Deep Tensor Attention Prior Network for Hyperspectral Image Denoising

Weilin Shen, Junmin Liu, *Member, IEEE*, Jinhai Li, and Chao Tian

***Abstract*—Hyperspectral imaging techniques can generate continuous narrowband images with a high spectral resolution. However, owing to environmental disturbances, atmospheric effects, and hardware limitations of hyperspectral imaging sensors, captured hyperspectral images (HSIs) often contain complex noises that significantly degrades their quality and limits their utility. In this article, we design a deep tensor attention module based on a canonical-polyadic (CP) decomposition of a feature tensor, referred to as a deep CP attention module (DCPAM), which can disentangle spatial and channel information and further enhance the topological structure of features. It has been shown that the DCPAM is effective and relatively simple to integrate into some well-known network architectures. And taking the network with the DCPAM as a deep prior, we propose a deep tensor attention prior network (DTAPNet) for HSI denoising tasks. Extensive experiments on simulated and real HSIs demonstrate that our proposed DTAPNet outperforms existing state-of-the-art HSI denoising models.**

***Index Terms*—Attention module, deep prior, hyperspectral image (HSI) denoising, tensor decomposition.**

I. Introduction

Hyperspectral images (HSIs), which capture spectral features and spatial distributions of materials simultaneously [1], [2], are composed of hundreds of grayscale images at different wavelengths [3]. Therefore, HSIs have the potential for discriminative abilities in measuring, monitoring, and probing in a wide variety of applications [4], [5], and has been widely adopted in many remote sensing tasks such as demixing [6] and object classification [7]. Nonetheless, different types of noise are generated on the final captured HSI owing to the complex environments and acquisition errors of the sensors [8]. The coexistence of noise and interference leads to significant degradation of HSIs [9]. Therefore, before performing high-level tasks, denoising original hyperspectral images is important.

To recover clean HSIs from noisy images, extracting the topological or algebraic structures or some priors under the HSI to be reconstructed reasonably is crucial. Classical methods such as Fourier and wavelet transformations simply denoise noisy hyperspectral images band-by-band [10] without considering spectral correlations. This strategy often leads to spectral distortion, which cause poor denoising performance. With the rapid development of sparse and low-rank learning, more sophisticated approaches for HSI denoising with reasonable priors have been proposed, such as local smoothness constrained by total variation (TV) [11], [12], sparse representation (SR) [13], [14], and low-rank structure (LR) [15], [16].

Despite achieving satisfactory restoration performance, most existing methods [17] require converting the HSI cube into multiple 2-D matrices for denoising. This behavior tends to damage or destroy the inherent structure of the data itself [18]. Introducing tensor theory in HSI restoration tasks is an effective approach to address this issue. This method [19] can fully utilize the high contextual relevance of high-dimensional data structures to achieve hyperspectral denoising. Despite these significant improvements, notable problems remain in terms of issues like complex parameter tuning and computational inefficiency.

Recently, fitting the prior information of an HSI through a computational network model has become an effective method to achieve HSI denoising. For example, DnCNN [20] learns the nonlinear structure of data through residual connections. Sidorov and Hardeberg [21] proposed a new approach that uses the intrinsic properties of CNNs for HSI reconstruction without training. Additionally, Zhao et al. [22] and Kan et al. [23] introduced attention mechanisms to achieve HSI denoising. Although attention-based deep learning methods are helpful to learn the significant features, but most ignore the entangled spatial and spectral or channel information in HSIs or extracted features. Recently, Zhang et al. [24] found that low-rank priors in tensor structures can be explicitly characterised using image features. Based on this approach, we considered that tensor decomposition methods can be introduced on the feature maps of networks to disentangle spatial and channel information to obtain high-dimensional structural information.

In addition, training a deep neural network on an external dataset of noisy and clean image pairs is a powerful approach for image denoising. However, large high-quality training datasets are typically required, which limits the applicability of denoising methods. Therefore, the development of single-image denoising methods is of considerable interest, as they have no prerequisites of large training datasets. That is, denoising algorithms and

Manuscript received 12 April 2023; revised 26 May 2023 and 15 June 2023; accepted 2 July 2023. Date of publication 12 July 2023; date of current version 21 July 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2020AAA0105601 and in part by the National Natural Science Foundation of China under Grant 62276208, Grant U20B2075, and Grant 11991023. *(Corresponding author: Junmin Liu.)*

Weilin Shen and Junmin Liu are with the School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: swl970130@stu.xjtu.edu.cn; junminliu@mail.xjtu.edu.cn).

Jinhai Li is with the Data Science Research Center, Kunming University of Science and Technology, Kunming 650500, China, and also with the Faculty of Science, Kunming University of Science and Technology, Kunming 650500, China (e-mail: jhlixjtu@163.com).

Chao Tian is with the First Institute of Photogrammetry and Remote Sensing, MNR, Xi'an 710054, China (e-mail: gistc@163.com).

Digital Object Identifier 10.1109/JSTARS.2023.3294619

networks learn only from the input image themselves. Currently, very little research has been conducted on this topic. Ulyanov et al. [25] proposed a method for recovering an image using a deep neural network. Based on this, Sidorov and Hardeberg [21] developed a deep image prior method for HSI data. In this study, we also focus on a single HSI denoising task.

In contrast to previous works that directly used the entangled feature tensor as the input of subsequent stages, we propose a novel scheme to enhance significant features by combining canonical-polyadic (CP) tensor decomposition with attention mechanisms on feature tensors. We refer to this scheme as deep CP attention module (DCPAM). Considering the idea of RPCA, learning the low-rank information of an image itself has a certain denoising ability. In this article, for HSI denoising, we design such a module to learn clean HSIs from noisy images. We use tensor low-rank decomposition to learn low-rank information of clear images, and apply attention mechanism to provide more effective feature information, so as to learn a better prior structure for clean images. Specifically, DCPAM first disentangles the feature tensor into three components by CP decomposition and utilizes these components to generate spatial and channel attention maps, respectively. Subsequently, it refines the features by aggregating the two attention maps into a 3-D attention tensor. It has been shown that DCPAM can be easily integrated into well-known network architectures. On the one hand, compared with traditional sparse and low-rank penalties, DCPAM utilizes the powerful computing power and adaptability of the network to obtain more accurate prior information. On the other hand, compared with the prior methods using networks in the past, DCPAM considers the high-dimensional tensor structure on the feature domain of hyperspectral data, which is conducive to more fully extracting the intrinsic correlation structure of the data to obtain better denoising quality. By taking the networks with DCPAM as a deep prior to regularize the HSI denoising problem, a deep tensor attention prior network (DTAPNet) is proposed. Compared with the method of denoising directly through the network or a certain set of model calculation, our method combines the prior penalty term of the data itself and the loss term related to the denoising problem, which can be more beneficial from recovering clear images from noisy images. We experimentally compared our approach with the state-of-the-art (SOTA) methods to demonstrate that it exhibited superior performance on both simulated and real-world data.

The key contributions of this study are summarized as follows.

1) We design a deep tensor attention module (DCPAM) based on a spatial and channel information disentangled by the CP tensor decomposition of features.
2) We show that the proposed DCPAM can be used as a play-and-plug block to easily integrate into three popular network architectures.
3) By taking a network with DCPAM as deep prior, we construct the DTAPNet for the HSI denoising task.
4) We present the results of extensive experiments conducted to verify the effectiveness of DTAPNet on both simulated and real-world data.

TABLE I
NOTATIONS

Definition	Notation
Scalar	a
Vector	$\boldsymbol{a}$
Matrix	$\boldsymbol{A}$
Tensor	$\boldsymbol{\mathcal{A}}$
Element of 3D tensor	$\boldsymbol{\mathcal{A}}_{ijk}$
Inner product (3D tensor)	$\langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \rangle = \sum_{ijk} \boldsymbol{\mathcal{A}}_{ijk} \cdot \boldsymbol{\mathcal{B}}_{ijk}$
Frobenius norm (3D tensor)	$\parallel \boldsymbol{\mathcal{A}} \parallel_F = \sqrt{\sum_{ijk} \lvert \boldsymbol{\mathcal{A}}_{ijk} \rvert^2}$

The rest of this article is organized as follows. Section II provides some notation along with relevant preliminaries. Section III describes the structure of the deep CP attention module. Section IV introduces the denoising method and its network structure. Section V presents experimental results and comparisons with several benchmarks. Finally, Section VI concludes this article.

II. Related Works

In this section, we introduce some basic concepts and related work. Here, we first provide some notations in Table I.

A. HSI Denoising Prior

HSI denoising tasks have been extensively studied in recent years. These methods considered different prior structures of HSIs.

In addition to classical 2-D image denoising methods such as BM3D [26] and NLM [27], several extended approaches have been developed for HSI denoising. They considered small local patches of HSI, building HSI's as small 3-D cubes instead of 2-D patches of traditional images, such as BM4D [28] and NLM3D [29], which usually consider smoothing the prior structures of hyperspectral data.

Another widely used method is to reconstruct the desired signal by solving a prior regularized convex optimization problem. Constraining the sparse prior structure of the data is a widely discussed approach. Tan et al. [30] introduced a method using approximate message passing to maintain the sparse properties of HSI. Peng et al. [31] combined two properties of MSI: nonlocal similarity in space and global correlation of spectra, and proposed a new tensor dictionary learning model (TDL) for MSI denoising.

However, instead of sparsity, low-dimensional structure suggested by spatial correlations among pixels and spectral correlations among bands is also the ubiquitous prior information in HSIs. More and more works have emerged that jointly consider the sparsity and low rank of hyperspectral data. He et al. [32] proposed local low-rank matrix recovery and global spatial–spectral total variation (LLRSSTV), which adopts matrix SVD to obtain low-rank information. In order to fully utilize the prior spectral information of hyperspectral data, another point is to regard HSIs as 3-D tensors and to apply tensor factorization techniques. For example, Renard et al. [33] obtained a low-rank approximation of input data using their proposed low-rank tensor approximation (LRTA) method. Regarding parallel factor analysis, Liu et al. [34] proposed the PARAFAC method to eliminate the spectral redundancy of HSI. Depending on the optimal singular value thresholding of low-rank tensor unfoldings, Zhuang et al. [35] presented a global and nonlocal low-rank factorizations (GLF) method, which can represent hyperspectral images in low-dimensional subspaces. He et al. [36] developed the nonlocal meets global (NGmeet) algorithm, which is an efficient rank adaptive alternating minimization method. Wu et al. [37] considered the spatial and spectral information of individual regions to constrain the low-rank and sparse properties of HSIs.

Despite promising results, laborious parameter tuning and computational constraints prevent most methods from being used in practice. Recently, deep learning has shown considerable potential in denoising tasks [20], [38]. Zhang et al. [20] proposed a DnCNN to obtain the prior structure by adding batch normalization [39] and residual connections [40]. Yuan et al. [41] trained a model with remote sensing images prior to using a deep residual network, which adopted a sliding window strategy to deal with HSIs. He et al. [42] decomposed 3-D convolutions into lightweight 2-D spatial and spectral convolutions to exploit the spatial spectral information of HSI. Xiong et al. [43] proposed an end-to-end network SMDS-Net to learn the low rank and sparsity of HSI. Basing on deep image prior [25], Sidorov and Hardeberg [21] developed a single-hyperspectral-image recovery method (deep HS prior). Bodrito et al. [44] advocated a deep prior based method with sparse coding principles. In addition to CNN-based methods, attention-based methods have recently been developed. Kan et al. [23] presented an AODN for HSI denoising that extracted spatial–spectral prior features using an attention module.

However, these deep learning methods are matrix-based and do not fully consider the high-dimensional correlation structure of HSIs or consider the high-dimensional structure of hyperspectral data one-sidedly, such as taking the spectral low rank of hyperspectral data into account separately, which limited the improvement of denoising quality. Since then, in this work, we represent the tensor low-rank prior of HSI by designing a tensor-based low-rank attention structure in the feature domain to better constrain the high-dimensional prior information of the data and achieve better denoising quality.

B. Attention Mechanisms

Traditional convolutional neural networks can learn the local features of the data. To learn the global information of the data, we consider introducing a self-attention mechanism, which is a global encoding approach that can incorporate long-distance features to associate the weights and calculate the weighted sum of inputs [45]. Many computer vision applications are based on variants of self-attention. For better image classification, Hu et al. [46] proposed SE-Net to utilize channel information. Woo et al. [47] presented CBAM by combining channel and spatial attention to obtain more comprehensive feature information for better performance. Wang et al. [48] introduced a nonlocal neural network, which was proposed as plug and play extension to existing architectures. Attention mechanisms can provide knowledge for more effective feature information, and have made great progress in recent years. In our article, we combine the attention structure with tensor-based low-rank representation for better prior information.

