CS2DIPs: Unsupervised HSI Super-Resolution Using Coupled Spatial and Spectral DIPs

Yuan Fang, Yipeng Liu, *Senior Member, IEEE*, Chong-Yung Chi, *Life Fellow, IEEE*, Zhen Long, *Student Member, IEEE*, and Ce Zhu, *Fellow, IEEE*

***Abstract*— In recent years, fusing high spatial resolution multispectral images (HR-MSIs) and low spatial resolution hyperspectral images (LR-HSIs) has become a widely used approach for hyperspectral image super-resolution (HSI-SR). Various unsupervised HSI-SR methods based on deep image prior (DIP) have gained wide popularity thanks to no pre-training requirement. However, DIP-based methods often demonstrate mediocre performance in extracting latent information from the data. To resolve this performance deficiency, we propose a coupled spatial and spectral deep image priors (CS2DIPs) method for the fusion of an HR-MSI and an LR-HSI into an HR-HSI. Specifically, we integrate the nonnegative matrix-vector tensor factorization (NMVTF) into the DIP framework to jointly learn the abundance tensor and spectral feature matrix. The two coupled DIPs are designed to capture essential spatial and spectral features in parallel from the observed HR-MSI and LR-HSI, respectively, which are then used to guide the generation of the abundance tensor and spectral signature matrix for the fusion of the HSI-SR by mode-3 tensor product, meanwhile taking some inherent physical constraints into account. Free from any training data, the proposed CS2DIPs can effectively capture rich spatial and spectral information. As a result, it exhibits much superior performance and convergence speed over most existing DIP-based methods. Extensive experiments are provided to demonstrate its state-of-the-art overall performance including comparison with benchmark peer methods.**

***Index Terms*— Hyperspectral image, multispectral image, deep image prior, super-resolution, nonnegative matrix-vector tensor factorization.**

I. Introduction

HYPERSPECTRAL images (HSIs) capture scenes across continuous wavelengths of the electromagnetic spectrum [1], [2]. They contain hundreds of spectral bands with rich spectral information, and have been widely applied in remote sensing for target recognition [3], [4], geological surveying [5], [6], plant disease detection [7], [8], etc.

Manuscript received 4 September 2023; revised 7 March 2024; accepted 13 April 2024. Date of publication 24 April 2024; date of current version 29 April 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62171088 and Grant 62020106011 and in part by the Ministry of Science and Technology under Grant MOST 111-2221-E-007-047-MY2. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Zhenzhong Chen. *(Corresponding author: Yipeng Liu.)*

Yuan Fang, Yipeng Liu, Zhen Long, and Ce Zhu are with the School of Communication and Information Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: rock_fy@163.com; yipengliu@uestc.edu.cn; zhen.long@uestc.edu.cn; eczhu@uestc.edu.cn).

Chong-Yung Chi is with the Institute of Communications Engineering and the Department of Electrical Engineering, National Tsing Hua University, Hsinchu 30013, China (e-mail: cychi@ee.nthu.edu.tw).

Digital Object Identifier 10.1109/TIP.2024.3390582

Due to limited solar irradiance, there is a trade-off between the spatial and spectral resolution of HSI [9]. To maximize spectral resolution, the spatial resolution is often relatively low. Due to hardware limitations, it is difficult and costly to enhance the spatial resolution of hyperspectral imaging systems. In contrast, normally multispectral image (MSI) has a much higher spatial resolution than that of HSI [10]. The fusion of HSI and MSI has emerged as an effective HSI-SR approach [11]. Existing HSI-SR methods basically fall into two categories, i.e. model-based methods, and deep learning-based methods.

Most model-based HSI-SR methods are based on the linear observation model linking the observed image and the original spectral scene. These include total variation (TV)-based methods [12], [13], [14], sparsity-based methods [15], [16], [17], [18], and low rank-based methods [19], [20], [21], [22], [23], [24]. Among all low rank-based methods, matrix decomposition-based methods unfold the HR-HSI into matrices [22], [23], [24], which may not be very effective due to lacking the precise spatial-spectral information of the underlying materials. In contrast, tensor decomposition directly processes the HSIs in their original forms [19], [20], [21], which effectively preserves both spectral and spatial correlation. While it has a theoretically grounded explanation, the performance of model-based methods is often limited to reliable priors that can be handled mathematically in optimization.

