J. Chanussot, "Noise reduction in hyperspectral imagery: Overview and application," *Remote Sens.*, vol. 10, no. 3, p. 482, Mar. 2018.

[6] B. Rasti, Y. Chang, E. Dalsasso, L. Denis, and P. Ghamisi, "Image restoration for remote sensing: Overview and toolbox," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 201–230, Jun. 2022.

[7] L. Zhuang and M. K. Ng, "FastHyMix: Fast and parameter-free hyperspectral image mixed noise removal," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 8, pp. 4702–4716, Aug. 2023.

[8] T.-X. Jiang, L. Zhuang, T.-Z. Huang, X.-L. Zhao, and J. M. Bioucas-Dias, "Adaptive hyperspectral mixed noise removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5511413.

[9] Y. Ma et al., "Hyperspectral unmixing with Gaussian mixture model and low-rank representation," *Remote Sens.*, vol. 11, no. 8, p. 911, Apr. 2019.

[10] X. Li, M. Ding, and A. Pižurica, "Fully group convolutional neural networks for robust spectral–spatial feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5509314.

[11] X. Zhao, R. Tao, W. Li, W. Philips, and W. Liao, "Fractional Gabor convolutional network for multisource remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2021.

[12] C. Wu, B. Du, and L. Zhang, "Hyperspectral anomalous change detection based on joint sparse representation," *ISPRS J. Photogramm. Remote Sens.*, vol. 146, pp. 137–150, Dec. 2018.

[13] Q. Zhang, Y. Zheng, Q. Yuan, M. Song, H. Yu, and Y. Xiao, "Hyperspectral image denoising: From model-driven, data-driven, to Model-Data-driven," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 6, 2023, doi: 10.1109/TNNLS.2023.3278866. [Online]. Available: https://ieeexplore.ieee.org/document/10144690

[14] W. He, H. Zhang, H. Shen, and L. Zhang, "Hyperspectral image denoising using local low-rank matrix recovery and global spatial–spectral total variation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 3, pp. 713–729, Mar. 2018.

[15] 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. 2015.

[16] T. Xie, S. Li, and J. Lai, "Adaptive rank and structured sparsity corrections for hyperspectral image restoration," *IEEE Trans. Cybern.*, vol. 52, no. 9, pp. 8729–8740, Sep. 2022.

[17] Y. Chen, W. He, N. Yokoya, and T. Huang, "Hyperspectral image restoration using weighted group sparsity-regularized low-rank tensor decomposition," *IEEE Trans. Cybern.*, vol. 50, no. 8, pp. 3556–3570, Aug. 2020.

[18] 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. 2013.

[19] C. Li, Y. Ma, J. Huang, X. Mei, and J. Ma, "Hyperspectral image denoising using the robust low-rank tensor recovery," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 32, no. 9, pp. 1604–1612, 2015.

[20] 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. 4260–4268.

[21] Q. Zhang, Y. Dong, Q. Yuan, M. Song, and H. Yu, "Combined deep priors with low-rank tensor factorization for hyperspectral image restoration," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[22] Y. Chen, M. Chen, W. He, J. Zeng, M. Huang, and Y.-B. Zheng, "Thick cloud removal in multitemporal remote sensing images via low-rank regularized self-supervised network," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5506613.

[23] W. He, Q. Yao, C. Li, N. Yokoya, and Q. Zhao, "Non-local meets global: An integrated paradigm for hyperspectral denoising," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2089–2107, Jan. 2022.

[24] Q. Zhang, Q. Yuan, Z. Li, F. Sun, and L. Zhang, "Combined deep prior with low-rank tensor SVD for thick cloud removal in multitemporal images," *ISPRS J. Photogramm. Remote Sens.*, vol. 177, pp. 161–173, Jul. 2021.

[25] 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.