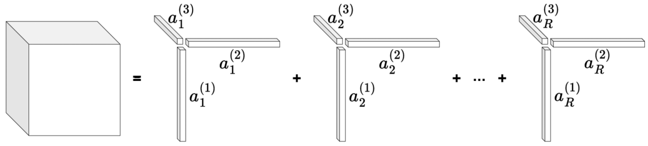


Fig. 1. CP decomposition of a 3-D tensor.

C. Tensor Low-Rank Representation

Matrix methods tend to ignore the high-dimensional inner structure of data. A useful approach is to treat the data as multidimensional arrays known as tensors [49]. For example, colour images are 3-D tensors, and functional magnetic resonance images are 4-D tensors. For tensors, finding low-rank structures is an important task and such methods have been extensively developed in recent years [49], [50]. Tucker decomposition, tensor singular value decomposition (tSVD) and CP decomposition are the most widely used decomposition methods. Tucker decomposition [51] represents a tensor as the product of a core tensor and projection matrices. However, it is hard to adjust and ensure the ranks in the experiments. TSVD [52] is an elegant and natural extension of matrix SVD. Nevertheless, components in the form of high-order tensors are difficult to design related attention network modules. CP decomposition [53] expresses an arbitrary tensor as the sum of rank-1 tensors. CP decomposition is widely used in tasks with high-dimensional data such as convolution speed-up and model compression [54]. In this study, we applied CP decomposition to model a tensor low-rank attention module to fully utilize the high-dimensional prior information of data in the feature domain.

III. Deep CP Attention Module and Networks

In this section, based on CP decomposition, we design a deep CP attention module.

A. DCPAM

The following theorem defines CP decomposition.

Theorem 1: Any tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ can be decomposed into the sum of some rank-1 tensors as follows:

$$\begin{aligned} \boldsymbol{\mathcal{X}} &\approx \left[\!\left[\boldsymbol{A}^{(1)}, \boldsymbol{A}^{(2)}, \ldots, \boldsymbol{A}^{(N)} \right]\!\right] \\ &\equiv \sum_{r=1}^{R} \boldsymbol{a}_r^{(1)} \circ \boldsymbol{a}_r^{(2)} \circ \cdots \circ \boldsymbol{a}_r^{(N)} \end{aligned} \quad (1)$$

where $\boldsymbol{A}^{(n)} = \left[\boldsymbol{a}_1^{(n)}, \boldsymbol{a}_2^{(n)}, \ldots, \boldsymbol{a}_R^{(n)} \right] \in \mathbb{R}^{I_n \times R}, n = 1, 2, \ldots,$ N, is a column normalization matrix, R represents the CP rank which is typically predefined in experiments, and $\circ$ denotes the Kronecker product. Fig. 1 shows the visual display for 3-D tensors.

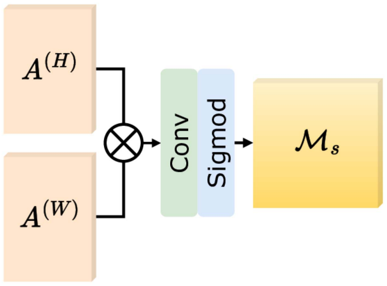


Fig. 2. Architecture of the spatial attention map based on CP decomposition (CPSAM).

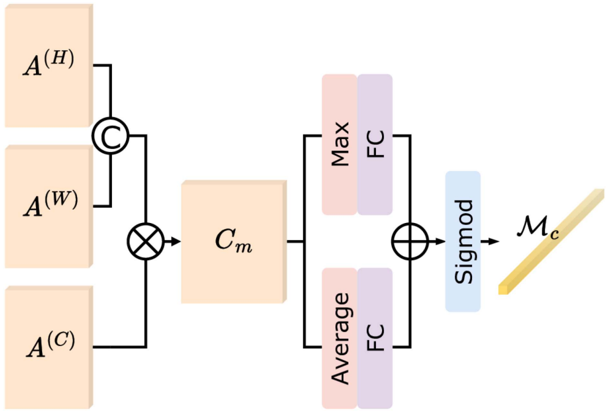


Fig. 3. Architecture of the channel attention map based on CP decomposition (CPCAM), where “C” expresses “Concat.”

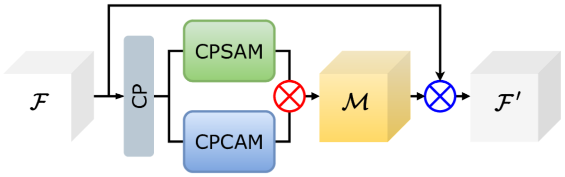


Fig. 4. Architecture of the deep CP attention module (DCPAM). The first red product label denotes Kronecker product and the blue product label denotes elementwise multiplication.

We applied this theory to extract the principal components of feature tensors to obtain better descriptive attention weights. According to the CP decomposition, the matrix $\boldsymbol{A}^{(n)}$ reveals the low-rank information of nth dimension. Hence, CP decomposition can effectively characterize information under different dimensions on the feature maps. In our work, we calculate the components of the CP decomposition by the alternating least squares method [49] with a given rank R.

Given an intermediate feature tensor $\boldsymbol{\mathcal{F}} \in \mathbb{R}^{C\times H\times W}$, considering CP decomposition, we have that $\boldsymbol{\mathcal{F}} = \boldsymbol{A}^{(C)} \circ \boldsymbol{A}^{(H)} \circ \boldsymbol{A}^{(W)}$, where $\boldsymbol{A}^{(C)} \in \mathbb{R}^{C\times R}$, $\boldsymbol{A}^{(H)} \in \mathbb{R}^{H\times R}$, and $\boldsymbol{A}^{(W)} \in \mathbb{R}^{W\times R}$. We utilize $\boldsymbol{A}^{(H)}$ and $\boldsymbol{A}^{(W)}$ to obtain the spatial attention map $\boldsymbol{\mathcal{M}}_s$, and make full use of $\boldsymbol{A}^{(C)}$, $\boldsymbol{A}^{(H)}$, and $\boldsymbol{A}^{(W)}$ to generate the channel attention map $\boldsymbol{\mathcal{M}}_c$. We then combine the two parts as the entire attention tensor $\boldsymbol{\mathcal{M}}$. The details are as follows.

1) Spatial Attention: Storing the spatial information of feature maps is of great significance. In our study, the spatial attention map is generated by the spatial-dependent components of the feature tensor, as shown in Fig. 2. According to CP decomposition, $\boldsymbol{A}^{(H)}$ and $\boldsymbol{A}^{(W)}$ contain low-rank information on the dimensions of height and weight. The product of $\boldsymbol{A}^{(H)}$ and $\boldsymbol{A}^{(W)}$ simply implies low-rank information of the spatial dimension. We then apply convolutional layers to generate spatial attention maps $\boldsymbol{\mathcal{M}}_s \in \mathbb{R}^{1\times H\times W}$. Thus, we can compute the spatial attention by

$$\boldsymbol{\mathcal{M}}_s\left(\boldsymbol{A}^{(H)}, \boldsymbol{A}^{(W)}\right) = \sigma\left(f_{conv}\left(\boldsymbol{A}^{(H)} \cdot \left(\boldsymbol{A}^{(W)}\right)^T\right)\right) \quad (2)$$

where σ and f_{conv} represent the sigmoid function and the convolution operation, respectively.

2) Channel Attention: Considering the relationship between the corresponding channel component $\boldsymbol{A}^{(C)}$ and the corresponding spatial components $\boldsymbol{A}^{(H)}$ and $\boldsymbol{A}^{(W)}$, we generate the following channel attention map. The operations are described in detail below. We consider $\boldsymbol{A}^{(H)}$ and $\boldsymbol{A}^{(W)}$ as query matrices and consider $\boldsymbol{A}^{(C)}$ as the key matrix. The entire process involves multihead attention. We obtain the correlation matrix $\boldsymbol{C}_m = \boldsymbol{A}^{(C)} \cdot \text{Concat}\left(\boldsymbol{A}^{(H)}, \boldsymbol{A}^{(W)}\right)^T$. The dimension related to the spatial information in the correlation matrix is squeezed to efficiently compute channel attention. Here, we gather aggregated spatial information by averaging and maximising the corresponding row vectors of each channel in the correlation matrix. To output our channel attention map $\boldsymbol{\mathcal{M}}_c \in \mathbb{R}^{C\times 1\times 1}$, a multilayer perceptron was provided. Fig. 3 shows the architecture in detail. We compute the channel attention by

$$\begin{aligned}\boldsymbol{\mathcal{M}}_c&\left(\boldsymbol{A}^{(C)}, \boldsymbol{A}^{(H)}, \boldsymbol{A}^{(W)}\right) = \boldsymbol{\mathcal{M}}_c(\boldsymbol{C}_m)\\ &= \sigma(f_{MLP}(\text{avg}(\boldsymbol{C}_m)) + f_{MLP}(\max(\boldsymbol{C}_m))) \quad (3)\end{aligned}$$

where f_{MLP} denotes a multilayer perceptron with one hidden layer.

3) Arrangement of Attention Module: For the given feature map $\boldsymbol{\mathcal{F}} = \boldsymbol{A}^{(C)} \circ \boldsymbol{A}^{(H)} \circ \boldsymbol{A}^{(W)}$, we produce the whole attention tensor $\boldsymbol{\mathcal{M}} = \boldsymbol{\mathcal{M}}_c \circ \boldsymbol{\mathcal{M}}_s \in \mathbb{R}^{C\times H\times W}$ by spatial attention map $\boldsymbol{\mathcal{M}}_s$ and channel attention map $\boldsymbol{\mathcal{M}}_c$. Then the output feature tensor can be calculated by $\boldsymbol{\mathcal{F}}' = \boldsymbol{\mathcal{F}} \otimes \boldsymbol{\mathcal{M}}$, where $\otimes$ represents element-wise multiplication. The entire tensor attention module is shown in Fig. 4.

B. Deep Networks With DCPAM

In this section, we integrate DCPAM into several deep networks and construct a deep CP attention module based network (simply denoted as DCPAMNet) to learn the inner structure of the feature maps.

We consider three network architectures, as shown in Fig. 5. The first architecture contained a simple ResNet bottleneck block and added a deep CP attention module. The second architecture is a skip-connection-based network with two downsampling and upsampling layers. In fact, we add 1 × 1 convolution to the skip connection layers. In addition to the convolutional layers in the skip connection layers and the last 1 × 1 convolutional layer for maintaining the dimensions of the output

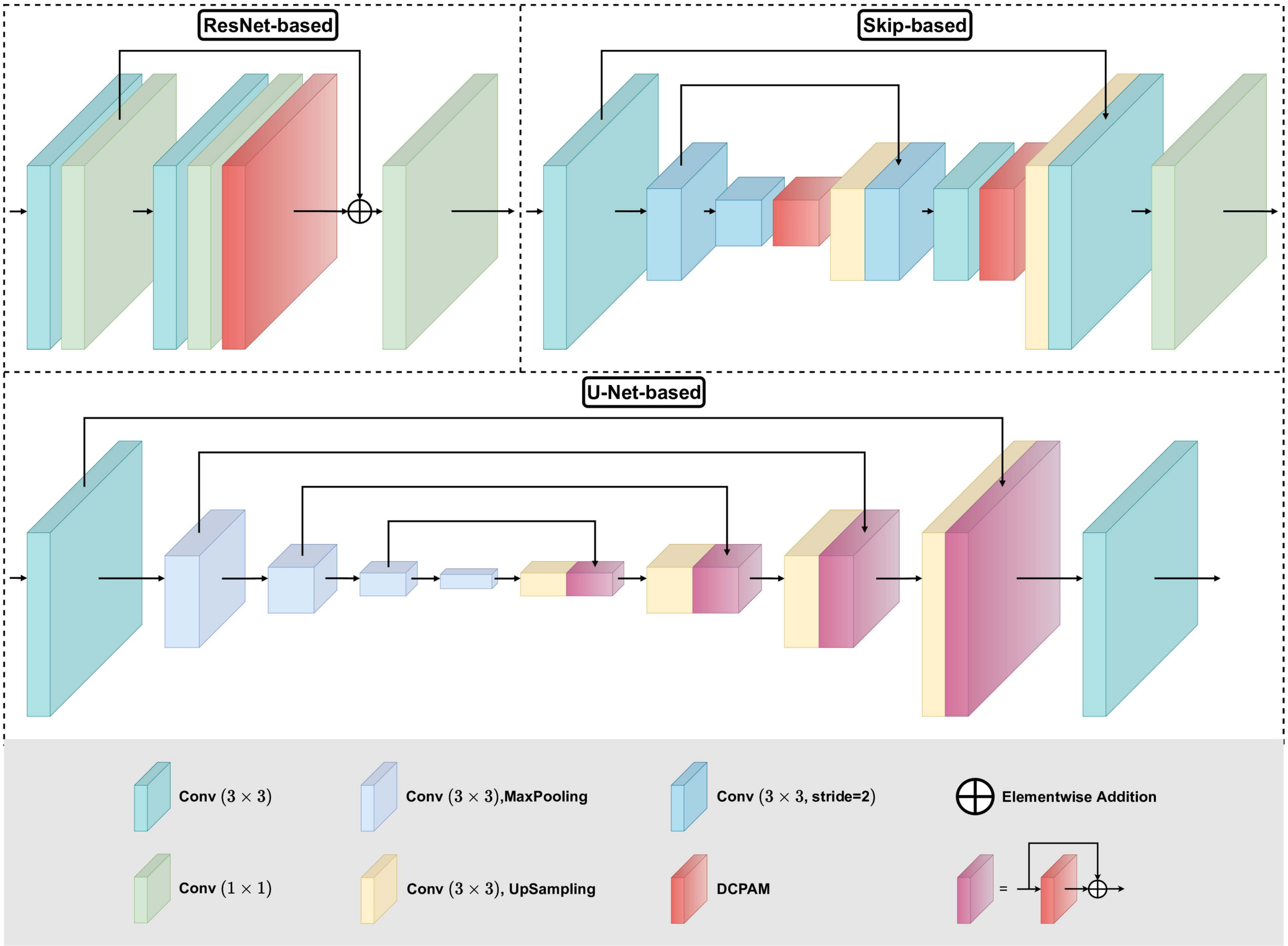


Fig. 5. Architecture of the deep network with DCPAM (DCPAMNet) based on ResNet, skip connection, and U-net model.

feature tensors, the values of the output channel in all other convolutional layers are the same and artificially set. The last network architecture was built based on U-Net. The convolutional block includes two 3 × 3 convolutional layers, as shown in the lower right corner. We compared their performances in ablation experiments.

Of note, the variables usually have different inner structures. In this section, we show that our proposed DCPAM can easily be integrated into different popular networks to obtain special information on variables, which demonstrates that our proposed module has higher applicability.

IV. HSI Denoising Model and Algorithm

In this section, we apply the well-known alternating direction method of multipliers (ADMM) algorithm [55] to the traditional HSI denoising optimization problem with a prior penalty and unfold our proposed optimization process into a deep network to solve the hyperspectral image-denoising problem.

A. Model Optimization

Before presenting our final network model, we consider the general optimization problem of HSI denoising.

Let $\boldsymbol{\mathcal{H}} \in \mathbb{R}^{H\times W\times S}$ be the observed HSI, where H and W represents the spatial dimensions, and S is the spectral dimension. We describe the HSI noise degradation model using the following formula:

$$\boldsymbol{\mathcal{H}} = \boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{N}} \tag{4}$$

where $\boldsymbol{\mathcal{X}}$ denotes the ideal clean HSI, and $\boldsymbol{\mathcal{N}}$ represents the noise tensor. The HSI denoising task aims to reconstruct $\boldsymbol{\mathcal{X}}$ from the noisy observation $\boldsymbol{\mathcal{H}}$.

A classical approach is to apply penalized least-squares optimization to the initial model (4), which leads to the following problem:

$$\min_{\boldsymbol{\mathcal{X}}} \frac{1}{2} \parallel \boldsymbol{\mathcal{H}} - \boldsymbol{\mathcal{X}} \parallel_F^2 + \lambda\phi(\boldsymbol{\mathcal{X}}) \tag{5}$$

where λ implies a tradeoff between fidelity term $\frac{1}{2} \parallel \boldsymbol{\mathcal{H}} - \boldsymbol{\mathcal{X}} \parallel_F^2$ and penalty term $\phi(\boldsymbol{\mathcal{X}})$.

To optimize the reconstruction problem (5), we introduce a well-known ADMM algorithm. By adding auxiliary variable $\boldsymbol{\mathcal{S}} = \boldsymbol{\mathcal{X}}$, (5) is rewritten as the constrained optimization problem

$$\min_{\boldsymbol{\mathcal{X}}} \frac{1}{2} \parallel \boldsymbol{\mathcal{H}} - \boldsymbol{\mathcal{X}} \parallel_F^2 + \lambda\phi(\boldsymbol{\mathcal{S}}), \quad \text{s.t.}\ \boldsymbol{\mathcal{S}} = \boldsymbol{\mathcal{X}}. \tag{6}$$

Algorithm 1: Iterative Optimization Algorithm.

```
for $t = 1 : K$ do
    $\boldsymbol{\mathcal{X}}^{t+1} = \frac{1}{\mu+1}(\boldsymbol{\mathcal{H}} + \mu\boldsymbol{\mathcal{S}}^t + \boldsymbol{\mathcal{Y}}^t)$
    $\boldsymbol{\mathcal{S}}^{t+1} = \mathrm{prox}_{\lambda,\mu}\left(\boldsymbol{\mathcal{X}}^{t+1} - \frac{\boldsymbol{\mathcal{Y}}^t}{\mu}\right)$
    $\boldsymbol{\mathcal{Y}}^{t+1} = \boldsymbol{\mathcal{Y}}^t + \mu(\boldsymbol{\mathcal{S}}^{t+1} - \boldsymbol{\mathcal{X}}^{t+1})$
end for
```

The augmented Lagrange function for such optimization problem (6) is given as

$$L_\mu(\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{S}},\boldsymbol{\mathcal{Y}}) = \frac{1}{2} \parallel \boldsymbol{\mathcal{H}} - \boldsymbol{\mathcal{X}} \parallel_F^2 + \lambda\phi(\boldsymbol{\mathcal{S}}) + \langle \boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}} \rangle + \frac{\mu}{2} \parallel \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}} \parallel_F^2 \tag{7}$$

where μ is a penalty factor and $\boldsymbol{\mathcal{Y}}$ is the Lagrange multiplier. Next, we can obtain $\boldsymbol{\mathcal{S}}$, $\boldsymbol{\mathcal{X}}$, and $\boldsymbol{\mathcal{Y}}$ by following:

$$\boldsymbol{\mathcal{S}}^{t+1} = \arg\min_{\boldsymbol{\mathcal{S}}} \lambda\phi(\boldsymbol{\mathcal{S}}) + \frac{\mu}{2} \parallel \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}^t + \frac{\boldsymbol{\mathcal{Y}}^t}{\mu} \parallel_F^2 \tag{8}$$

$$\boldsymbol{\mathcal{X}}^{t+1} = \arg\min_{\boldsymbol{\mathcal{X}}} \frac{1}{2} \parallel \boldsymbol{\mathcal{H}} - \boldsymbol{\mathcal{X}} \parallel_F^2 + \frac{\mu}{2} \parallel \boldsymbol{\mathcal{S}}^{t+1} - \boldsymbol{\mathcal{X}} + \frac{\boldsymbol{\mathcal{Y}}^t}{\mu} \parallel_F^2 \tag{9}$$

$$\boldsymbol{\mathcal{Y}}^{t+1} = \boldsymbol{\mathcal{Y}}^t + \mu\left(\boldsymbol{\mathcal{S}}^{t+1} - \boldsymbol{\mathcal{X}}^{t+1}\right). \tag{10}$$

To address the optimization subquestion (8), we calculate the following:

$$\boldsymbol{\mathcal{S}}^{t+1} = \mathrm{prox}_{\lambda,\mu}\left(\boldsymbol{\mathcal{X}}^t - \frac{\boldsymbol{\mathcal{Y}}^t}{\mu}\right) \tag{11}$$

where $\mathrm{prox}_{\lambda,\mu}(\cdot)$ is a proximal operator with parameters λ and μ.

Obviously, (9) of updating variable $\boldsymbol{\mathcal{X}}^{t+1}$ represents a quadratic optimization problem with an explicit solution, that is

$$\boldsymbol{\mathcal{X}}^{t+1} = \frac{1}{\mu+1}\left(\boldsymbol{\mathcal{H}} + \mu\boldsymbol{\mathcal{S}}^{t+1} + \boldsymbol{\mathcal{Y}}^t\right). \tag{12}$$

The optimization algorithm guides us to update these three steps alternately, i.e., (10)–(12), until a convergence condition is reached, as briefly shown in Algorithm 1.

B. DTAPNet for HSI Denoising

In this section, we propose our deep tensor attention prior network (DTAPNet) by unfolding the optimization algorithm given before and give out some learning strategies in this section.

1) Algorithm Unfolding: We consider the algorithm unrolling method [56] and integrate our proposed three updating steps, i.e., (10)–(12), into an end-to-end network. The implicit proximal operator (11) can be easily expressed as a convolutional network module and automatically trained in an end-to-end manner, rather than derived from manually prespecified regularization terms. Here, we adopted our proposed deep CP attention module-based network (DCPAMNet) to formulate this operator. We note that the parameter λ is implied in the network to briefly write the proximal operator as

$$\boldsymbol{\mathcal{S}}^{t+1} = \mathrm{DCPAMNet}_\mu\left(\boldsymbol{\mathcal{X}}^t - \frac{\boldsymbol{\mathcal{Y}}^t}{\mu}\right). \tag{13}$$

Algorithm 2: Network Design.

```
for stage $t = 1 : K$ of the network do
    $\boldsymbol{\mathcal{X}}^{t+1} = \frac{1}{\mu+1}\left(\boldsymbol{\mathcal{H}} + \mu\boldsymbol{\mathcal{S}}^t + \boldsymbol{\mathcal{Y}}^t\right)$
    $\boldsymbol{\mathcal{S}}^{t+1} = \mathrm{DCPAMNet}_\mu\left(\boldsymbol{\mathcal{X}}^{t+1} - \frac{\boldsymbol{\mathcal{Y}}^t}{\mu}\right)$
    $\boldsymbol{\mathcal{Y}}^{t+1} = \boldsymbol{\mathcal{Y}}^t + \mu\left(\boldsymbol{\mathcal{S}}^{t+1} - \boldsymbol{\mathcal{X}}^{t+1}\right)$
end for
```

Since then, the unrolled network is to integrate (10), (12), and (13), as summarized below in Algorithm 2.

Compared to Algorithm 1, the marked red part exhibits the most important changes. Specifically, we initialized $\boldsymbol{\mathcal{X}}^0 = \boldsymbol{\mathcal{H}}$ and $\boldsymbol{\mathcal{Y}}^0 = \mathbf{0}$. Then, we alternately updated the above three equations K times. Finally, our proposed network outputs denoised HSI $\boldsymbol{\mathcal{X}}^K$. Fig. 6 shows the whole deep tensor attention prior network (DTAPNet) structure based on ADMM. The dotted box represents the update process of the tth stage corresponding to the unfolding algorithm.

2) Learning Strategy: In the optimization algorithm, the hyperparameter μ has a significant impact on accuracy. Nevertheless, in our actual operations, it is of great challenge and difficulty to tune hyperparameters for different conditions. Here, we set the hyper-parameter μ as a learnable variable to obtain adaptive parameters along the end-to-end network.

In our work, the MSE loss is introduced as the evaluation criterion, which leads to the following expression of the loss function:

$$\mathcal{L}(\mu, \boldsymbol{\Theta}_{net}) = \mathbb{E} \parallel F(\boldsymbol{\mathcal{H}}, \mu, \boldsymbol{\Theta}_{net}) - \boldsymbol{\mathcal{H}} \parallel_F^2 \tag{14}$$

where μ is a hyperparameter which can be learned automatically in the net, $F(\cdot)$ denotes the whole network and $\boldsymbol{\Theta}_{net}$ denotes the set of all network parameters. We minimize the loss function with Adam and set the learning rate initially as 0.01. In the following experiments of our work, we set $K = 10$.

V. Experiments

In this part, we describe the results of some simulation experiments performed with two datasets including OHS-3D HSI and Pavia University, and two real data experiments on the OHS-3D HSI and Indian Pines dataset to demonstrate the performance of our proposed DTAPNet model and compare it with several state-of-the-art (SOTA) methods.