[26] Y. Chen, W. Lai, W. He, X.-L. Zhao, and J. Zeng, "Hyperspectral compressive snapshot reconstruction via coupled low-rank subspace representation and self-supervised deep network," *IEEE Trans. Image Process.*, vol. 33, pp. 926–941, 2024.

[27] O. Sidorov and J. Y. Hardeberg, "Deep hyperspectral prior: Single-image denoising, inpainting, super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3844–3851.

[28] Q. Zhang, Q. Yuan, J. Li, Z. Li, H. Shen, and L. Zhang, "Thick cloud and cloud shadow removal in multitemporal imagery using progressively spatio-temporal patch group deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 148–160, Apr. 2020.

[29] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[30] 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.

[31] Q. Yuan, Q. Zhang, J. Li, H. Shen, and L. Zhang, "Hyperspectral image denoising employing a spatial–spectral deep residual convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1205–1218, Sep. 2018.

[32] Q. Zhang, Q. Yuan, J. Li, X. Liu, H. Shen, and L. Zhang, "Hybrid noise removal in hyperspectral imagery with a spatial–spectral gradient network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7317–7329, Oct. 2019.

[33] 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.

[34] W. Liu and J. Lee, "A 3-D atrous convolution neural network for hyperspectral image denoising," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5701–5715, Aug. 2019.

[35] 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.

[36] E. Pan, Y. Ma, X. Mei, F. Fan, J. Huang, and J. Ma, "SQAD: Spatial–spectral quasi-attention recurrent network for hyperspectral image denoising," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5524814.

[37] Z. Lai and Y. Fu, "Mixed attention network for hyperspectral image denoising," 2023, *arXiv:2301.11525*.

[38] 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. 568–578.

[39] T. Ye et al., "Perceiving and modeling density is all you need for image dehazing," 2021, *arXiv:2111.09733*.

[40] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[41] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[42] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5728–5739.

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

[44] M. Li, Y. Fu, and Y. Zhang, "Spatial–spectral transformer for hyperspectral image denoising," in *Proc. AAAI*, 2023, pp. 1368–1376.

[45] A. Dixit, A. K. Gupta, P. Gupta, S. Srivastava, and A. Garg, "UNFOLD: 3-D U-Net, 3-D CNN, and 3-D transformer-based hyperspectral image denoising," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5529710.

[46] 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, Jan. 2013.

[47] Y. Peng, D. Meng, Z. Xu, C. Gao, Y. Yang, and B. Zhang, "Decomposable nonlocal tensor dictionary learning for multispectral image denoising," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2949–2956.

[48] J. Xue, Y. Zhao, W. Liao, and J. C. Chan, "Nonlocal low-rank regularized tensor decomposition for hyperspectral image denoising," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 5174–5189, Jul. 2019.

[49] Y. Chen, X. Cao, Q. Zhao, D. Meng, and Z. Xu, "Denoising hyperspectral image with non-i.I.d. noise structure," *IEEE Trans. Cybern.*, vol. 48, no. 3, pp. 1054–1066, Mar. 2018.

[50] 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.

[51] Y. Chang, L. Yan, and W. Liao, "HSI-DeNet: Hyperspectral image restoration via convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 667–682, Feb. 2018.

[52] J. Guan, R. Lai, H. Li, Y. Yang, and L. Gu, "DnRCNN: Deep recurrent convolutional neural network for HSI destriping," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 7, pp. 3255–3268, Jul. 2023.

[53] W. Dong, H. Wang, F. Wu, G. Shi, and X. Li, "Deep spatial–spectral representation learning for hyperspectral image denoising," *IEEE Trans. Comput. Imag.*, vol. 5, no. 4, pp. 635–648, Dec. 2019.

[54] L. Pang, W. Gu, and X. Cao, "TRQ3DNet: A 3D quasi-recurrent and transformer based network for hyperspectral image denoising," *Remote Sens.*, vol. 14, no. 18, p. 4598, Sep. 2022.

[55] Z. Liu et al., "Video Swin transformer," 2021, *arXiv:2106.13230*.