A. Datasets

Training dataset: In order to train our model, we selected hyperspectral remote sensing images taken by the “Zhuhai-1” OHS-3D hyperspectral satellite in 2020, which centered at $45.9312^\circ N$, $84.537874^\circ E$ (in Xinjiang Uygur Autonomous Region, China). The sensor records 32 spectral bands in the wavelength range from 439 to 944 nm. The size of

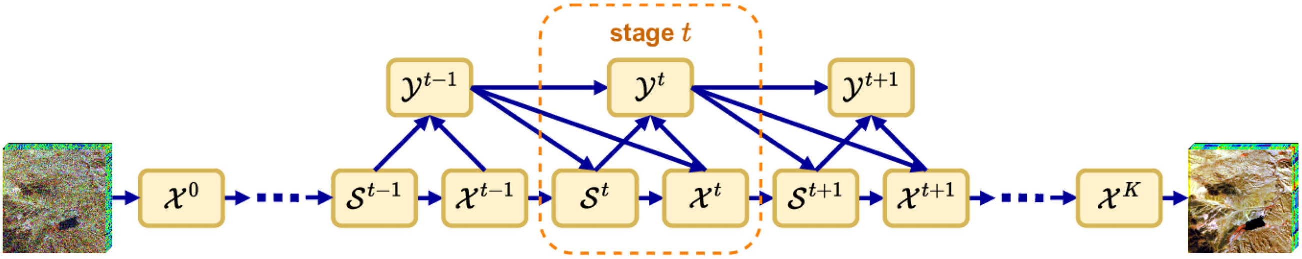


Fig. 6. Architecture of the deep tensor attention prior network (DTAPNet).

the entire hyperspectral image is 5057 × 5056 × 32. In this article, the dataset is simply denoted as OHS-3D dataset. We splited the images into two parts: one for training the proposed network and the other for testing. For the testing part, we crop a region of size 500 × 500 × 32 from the full HSI, and the rest are used for training.

Testing dataset: In this work, we conducted simulation experiments on the following three datasets to evaluate the effectiveness of the proposed method in real scenarios.

1) *OHS-3D HSI Dataset:* For experiments, a patch of 500 × 500 pixels was cropped from the images.
2) *Pavia University Dataset:* The scene was captured by the reflective optical spectroscopic imaging system (ROSIS). Here, we employed a cropped part of 300 × 300 pixels size with 103 channels for experiments.
3) *OHS-3 C HSI Dataset:* This dataset (simply denoted as OHS-3 C HSI dataset), which was originally acquired by the "Zhuhai-1" OHS-3 C hyperspectral satellite in 2020 and centered at $43.939689^{\circ}N$, $87.56131^{\circ}E$ (in Xinjiang Uygur Autonomous Region, China), reserved 32 bands for analysis. A cropped part of 500 × 500 pixels size was employed for experiments in this work.

B. Experimental Setup

In this work, our proposed DTAPNet is compared with several SOTA HSI denoising method, including three sparse methods: BM3D [26], BM4D [28], TDL [31]; three low-rank methods: LRTA [33], GLF [35], NGmeet [36]; and five deep learning-based methods HSID-CNN [41], deep HS prior [21], SMDS-Net [43], FastHyMix [57], T3SC [44]. The gray values of each HSI band are normalized before denoising.

To evaluate the effectiveness of our proposed method and obtain quantitative comparisons with these current mainstream methods, we adopted four commonly employed metrics, including peak signal-to-noise-ratio (PSNR), structural similarity index measure (SSIM) [58], spectral angle mapper (SAM) [59] and erreur relative globale adimensionnelle de synthèse (ERGAS) [60]. PSNR is usually used to evaluate the visual quality and SSIM evaluates structural similarity based on the highly structured features of images. SAM describes the spectral fidelity and ERGAS computes a weighted sum of MSE in each band to obtain the fidelity of restored images. Generally speaking, higher PSNR and SSIM, and lower SAM, and ERGAS represent better performance.

Next, we generated several different types of noises to simulate real-world noise of HSIs. Specifically, the additional noise was generated in the three cases as follows.

1) Each band was added by i.i.d Gaussian noise with same noise level $\sigma = 25, 50, 75$, or 100.
2) Each band was added by a random probability distribution noise. For example, we select $\sigma = rand(25)$ and $\sigma = rand(50)$.
3) Add salt and pepper noise with a noise density of 0.1.

C. Denoising Performance

We provide the denoising results of our experiments in this section. Details are shown in Tables II–IV. For a detailed comparison, we chose several cases from each dataset to show the visual performance, i.e. Figs. 7, 9, and 11. Although HSIs usually have large number of bands, only three bands were selected in each case to illustrate visual performance with pseudocolour. For presentation, we choose bands 26, 16 and 6 for OHS-3D HSI dataset, bands 57, 27, and 17 for Pavia University dataset, and bands 16, 10, and 3 for OHS-3 C HSI dataset. Fig. 7 reveals the denoising results on OHS-3D HSI dataset with noise level of $\sigma = 100$. Fig. 9 shows the denoising performance of different methods on Pavia University dataset with random noise of $\sigma = rand(25)$. Fig. 11 exhibits the corresponding denoising effect with salt and pepper noise on OHS-3 C HSI dataset. Fig. 8 illustrates the spectral signature of pixel (75,475) on OHS-3D HSI denoising task mentioned before. The spectral signature of pixel (175,145) on Pavia University dataset with random noise of $\sigma = rand(25)$ is revealed in Fig. 10. Similarly, Fig. 12 shows the spectral signature of pixel (285,235) on OHS-3 C HSI dataset with salt and pepper noise. The per-band denoising results, which are shown in Figs. 13 and 14, describe the values of PSNR and SSIM of recovered HSI at different bands.

To observe the performance more clearly, the top two results for each quality metric are marked red and blue in Tables II–IV. Clearly, our proposed DTAPNet achieved the highest SSIM and the lowest ERGAS among all noise levels, and performed better on PSNR and SAM in most cases among all comparing methods. For example, our method achieved a PSNR value of 30.58, an SSIM value of 0.9615, an SAM value of 0.0530 and an ERGAS value of 1.1077 on the OHS-3D HSI denoising task with noise level of $\sigma = 100$, which represented the best on all metrics.

Figs. 7, 9, and 11 show the full images and the local magnification maps of our proposed model and the comparison

TABLE II
QUANTITATIVE EVALUATION OF DENOISING RESULTS OF OHS-3D HSI DATASET

			Sparse methods			Low-rank methods			DL methods					
Noise	Index	Noisy	BM3D	BM4D	TDL	LRTA	GLF	NGmeet	HSID-CNN	deepHSprior	SMDS-Net	FastHyMix	T3SC	Proposed
$\sigma = 25$	PSNR	20.21	31.56	32.45	32.29	30.96	31.98	33.25	31.00	32.86	33.20	33.09	33.18	33.90
	SSIM	0.7351	0.9725	0.9775	0.9767	0.9689	0.9693	0.9814	0.9668	0.9799	0.9774	0.9772	0.9780	0.9830
	SAM	0.2027	0.0478	0.0459	0.0462	0.0530	0.0522	0.0418	0.0558	0.0415	0.0454	0.0461	0.0451	0.0403
	ERGAS	3.4676	0.9339	0.8458	0.8626	1.0049	0.9794	0.7698	1.03001	0.8050	0.8407	0.8482	0.8124	0.7528
$\sigma = 50$	PSNR	14.40	28.88	29.27	30.42	28.11	30.47	30.54	27.81	30.53	30.99	31.21	31.05	31.78
	SSIM	0.4188	0.9470	0.9514	0.9629	0.9381	0.9560	0.9633	0.9343	0.9652	0.9644	0.9648	0.9649	0.9712
	SAM	0.3792	0.0619	0.0662	0.0540	0.0669	0.5938	0.0531	0.0811	0.0509	0.0530	0.0546	0.0521	0.0491
	ERGAS	6.7431	1.2785	1.2179	1.0717	1.3960	1.1548	1.0626	1.4569	1.0539	1.0533	1.0442	1.0510	0.9634
$\sigma = 75$	PSNR	11.54	26.71	27.08	28.54	26.38	28.70	28.89	23.46	28.64	29.23	29.27	30.06	31.28
	SSIM	0.2574	0.9046	0.9156	0.9401	0.9009	0.9317	0.9331	0.8313	0.9438	0.9439	0.9435	0.9468	0.9674
	SAM	0.5029	0.0886	0.0722	0.0643	0.0750	0.0687	0.0659	0.1320	0.0617	0.0648	0.0649	0.0592	0.0505
	ERGAS	9.3043	1.6254	1.5712	1.3285	1.7009	1.4013	1.3823	2.3731	1.3003	1.2957	1.2957	1.2430	1.0197
$\sigma = 100$	PSNR	9.98	24.44	25.55	26.63	24.82	26.79	26.30	22.30	26.72	27.17	28.28	29.19	30.58
	SSIM	0.1776	0.8236	0.8730	0.9023	0.8449	0.8906	0.8711	0.7702	0.9077	0.9093	0.9266	0.9417	0.9615
	SAM	0.5800	0.1145	0.0823	0.0771	0.0832	0.0801	0.0817	0.1464	0.0739	0.0776	0.0726	0.0654	0.0530
	ERGAS	11.0596	2.1036	1.8719	1.6565	2.0337	1.7274	1.8377	2.7022	1.6291	1.6192	1.4746	1.3130	1.1077
$\sigma = rand(25)$	PSNR	24.97	33.37	34.12	34.10	32.30	32.58	33.16	32.69	34.25	34.04	33.83	34.10	34.60
	SSIM	0.8919	0.9821	0.9846	0.9849	0.9773	0.9733	0.9813	0.9767	0.9854	0.9846	0.9804	0.9844	0.9857
	SAM	0.1174	0.0419	0.0384	0.0391	0.0483	0.0498	0.0430	0.0461	0.0370	0.0382	0.0429	0.0377	0.0365
	ERGAS	2.0058	0.7612	0.7110	0.7193	0.8605	0.9186	0.7781	0.8625	0.6855	0.7041	0.7874	0.7032	0.6809
$\sigma = rand(50)$	PSNR	19.04	30.87	31.86	31.90	30.77	31.78	32.30	30.25	32.50	32.23	32.80	32.75	33.46
	SSIM	0.6794	0.9677	0.9742	0.9743	0.9672	0.9678	0.9769	0.9608	0.9783	0.9759	0.9756	0.9776	0.9807
	SAM	0.2285	0.0530	0.0497	0.0478	0.0526	0.0532	0.0461	0.0613	0.0432	0.0464	0.0473	0.0456	0.0424
	ERGAS	3.9659	1.0154	0.9050	0.9024	1.0281	1.0007	0.8600	1.1153	0.8401	0.8662	0.8767	0.8329	0.7878
Salt & Pepper	PSNR	15.74	28.88	29.22	30.08	28.26	30.18	30.39	28.82	30.68	29.98	30.54	30.77	31.85
	SSIM	0.4806	0.9448	0.9485	0.9590	0.9382	0.9508	0.9581	0.9456	0.9654	0.9580	0.9648	0.9636	0.9720
	SAM	0.3324	0.0558	0.0638	0.0554	0.0661	0.0614	0.0535	0.0708	0.0511	0.0580	0.0560	0.0536	0.0496
	ERGAS	5.7717	1.2757	1.2222	1.1091	1.3707	1.1972	1.1122	1.3102	1.0295	1.1221	1.0595	1.0254	0.9561

The best-performing results are marked in red, and the second-best performing results are marked in blue. Most of the denoising results of our proposed method are marked in red, which means that our method has more advantages than SOTA methods. It indicates that our method has achieved remarkable results in the field of hyperspectral denoising and has high applicability.

TABLE III
QUANTITATIVE EVALUATION OF DENOISING RESULTS OF PAVIA UNIVERSITY DATASET

			Sparse methods			Low-rank methods			DL methods					
Noise	Index	Noisy	BM3D	BM4D	TDL	LRTA	GLF	NGmeet	HSID-CNN	deepHSprior	SMDS-Net	FastHyMix	T3SC	Proposed
$\sigma = 25$	PSNR	20.89	29.44	33.71	33.40	27.72	34.25	34.59	29.43	33.21	32.84	34.72	34.53	34.29
	SSIM	0.6834	0.9452	0.9808	0.9742	0.9067	0.9819	0.9817	0.9400	0.9769	0.9758	0.9823	0.9817	0.9837
	SAM	0.4520	0.1344	0.0986	0.0987	0.2192	0.0912	0.0781	0.1552	0.0917	0.1179	0.0909	0.0856	0.0875
	ERGAS	9.1533	3.2038	1.9553	1.9992	4.1168	1.9059	1.8602	3.2817	2.0685	2.1611	1.8521	1.8917	1.8508
$\sigma = 50$	PSNR	15.69	25.22	27.53	27.47	24.43	28.46	27.70	26.01	27.44	27.96	28.31	28.60	28.17
	SSIM	0.3738	0.8614	0.9296	0.9354	0.8219	0.9389	0.9224	0.8704	0.9291	0.9413	0.9410	0.9410	0.9415
	SAM	0.6715	0.1884	0.1572	0.1583	0.2511	0.1458	0.1510	0.2030	0.1510	0.1561	0.1739	0.1504	0.1502
	ERGAS	14.3354	4.6732	3.5957	3.5650	5.3094	3.6305	3.8772	4.4459	3.5941	3.4007	3.6117	3.4627	3.3625
$\sigma = 75$	PSNR	12.72	22.17	23.37	23.01	21.75	23.63	22.86	23.51	23.26	23.28	23.45	23.60	23.87
	SSIM	0.2136	0.7735	0.8581	0.8623	0.7370	0.8658	0.8121	0.7822	0.8714	0.8754	0.8711	0.8777	0.8790
	SAM	0.7799	0.2301	0.2040	0.2076	0.2594	0.2069	0.2068	0.2497	0.2008	0.2090	0.2418	0.2191	0.1926
	ERGAS	17.3092	5.8344	5.0915	5.1866	6.3246	5.3499	5.7010	5.6277	5.0656	5.0554	5.3298	5.0600	4.9476
$\sigma = 100$	PSNR	10.76	19.74	20.38	19.88	19.32	20.33	19.75	20.90	20.20	20.08	20.42	21.17	21.53
	SSIM	0.1327	0.6873	0.7788	0.7688	0.6343	0.7740	0.6808	0.6905	0.7904	0.7891	0.7872	0.7906	0.8011
	SAM	0.8374	0.2610	0.2391	0.2449	0.2732	0.2364	0.2490	0.2886	0.2560	0.2454	0.2934	0.2436	0.2319
	ERGAS	19.0251	6.8704	6.4478	6.5497	7.3437	6.7141	7.0558	6.5298	6.4013	6.4314	6.6766	6.3880	6.2373
$\sigma = rand(25)$	PSNR	25.43	32.40	37.03	38.05	30.05	34.90	35.12	31.10	36.97	38.21	38.50	38.23	38.64
	SSIM	0.8626	0.9740	0.9912	0.9927	0.9448	0.9825	0.9833	0.9739	0.9904	0.9929	0.9910	0.9909	0.9940
	SAM	0.2880	0.1180	0.0766	0.0635	0.1810	0.0842	0.0820	0.1076	0.0663	0.0657	0.0630	0.0669	0.0608
	ERGAS	5.6761	2.3211	1.3603	1.2197	3.2813	1.9172	1.8999	2.4115	1.3758	1.1962	1.3106	1.3518	1.1304
$\sigma = rand(50)$	PSNR	20.05	29.01	33.02	33.38	27.20	33.39	30.72	29.05	33.21	33.10	33.76	33.29	34.16
	SSIM	0.3694	0.9403	0.9780	0.9807	0.8965	0.9796	0.9498	0.9345	0.9769	0.9770	0.9818	0.9797	0.9827
	SAM	0.4831	0.1445	0.1044	0.0977	0.2259	0.0981	0.1032	0.1642	0.0963	0.0992	0.0980	0.0984	0.0907
	ERGAS	9.8642	3.3295	2.0954	2.0447	4.2791	1.9942	2.8864	3.4076	2.0685	2.2084	1.9560	2.0017	1.8508
Salt & Pepper	PSNR	14.33	26.13	28.01	27.40	24.05	27.97	28.04	25.22	27.55	27.67	27.43	27.72	28.05
	SSIM	0.3147	0.8657	0.9420	0.9483	0.8552	0.9593	0.9607	0.8711	0.9522	0.9543	0.9518	0.9556	0.9680
	SAM	0.7263	0.1798	0.1316	0.1354	0.2422	0.1273	0.1233	0.2422	0.1374	0.1356	0.1373	0.1306	0.1159
	ERGAS	17.0816	4.7107	3.4075	3.4985	5.1301	3.2789	3.2510	4.8144	3.4331	3.3923	3.4266	3.3418	3.1759

The best-performing results are marked in red, and the second-best performing results are marked in blue. Most of the denoising results of our proposed method are marked in red, which means that our method has more advantages than SOTA methods. It indicates that our method has achieved remarkable results in the field of hyperspectral denoising and has high applicability.

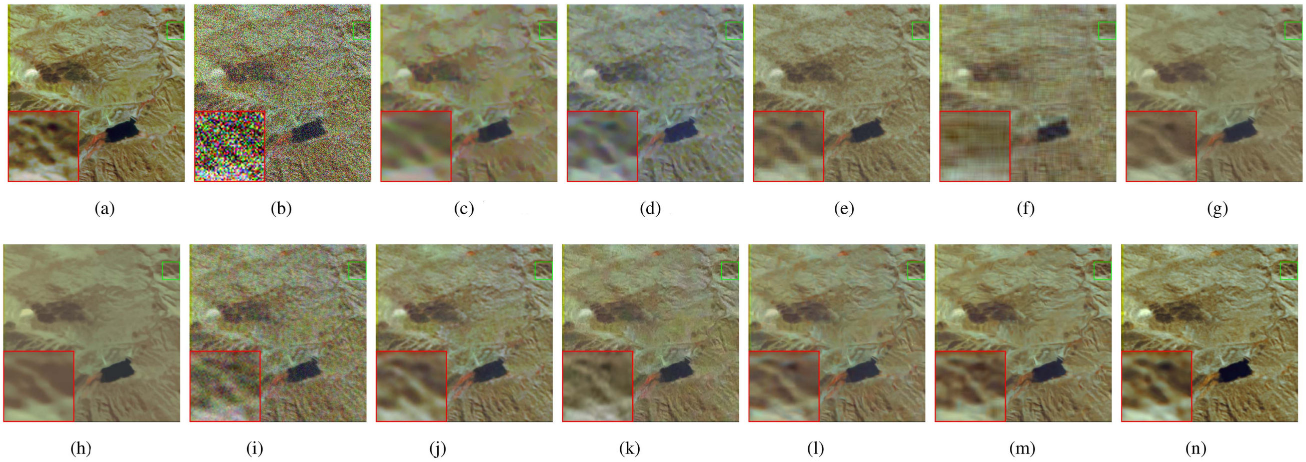


Fig. 7. Visual results for the OHS-3-D HSI with noise level $\sigma = 100$. (a) GT. (b) Noisy. (c) BM3D. (d) BM4D. (e) TDL. (f) LRTA. (g) GLF. (h) NGmeet. (i) HSID-CNN. (j) deepHSprior. (k) SMDS-Net. (l) FastHyMix. (m) T3SC. (n) Proposed.

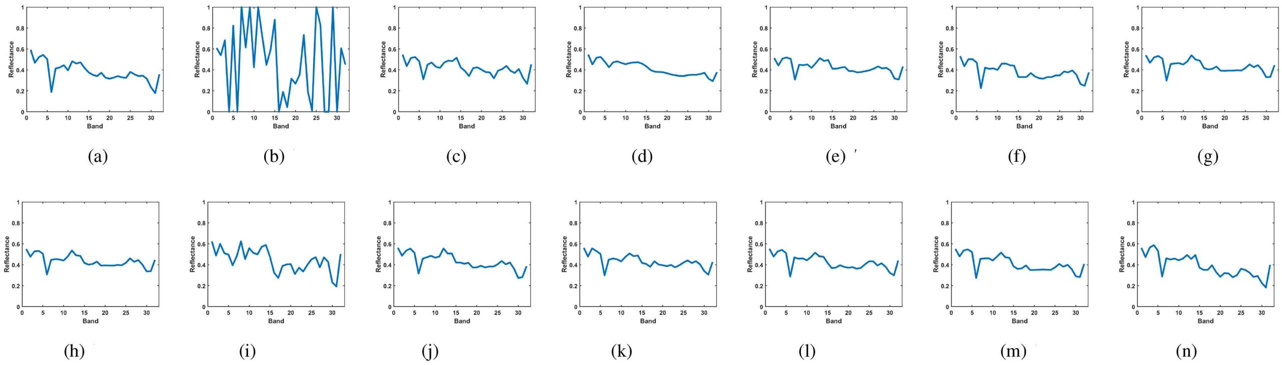


Fig. 8. Denoising results of pixel (75, 475) in the OHS-3D HSI with noise level $\sigma = 100$. (a) GT. (b) Noisy. (c) BM3D. (d) BM4D. (e) TDL. (f) LRTA. (g) GLF. (h) NGmeet. (i) HSID-CNN. (j) deepHSprior. (k) SMDS-Net. (l) FastHyMix. (m) T3SC. (n) Proposed.

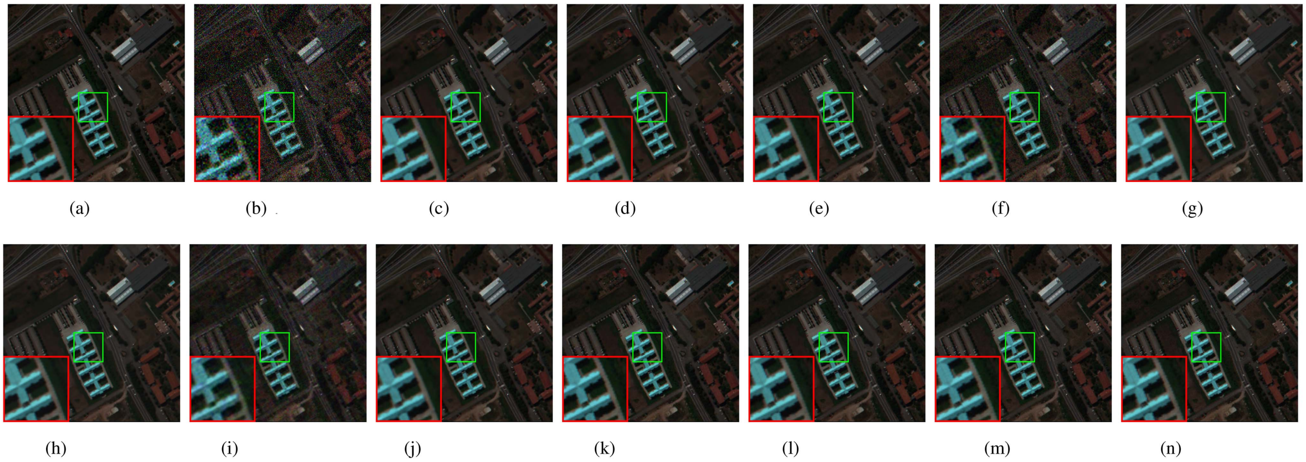


Fig. 9. Visual results for the Pavia University dataset. (a) GT. (b) Noisy. (c) BM3D. (d) BM4D. (e) TDL. (f) LRTA. (g) GLF. (h) NGmeet. (i) HSID-CNN. (j) deepHSprior. (k) SMDS-Net. (l) FastHyMix. (m) T3SC. (n) Proposed.