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

[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

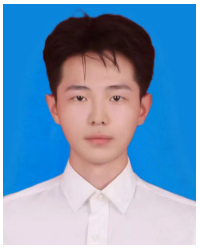
[58] J. Peng et al., "Fast noise removal in hyperspectral images via representative coefficient total variation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5546017.

Qiang Zhang (Member, IEEE) received the B.E. degree in surveying and mapping engineering and the M.E. and Ph.D. degrees in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2017, 2019, and 2022, respectively.

He is currently an Associate Professor with the Center of Hyperspectral Imaging in Remote Sensing (CHIRS), Information Science and Technology College, Dalian Maritime University, Dalian, China. He has authored more than ten journal articles in IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *Earth System Science Data*, and *ISPRS Journal of Photogrammetry and Remote Sensing*. His research interests include remote sensing information processing, computer vision, and machine learning. More details can be found at https://qzhang95.github.io.

Yushuai Dong received the B.S. degree in food science and engineering from Huazhong Agricultural University, Wuhan, China, in 2022. He is currently pursuing the M.S. degree with the School of Information Science and Technology College, Dalian Maritime University, Dalian, China.

His research interests include hyperspectral image processing and machine learning.

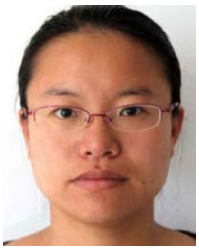


Yaming Zheng (Graduate Student Member, IEEE) received the B.S. degree from the School of Information Engineering (School of Software), Henan Animal Husbandry and Economics University, Zhengzhou, China, in 2022. He is currently pursuing the M.S. degree with the College of Information Science and Technology, Dalian Maritime University, Dalian, China.

His major research interests include remote sensing information processing.

Haoyang Yu (Member, IEEE) received the B.S. degree in information and computing science from Northeastern University, Shenyang, China, in 2013, and the Ph.D. degree in cartography and geographic information systems from the Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences (CAS), Beijing, China, in 2019.

He is currently an Associate Professor with the Center of Hyperspectral Imaging in Remote Sensing (CHIRS), Information Science and Technology College, Dalian Maritime University, Dalian, China. His research focuses on models and algorithms for hyperspectral image processing, analysis, and applications.



Meiping Song received the Ph.D. degree from the College of Computer Science and Technology, Harbin Engineering University, Harbin, China, in 2006.

She has been a Professor with the College of Information Science and Technology, Dalian Maritime University, Dalian, China, since 2020. Her research include remote sensing and hyperspectral image processing.

Lifu Zhang (Senior Member, IEEE) received the B.E. degree in photogrammetry and remote sensing from the Department of Airborne Photogrammetry and Remote Sensing, Wuhan Technical University of Surveying and Mapping, Wuhan, China, in 1992, the M.E. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan Technical University of Surveying and Mapping, in 2000, and the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, in 2005.

He is currently a full-time Professor and the Dean of the Hyperspectral Remote Sensing Division, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include hyperspectral remote sensing and imaging spectrometer system development and its applications.

Dr. Zhang is a member of the International Society for Optical Engineering (SPIE), the Academy of Space Science of China, and Chinese National Committee of the International Society for Digital Earth (CNISDE); the Vice-Chairperson of the Hyperspectral Earth Observation Committee, CNISDE; and a Standing Committeeman of the Expert Committee of China Association of Remote Sensing Applications.

Qiangqiang Yuan (Member, IEEE) received the B.S. degree in surveying and mapping engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2006 and 2012, respectively.

In 2012, he joined the School of Geodesy and Geomatics, Wuhan University, where he is a Professor. He published more than 90 research papers.

Prof. Yuan was a recipient of the Youth Talent Support Program of China in 2019. He was recognized as the Best Reviewer of IEEE GEOSCIENCE AND REMOTE SENSING LETTERS in 2019. He is an Associate Editor of five international journals and has served as a referee for more than 40 journals.