TABLE IV
QUANTITATIVE EVALUATION OF DENOISING RESULTS OF OHS-3 C HSI DATASET

			Sparse methods			Low-rank methods			DL methods					
Noise	Index	Noisy	BM3D	BM4D	TDL	LRTA	GLF	NGmeet	HSID-CNN	deepHSprior	SMDS-Net	FastHyMix	T3SC	Proposed
$\sigma = 25$	PSNR	20.50	31.91	32.80	33.04	30.44	32.25	34.38	32.28	33.93	33.90	34.08	34.11	34.81
	SSIM	0.6481	0.9591	0.9667	0.9685	0.9445	0.9528	0.9772	0.9407	0.9751	0.9711	0.9756	0.9760	0.9799
	SAM	0.3882	0.0965	0.0946	0.0876	0.1189	0.1083	0.0774	0.1111	0.0796	0.0874	0.0816	0.0788	0.0732
	ERGAS	7.0468	1.9169	1.7256	1.6829	2.2549	2.0308	1.4373	2.0063	1.5074	1.6112	1.4870	1.4755	1.3756
$\sigma = 50$	PSNR	15.24	28.17	28.78	28.73	27.07	28.98	29.25	28.47	29.66	29.81	29.90	30.45	30.63
	SSIM	0.3381	0.9061	0.9193	0.9233	0.8835	0.9190	0.9324	0.8666	0.9446	0.9414	0.9489	0.9422	0.9535
	SAM	0.6119	0.1277	0.1382	0.1244	0.1472	0.1320	0.1174	0.1594	0.1101	0.1165	0.1100	0.1124	0.1094
	ERGAS	12.1082	2.7917	2.5995	2.6079	3.1249	2.6333	2.4546	3.0312	2.3113	2.3584	2.2811	2.2006	2.1768
$\sigma = 75$	PSNR	12.40	24.85	25.20	24.80	23.80	25.10	24.64	24.61	25.30	25.45	25.66	25.77	26.21
	SSIM	0.1960	0.8407	0.8572	0.8675	0.8042	0.8636	0.8475	0.7936	0.8948	0.8877	0.8958	0.8954	0.9071
	SAM	0.7231	0.1554	0.1736	0.1538	0.1689	0.1605	0.1565	0.1891	0.1443	0.1499	0.1496	0.1437	0.1389
	ERGAS	15.1998	3.7443	3.5920	3.6904	4.1241	3.6907	3.7651	4.2770	3.4926	3.5474	3.4826	3.4651	3.3767
$\sigma = 100$	PSNR	10.56	21.97	22.86	21.63	21.00	21.92	21.32	20.81	21.98	22.12	22.20	22.54	23.4538
	SSIM	0.1268	0.7713	0.7913	0.8024	0.7154	0.7973	0.7411	0.6847	0.8340	0.8294	0.8337	0.8352	0.8429
	SAM	0.7804	0.1794	0.1989	0.1802	0.1904	0.1849	0.1857	0.2212	0.1724	0.1790	0.1777	0.1700	0.1632
	ERGAS	17.0083	4.7300	4.6145	4.7898	5.1447	4.7852	4.9787	5.6814	4.6357	4.6808	4.6531	4.6233	4.5100
$\sigma = rand(25)$	PSNR	25.14	34.01	35.13	35.01	32.58	33.00	34.97	34.59	35.41	35.86	36.28	35.99	36.43
	SSIM	0.8432	0.9756	0.9810	0.9804	0.9665	0.9586	0.9805	0.9656	0.9815	0.9823	0.9822	0.9818	0.9836
	SAM	0.2360	0.0817	0.0742	0.0737	0.0982	0.1038	0.0743	0.0853	0.0705	0.0701	0.0724	0.0719	0.0687
	ERGAS	4.1706	1.5116	1.3263	1.3462	1.7768	1.9315	1.3494	1.5394	1.2818	1.3129	1.2815	1.3073	1.2399
$\sigma = rand(50)$	PSNR	19.62	31.34	32.36	32.49	30.01	31.90	33.72	31.74	33.39	33.43	34.07	33.69	35.15
	SSIM	0.5991	0.9537	0.9633	0.9648	0.9387	0.9502	0.9741	0.9335	0.9723	0.9687	0.9734	0.9724	0.9814
	SAM	0.4159	0.1027	0.0988	0.0918	0.1219	0.1104	0.0813	0.1170	0.0826	0.0903	0.0843	0.0875	0.0722
	ERGAS	7.7228	2.0323	1.8014	1.7750	2.3477	2.0694	1.5356	2.1164	1.5880	1.6733	1.5323	1.5597	1.3239
Salt & Pepper	PSNR	14.72	27.58	28.01	27.74	26.20	27.81	27.87	27.16	28.34	27.84	29.85	30.38	30.61
	SSIM	0.3226	0.9052	0.9274	0.9323	0.8865	0.9246	0.9320	0.8475	0.9480	0.9304	0.9460	0.9436	0.9607
	SAM	0.6354	0.1199	0.1311	0.1200	0.1433	0.1290	0.1205	0.1706	0.1055	0.1287	0.1137	0.1019	0.0824
	ERGAS	12.3223	2.9145	2.6292	2.7729	3.2999	2.8223	2.7326	3.4196	2.5727	2.7905	2.3811	2.3609	2.3516

The best-performing results are marked in red, and the second-best performing results are marked in blue. Most of the denoising results of our proposed method are marked in red, which means that our method has more advantages than SOTA methods. It indicates that our method has achieved remarkable results in the field of hyperspectral denoising and has high applicability.

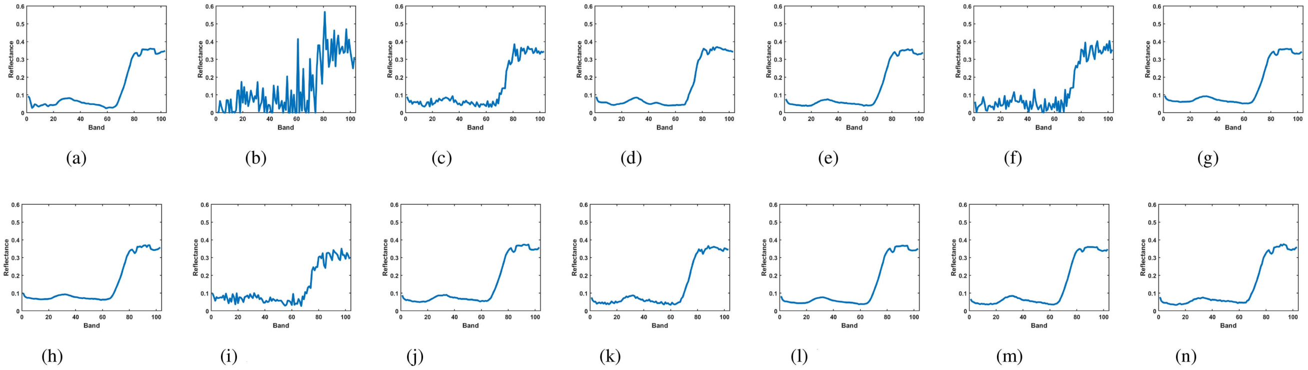
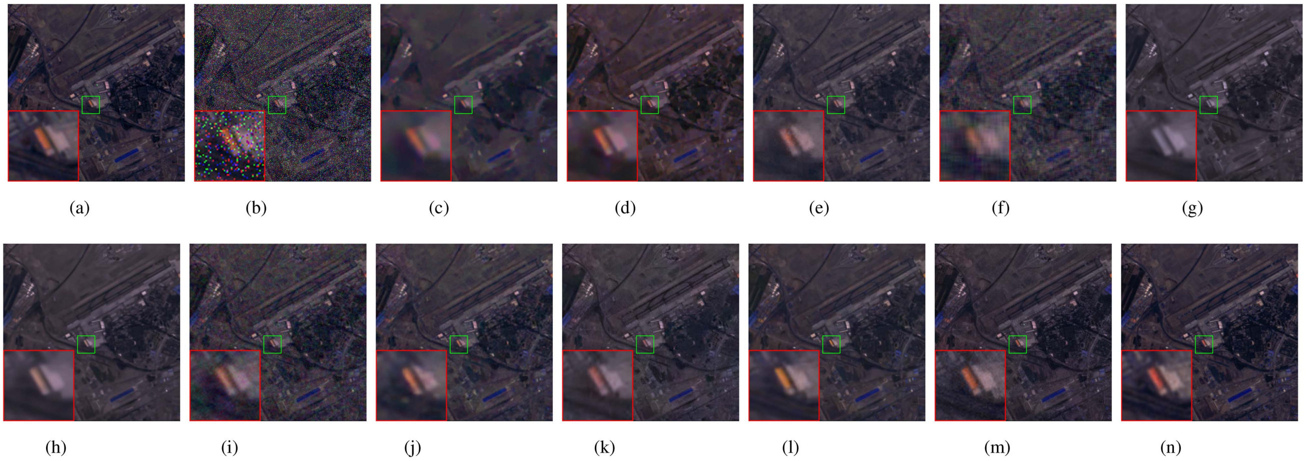


Fig. 10. Denoising results of pixel (175, 145) in the Pavia University dataset. (a) GT. (b) Noisy. (c) BM3D. (d) BM4D. (e) TDL. (f) LRTA. (g) GLF. (h) NGmeet. (i) HSID-CNN. (j) deepHSprior. (k) SMDS-Net. (l) FastHyMix. (m) T3SC. (n) Proposed.

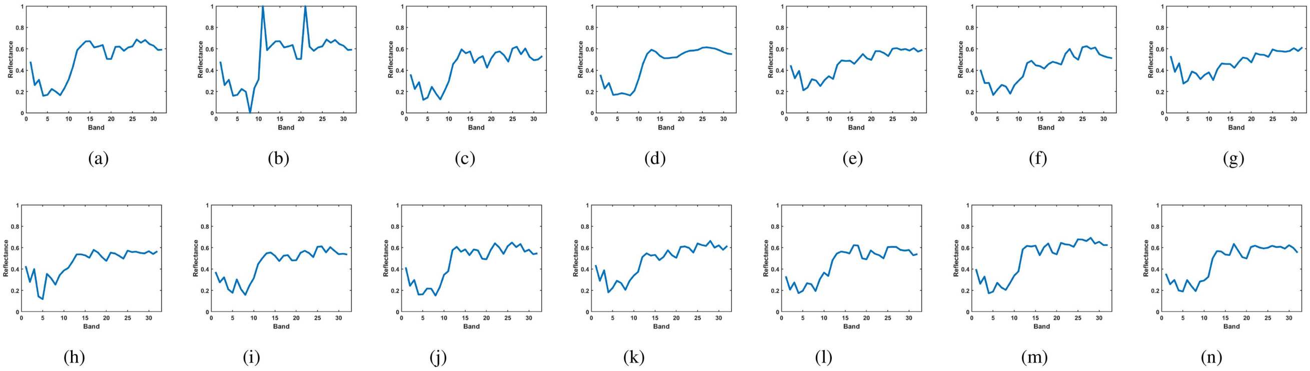
methods. It also shows that our proposed DTAPNet exhibited the best visual performance with sharpest textures and the most spatial details under different noises. The overall recovery performance of LRTA algorithm was not sufficient for this task. Some obvious noise artifacts are shown in Fig. 7. The sparse methods BM3D, BM4D, and TDL algorithms provide reasonable performance under weak noise levels. However, their performance deteriorated significantly under strong noise levels. Due to the lack of consideration of the low-rank structure of the spectrum, their performance is generally worse than low-rank methods, which is exhibited in Tables II–IV. GLF exhibited respectable noise reduction abilities under different noise levels and even obtained the best PSNR and SAM values in some cases. However, compared with our proposed method, they performed significantly worse in terms of SSIM and ERGAS, which shows that our proposed method was able to recover better structures from hyperspectral images. NGmeet makes full use of the nonlocal low-rank prior information of the data through the adaptive rank and obtains better results, especially at some week noise levels. But our method takes advantage of the strong strength of the network, considering their low-rank information for the feature layers, and can show better results in most cases and a wider range of noise. For deep learning methods, HSID-CNN did not achieve good results, since it did not combine spectral spatial structure correlation and global spectral correlation well compared to other deep learning models. The rest four deep learning methods performed well under most cases. However, our proposed method performed better.



(a) (b) (c) (d) (e) (f) (g)

(h) (i) (j) (k) (l) (m) (n)

Fig. 11. Visual results for the OHS-3 C HSI with salt & pepper noise. (a) GT. (b) Noisy. (c) BM3D. (d) BM4D. (e) TDL. (f) LRTA. (g) GLF. (h) NGmeet. (i) HSID-CNN. (j) deepHSprior. (k) SMDS-Net. (l) FastHyMix. (m) T3SC. (n) Proposed.



(a) (b) (c) (d) (e) (f) (g)

(h) (i) (j) (k) (l) (m) (n)

Fig. 12. Denoising results of pixel (285, 235) in the OHS-3 C HSI with salt & pepper noise. (a) GT. (b) Noisy. (c) BM3D. (d) BM4D. (e) TDL. (f) LRTA. (g) GLF. (h) NGmeet. (i) HSID-CNN. (j) deepHSprior. (k) SMDS-Net. (l) FastHyMix. (m) T3SC. (n) Proposed.

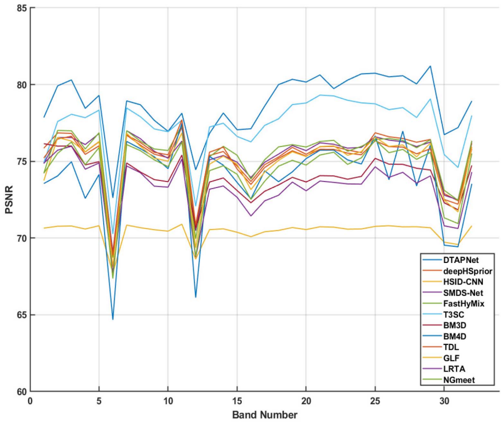


Fig. 13. PSNR values in each band with $\sigma = 100$.

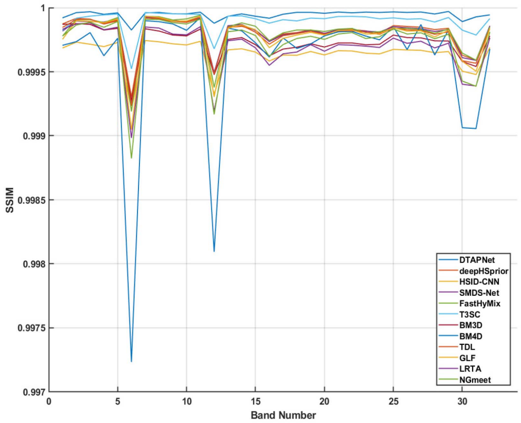


Fig. 14. SSIM values in each band with $\sigma = 100$.

One possible reason is that SMDS-Net and FastHyMix models relatively simply consider the low rank of the spectrum and the sparsity of the space, while ignoring the overall low rank structure of the data. For deepHSprior and T3SC models, this may be attributed to the fact that our method makes a better use of the multidimensional structure of feature maps by CP decomposition. Hence, our method performed better than other competing methods in terms of spatial and spectral accuracy and visual effect.

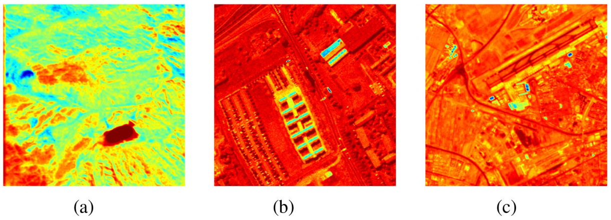


Fig. 15. Heatmaps of three different datasets. (a) OHS-3D HSI. (b) Pavia University. (c) OHS-3C HSI.

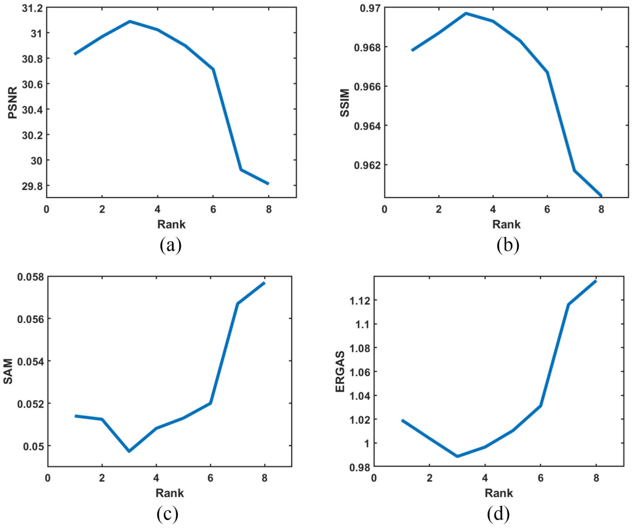


Fig. 16. 4 index under different CP ranks. (a) PSNR. (b) SSIM. (c) SAM. (d) ERGAS.

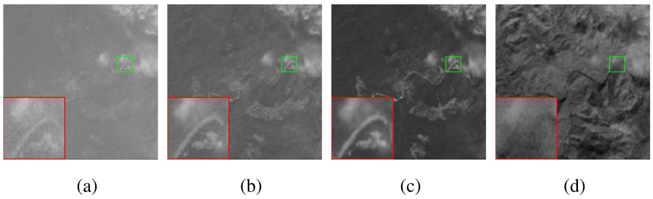


Fig. 17. Visual exhibitation for the OHS-2D HSI at band 2, 7, 14, 32. (a) Band 2. (b) Band 7. (c) Band 14. (d) Band 32.

In Figs. 8, 10, and 12 under different noises with weak or strong noise level or sparse added noise, our proposed method can obtain curves of the spectral signature closer to the ground truth among all compared methods. This reveals that our model can better recover the spectrum.

Meanwhile, the specific numerical results shown in Tables II–IV and the visual display effect exhibited in Figs. 7, 9, and 11 jointly illustrate that our proposed method has the best denoising results with different kinds of noise, including Gaussian noise with a large range, random probability distribution noise and salt & pepper noise. This reveals that our method is able to reach good denoising results under many different noises, and has wide applicability to different kinds of noise.

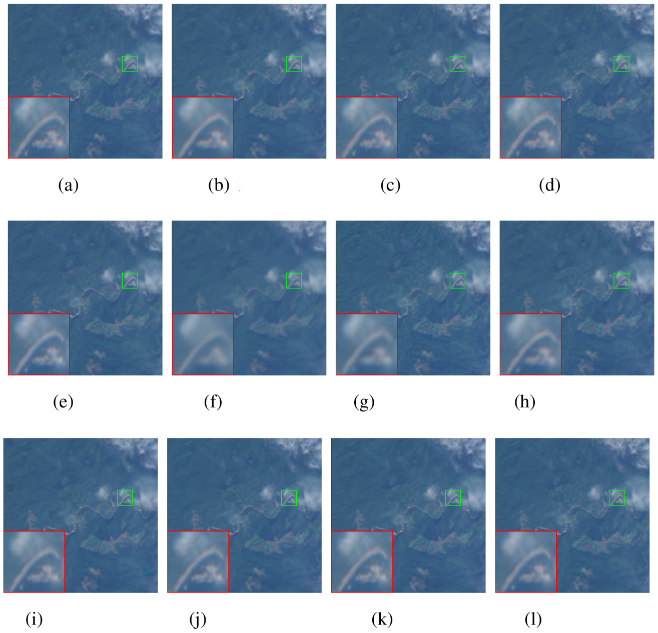


Fig. 18. Visual results for denoising the OHS-2D HSI. (a) GT. (b) BM4D. (c) TDL. (d) LRTA. (e) GLF. (f) NGmeet. (g) HSID-CNN. (h) deepHSprior. (i) SMDS-Net. (j) FastHyMix. (k) T3SC. (l) Proposed.

TABLE V
Denoising Performance of Different Prior Network

σ	Prior Network	PSNR	SSIM	SAM	ERGAS
50	ResNet based	27.87	0.8922	0.0612	1.6022
	Skip based	29.67	0.9307	0.0565	1.2843
	U-net(w/o DCPAM)	26.59	0.8745	0.0755	1.6569
	U-net based	**31.77**	**0.9712**	**0.0491**	**0.9634**
100	ResNet basedM	26.76	0.8537	0.0654	1.6782
	Skip based	28.27	0.9013	0.0596	1.4350
	U-net(w/o DCPAM)	23.56	0.7756	0.1220	2.6264
	U-net based	**30.58**	**0.9615**	**0.0530**	**1.1077**

The bold entities indicate the best performance.

TABLE VI
Quantitative Evaluation of Denoising Results of Indian Pines Dataset

	Original	NLM3D	BM4D	LRTA	GLF	deepHSprior	Proposed
OA	82.43%	92.16%	95.85%	85.30%	95.21%	98.62%	**99.13%**
kappa	0.7532	0.8900	0.9419	0.7934	0.9328	0.9807	**0.9879**

The bold entities indicate the best performance.

Moreover, in Figs. 13 and 14, the blue curves of our proposed model exhibited the highest throughout almost all bands, which also illustrates the superior performance of our proposed model. We further visualize the learned heatmaps of images to examine the partial features they activate, as shown in Fig. 15. Here, we consider the denoising experimental performance with noise level of $\sigma = 50$. For example, the first heatmap activates the various mountains and rivers in Xinjiang Uygur Autonomous Region. The heatmap on Pavia University dataset identifies the various housing areas of the land. The last heatmap identifies the

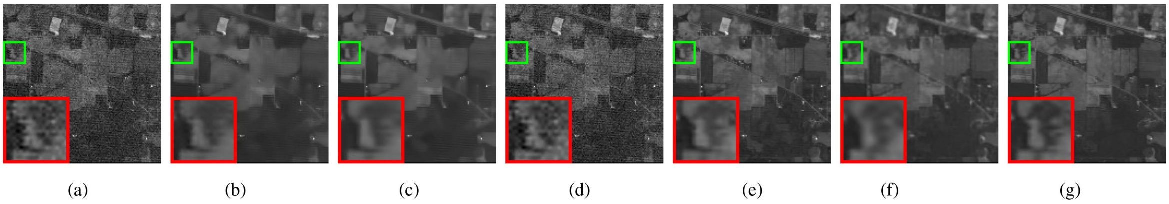


Fig. 19. Visual results for the Indian Pines dataset with band 2. (a) Noisy. (b) NLM3D. (c) BM4D. (d) LRTA. (e) GLF. (f) deepHSprior. (g) Proposed.

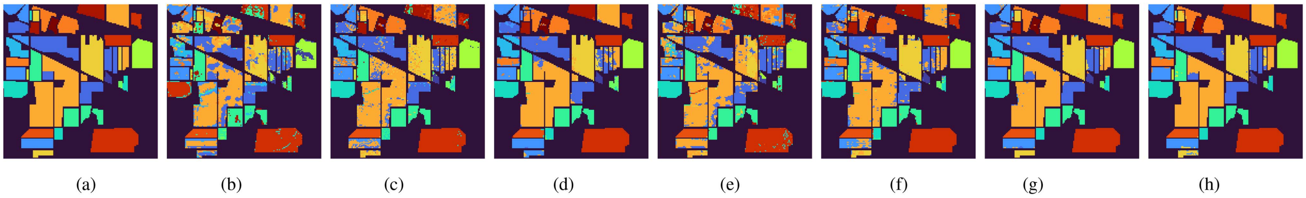


Fig. 20. Visual display of obtained classification maps. (a) GT. (b) Original. (c) NLM3D. (d) BM4D. (e) LRTA. (f) GLF. (g) deepHSprior. (h) Proposed.

highways and buildings on the land. The three hearmaps all distinguish different buildings or various terrains from each other. Therefore, we believe that our method can obtain reasonable denoising results.

D. Ablation Study

We perform a few extended experiments to investigate the effect of some parameters.

1) CP Rank: In fact, CP rank plays a significant role in supporting the denoising quality of our model. To study the influence of low CP rank on feature maps more clearly, we conducted experiments on OHS-3D HSI dataset with $\sigma = 50$ and add DCPAM separately at the second layer of a U-net based network, and calculated PSNR, SSIM, SAM and ERGAS values with different CP ranks. Fig. 16 shows the results of four metrics while tuning CP rank. The best results are obtained when rank $= 3$. This verifies that aggregating contextual information using low-rank constraints was effective.

2) Prior Network: The choice of the structure of prior network plays a important role in our model. Studies were conducted on OHS-3D HSI dataset with $\sigma = 50, 100$. Results of three mentioned prior network are presented in Table V, which shows clearly that U-net-based network performed best. Moreover, under U-net-based structure, we conducted ablation experiments where DCPAMs were removed to study the effect of DCPAMs on the denoising problem. The results are also shown in Table V. Obviously, our proposed DCPAM captures the high-dimensional intrinsic information of the feature tensors, and has a significant improvement in the denoising performance.

E. Real Data Experiments

Further, we compare our proposed model with several competing methods on two real data, including OHS-2D HSI and Indian Pines dataset.

1) OHS-2D HSI Dataset: This remote sensing HSI is taken by the "Zhuhai-1" OHS-2D hyperspectral satellite, which centered at $18.858084^\circ N$, $109.517711^\circ E$ (in Hainan Province, China). We simply denote the dataset as OHS-2D dataset in this article. We conduct our experiments on a cropped region of size $500 \times 500 \times 32$. It contains complex noises, for example, as shown in Fig. 17, the noise in bands 2, 32 is obviously stronger than that in bands 7, 14. We tune our model and apply it on the remote sensing HSI. We compare our proposed method with several sparse methods, low-rank methods and deep learning methods mentioned before. Fig. 18 presents the denoising pseudocolor images synthesized by bands 14, 7, 2. It can be seen that other methods either discard some detailed textures too much, e.g., BM4D, NGmeet, deep HS prior, or cannot remove the noise cleanly, e.g., TDL, LRTA, PARAFAC, HSID-CNN, FastHyMix. On the contrary, our proposed method removes most of the noises while preserving important structures, and performs the best. Experiments further verify that the proposed DTAPNet has strong denoising ability.

2) Indian Pines Dataset: Considering that the Indian Pines dataset contained real noise in several bands, we performed a real data experiment on Indian Pines to illustrate the flexibility of our method. Fig. 19 shows the obtained grayscale images with band 2 by different methods for visual assessment. It may be observed that the output of the LRTA method still contained some amount of residual noise and PARAFAC leaf stripes. The BM4D and NLM3D methods introduced significant blurring. The GLF and deep HS priors showed much better performance. However, as may be observed in the magnified regions shown in Fig. 19, our proposed method performed the best, with clearest outlines and details, which confirms its superior denoising performance.

For quantitative analysis, classification experiments were conducted with 16 ground-truth classes. Here, we used linear discriminant analysis (LDA) [61] as a classifier to conduct classification experiments. Table VI lists the results. The OA and kappa scores were additionally computed from the original noisy image and HSIs denoised using different methods, including our proposed model. Specifically, the OA obtained by our proposed method was 99.13%, whereas that of the original image was only 82.43%. Furthermore, compared with other methods, our proposed method has the best performance in terms of OA and kappa accuracy. As shown in Fig. 20, our model generated a map with fewer fragments.

VI. Conclusion

In this study, we have proposed a deep tensor attention prior network by introducing CP decomposition in the feature domain for the HSI denoising task. We use the traditional CP decomposition optimization algorithm to obtain more accurate low-rank decomposition components in the feature domain. By processing the components containing information of each dimension, the attention maps corresponding to the feature maps are calculated. Specifically, we built a tensor attention module based on the CP decomposition called DCPAM and added such a module to several networks. Subsequently, the entire tensor-attention-based network is integrated into an ADMM algorithm as a step of iteration. Finally, we unrolled the algorithm to obtain a prior network for the HSI denoising task. Compared with SOTA methods, several experiments on both simulated and real data demonstrated the superior performance and validity of our model.

Despite the progress made in this article on HSI denoising, several important challenges remain to be addressed. Future challenges include: exploring the effect and impact of denoising under richer noise models. Research the combination and application of tensor decomposition methods in the field of hyperspectral denoising, such as the determination of rank, the differences and actual effects of different tensor decomposition methods, the combination of tensor modules, and more cutting-edge networks, etc. Further studying the contribution of HSI denoising methods as a preprocessing step for the next step of HSI analysis, such as target detection, super-resolution, segmentation, etc., is also an interesting topic.

References

[1] Y. Gu et al., “Multimodal hyperspectral remote sensing: An overview and perspective,” *Sci. China Inf. Sci.*, vol. 64, no. 2, pp. 1–24, 2021.

[2] C.-I. Chang, *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*, vol. 1. Berlin, Germany: Springer, 2003.

[3] L. Gao and R. T. Smith, “Optical hyperspectral imaging in microscopy and spectroscopy–a review of data acquisition,” *J. Biophotonics*, vol. 8, no. 6, pp. 441–456, 2015.

[4] J. Li, Y. Li, L. He, J. Chen, and A. Plaza, “Spatio-temporal fusion for remote sensing data: An overview and new benchmark,” *Sci. China Inf. Sci.*, vol. 63, no. 4, pp. 1–17, 2020.

[5] W. Li, F. Feng, H. Li, and Q. Du, “Discriminant analysis-based dimension reduction for hyperspectral image classification: A survey of the most recent advances and an experimental comparison of different techniques,” *IEEE Geosci. Remote Sens. Mag.*, vol. 6, no. 1, pp. 15–34, Mar. 2018.

[6] X. Lu, H. Wu, Y. Yuan, P. Yan, and X. Li, “Manifold regularized sparse NMF for hyperspectral unmixing,” *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2815–2826, May 2013.

[7] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, “Deep learning-based classification of hyperspectral data,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.

[8] N. Acito, M. Diani, and G. Corsini, “Signal-dependent noise modeling and model parameter estimation in hyperspectral images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 8, pp. 2957–2971, Aug. 2011.

[9] D. Hong et al., “Interpretable hyperspectral artificial intelligence: When nonconvex modeling meets hyperspectral remote sensing,” *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 2, pp. 52–87, Jun. 2021.

[10] B. Rasti, J. R. Sveinsson, M. O. Ulfarsson, and J. A. Benediktsson, “Hyperspectral image denoising using first order spectral roughness penalty in wavelet domain,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2458–2467, Jun. 2014.

[11] H. Zhang, “Hyperspectral image denoising with cubic total variation model,” *Int. Soc. Photogrammetry Remote Sens.*, vol. 7, pp. 95–98, 2012.

[12] 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, 2016.

[13] Y. Fu, A. Lam, I. Sato, and Y. Sato, “Adaptive spatial-spectral dictionary learning for hyperspectral image restoration,” *Int. J. Comput. Vis.*, vol. 122, no. 2, pp. 228–245, 2017.

[14] T. Lu, S. Li, L. Fang, Y. Ma, and J. A. Benediktsson, “Spectral–spatial adaptive sparse representation for hyperspectral image denoising,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 373–385, Jan. 2016.

[15] 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.

[16] W. He, H. Zhang, L. Zhang, and H. Shen, “Hyperspectral image denoising via noise-adjusted iterative low-rank matrix approximation,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 3050–3061, Jun. 2015.

[17] Y. Chang, L. Yan, T. Wu, and S. Zhong, “Remote sensing image stripe noise removal: From image decomposition perspective,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7018–7031, Dec. 2016.

[18] Y. Chen, W. He, N. Yokoya, and T.-Z. 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.

[19] S. Zhang, L. Wang, Y. Fu, X. Zhong, and H. Huang, “Computational hyperspectral imaging based on dimension-discriminative low-rank tensor recovery,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 10183–10192.

[20] 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.

[21] O. Sidorov and J. Yngve Hardeberg, “Deep hyperspectral prior: Single-image denoising, inpainting, super-resolution,” in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2019, pp. 3844–3851.

[22] Y. Zhao, D. Zhai, J. Jiang, and X. Liu, “Adrn: Attention-based deep residual network for hyperspectral image denoising,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 2668–2672.

[23] Z. Kan, S. Li, M. Hou, L. Fang, and Y. Zhang, “Attention-based octave network for hyperspectral image denoising,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1089–1102, 2022.

[24] S. Zhang, L. Wang, L. Zhang, and H. Huang, “Learning tensor low-rank prior for hyperspectral image reconstruction,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12006–12015.

[25] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Deep image prior,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9446–9454.

[26] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, “Image denoising by sparse 3-D transform-domain collaborative filtering,” *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.

[27] A. Buades, B. Coll, and J.-M. Morel, “A non-local algorithm for image denoising,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 2, pp. 60–65.

[28] M. Maggioni and A. Foi, “Nonlocal transform-domain denoising of volumetric data with groupwise adaptive variance estimation,” Proc. SPIE, vol. 8296, pp. 133–140, 2012.

[29] J. V. Manjón, P. Coupé, L. Martí-Bonmatí, D. L. Collins, and M. Robles, “Adaptive non-local means denoising of mr images with spatially varying noise levels,” *J. Magn. Reson. Imag.*, vol. 31, no. 1, pp. 192–203, 2010.

[30] J. Tan, Y. Ma, H. Rueda, D. Baron, and G. R. Arce, “Compressive hyperspectral imaging via approximate message passing,” *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 2, pp. 389–401, Mar. 2016.

[31] 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.*, 2014, pp. 2949–2956.

[32] 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.

[33] N. Renard, S. Bourennane, and J. Blanc-Talon, "Denoising and dimensionality reduction using multilinear tools for hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 2, pp. 138–142, Apr. 2008.

[34] X. Liu, S. Bourennane, and C. Fossati, "Denoising of hyperspectral images using the PARAFAC model and statistical performance analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 10, pp. 3717–3724, Oct. 2012.

[35] L. Zhuang, X. Fu, M. K. Ng, and J. M. Bioucas-Dias, "Hyperspectral image denoising based on global and nonlocal low-rank factorizations," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10438–10454, Dec. 2021.

[36] W. He et al., "Non-local meets global: An iterative paradigm for hyperspectral image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2089–2107, Apr. 2022.

[37] H. Wu, K. Zhang, S. Wu, M. Zhang, and S. Shi, "Hyperspectral superresolution reconstruction via decomposition of low-rank and sparse tensor," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8943–8957, Oct. 2022.

[38] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 7, pp. 2480–2495, Jul. 2021.

[39] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[41] 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, Feb. 2019.

[42] W. He, Q. Yao, N. Yokoya, T. Uezato, H. Zhang, and L. Zhang, "Spectrum-aware and transferable architecture search for hyperspectral image restoration," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 19–37.

[43] 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, Aug. 2022.

[44] T. Bodrito, A. Zouaoui, J. Chanussot, and J. Mairal, "A trainable spectral-spatial sparse coding model for hyperspectral image restoration," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 5430–5442, 2021.

[45] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 5998–6008, 2017.

[46] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[47] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

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

[49] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, 2009.

[50] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, "Tensor decomposition for signal processing and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3551–3582, Jul. 2017.

[51] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, 1966.

[52] C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan, "Tensor robust principal component analysis with a new tensor nuclear norm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 925–938, Apr. 2020.

[53] R. A. Harshman, "Foundations of the parafac procedure: Models and conditions for an 'explanatory' multimodal factor analysis," vol. 16, no. 1, p. 84, 1970.

[54] Y. Panagakis et al., "Tensor methods in computer vision and deep learning," *Proc. IEEE*, vol. 109, no. 5, pp. 863–890, May 2021.

[55] S. Boyd et al., "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[56] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," *IEEE Signal Process. Mag.*, vol. 38, no. 2, pp. 18–44, Mar. 2021.

[57] L. Zhuang and M. K. Ng, "Fasthymix: Fast and parameter-free hyperspectral image mixed noise removal," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2021.3112577.

[58] T. Okamoto and I. Yamaguchi, "Simultaneous acquisition of spectral image information," *Opt. Lett.*, vol. 16, no. 16, pp. 1277–1279, 1991.

[59] F. A. Kruse et al., "The spectral image processing system (sips)–interactive visualization and analysis of imaging spectrometer data," *Remote Sens. Environ.*, vol. 44, no. 2/3, pp. 145–163, 1993.

[60] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[61] S. Balakrishnama and A. Ganapathiraju, "Linear discriminant analysis-a brief tutorial," *Inst. Signal Inf. Process.*, vol. 18, no. 1998, pp. 1–8, 1998.

Weilin Shen received the B.S. degree in statistics in 2019 from Xi'an Jiaotong University, Xi'an, China, where he is currently working toward the Ph.D. degree in statistics from Xi'an Jiaotong University, Xi'an, China.

His research interests include high-dimensional data analysis, statistical learning, and deep learning.

Junmin Liu (Member, IEEE) was born in 1982. He received the Ph.D. degree in applied mathematics from Xi'an Jiaotong University, Xi'an, China, in 2013.

From 2011 to 2012, he has served as a Research Assistant with the Department of Geography and Resource Management, The Chinese University of Hong Kong, Hong Kong, China. From 2014 to 2017, he worked as a Visiting Scholar with the University of Maryland, College Park, MD, USA. He is currently a Full Professor with the School of Mathematics and Statistics, Xi'an Jiaotong University. He has authored or coauthored over 60+ research papers in international conferences and journals. His research interests include machine learning, and image processing, typically for remotely sensed image fusion, hyperspectral unmixing, object detection, and so on.

Dr. Liu is a Member of CSIAM, the Topic Editor of *Remote Sensing*, and the Reviewer Editor of *Frontiers in Computational Neuroscience*.

Jinhai Li was born in 1984. He received the M.Sc. degree from Guangxi University, Nanning, China, in 2009, and the Ph.D. degree from Xi'an Jiaotong University, Xi'an, China, in 2012, both in science.

He is currently a Professor with the Kunming University of Science and Technology, Kunming, China. In recent years, he has published more than 40 papers in IEEE TKDE, IEEE TCYB, IEEE TFS, IEEE SMCA, Pattern Recognition, etc. His research interests include Big Data, cognitive computing, granular computing, and formal concept analysis.

Dr. Li is currently an Area Editor of *International Journal of Approximate Reasoning*, and Associate Editor of *International Journal of Machine Learning and Cybernetics*.

Chao Tian was born in 1982. He received the B.S. degree in geographic information system from Northwest University, Xi'an, China, in 2005 and the M.S. degree in geodesy and geomatics engineering from Chang'an University, Xi'an, China, in 2012.

Since 2012, he has been a Senior Engineer with the First Institute of Photogrammetry and Remote Sensing of the Ministry of Natural Resources, Xi'an, China. His research interests include the application of photogrammetry and remote sensing.