With the rapid development of deep learning, the corresponding methods have been gradually applied to HSI-SR. Convolutional neural networks have demonstrated powerful performance in HSI-SR [25], [26], [27], [28], [29], [30], owing to their excellent learning capabilities. Meanwhile, a method called deep unrolling has received more and more attention thanks to its integration of the interpretability of model-based methods and the powerful mapping capability of deep learning-based methods [29], [30]. However, deep learning-based methods often suffer from poor generalization and require a large number of training data. To deal with such problems, unsupervised self-encoder approaches have been proposed for HSI-SR by fusing HR-MSI and LR-HSI observations without requiring training [31], [32], [33], [34], [35]. However, unsupervised self-encoders need careful network design and optimization.

Recently, an unsupervised image restoration method called deep image prior (DIP) utilizes untrained CNN structures to fit contaminated data from scratch to image data restoration

[36], [37], [38], [39], [40], [41]. DIP's effectiveness stems from spectral shift [42], [43]. It first learns low-frequency data like real images, followed by high-frequency data like noise. However, DIP-based methods often may not effectively utilize the spatial-spectral structure of HSI, so resulting in slow convergence and mediocre performance. The performance enhancement of DIP-based methods for HSI remains an open challenge.

In this paper, we propose a novel coupled spatial and spectral deep image priors (CS2DIPs) method for HSI-SR. Specifically, we integrate the nonnegative matrix-vector tensor factorization (NMVTF) [44], [45] into the DIP framework to jointly train the abundance tensor and spectral signature matrix. Then we propose a new DIP-based network that guides the learning of these two with the given observations of HR-MSI and LR-HSI as the inputs to the network. Finally, non-negativity and sum-to-one constraints are applied to the learned abundance tensor and spectral signature matrix for physical realism.

The main contributions are summarized as follows:

1) The proposed novel CS2DIPs framework uses NMVTF to transform the learning of HSI-SR into the learning of abundance tensor (a DIP guided by HR-MSI) and spectral feature matrix (a DIP guided by LR-HSI), thereby effectively and efficiently extracting various HSI spatial and spectral attributes with super-resolution, followed by their fusion for the recovery of the desired HSI-SR, and meanwhile significantly improving performance and convergence speed of DIP.
2) As for the learning process of CS2DIPs (cf. Fig. 2), specifically, the observed HR-MSI and LR-HSI are fed into two separate U-Nets with skip connections. Their deep and surface features then jointly guide the learning of the upsampling-based abundance tensor and spectral signature matrix, meanwhile integrating their inherent constraints (i.e., non-negativity on both and sum-to-one on the abundance tensor).
3) Extensive experiments on both simulated data and real data are provided to demonstrate the efficacy of the proposed CS2DIPs, exhibiting state-of-the-art unsupervised HSI-SR fusion performance, as well as comparison with some benchmark approaches.

II. Notations and Problem Formulation

A. Notations

In this paper, a scalar, a vector, a matrix, and a tensor are denoted by a lowercase letter x, a boldface lowercase letter $\mathbf{x}$, a boldface capital letter $\mathbf{X}$, and a calligraphic letter $\mathcal{X}$, respectively. For a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, $\mathcal{X} \geq 0$ means that $\mathcal{X}(i_1, i_2, \ldots, i_N) = x_{i_1,i_2,\ldots,i_N} \geq 0$ for all $i_n \in [I_n] \triangleq \{1, 2, \ldots, I_n\}$ and $n \in [N]$. $\lceil \cdot \rceil$ denotes the ceiling function.

B. Preliminaries on Tensor Computation

Definition 1 (Mode-n Product): The mode-n product of tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ and martix $\mathbf{B} \in \mathbb{R}^{J \times I_n}$ is defined as:

$$\mathcal{C} = \mathcal{A} \times_n \mathbf{B} \in \mathbb{R}^{I_1 \times \cdots \times I_{n-1} \times J \times I_{n+1} \times \cdots \times I_N} \tag{1}$$

whose entries are defined as:

$$\mathcal{C}(i_1, \ldots, j, \ldots, i_N) = \sum_{i_n=1}^{I_n} \mathcal{A}(i_1, \ldots, j, \ldots, i_N)\mathbf{B}(j, i_n) \tag{2}$$

Definition 2 (Tensor Inner Product): The inner product of two tensors $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ with the same size is defined as:

$$\langle \mathcal{A}, \mathcal{B} \rangle = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_N=1}^{I_N} a_{i_1,\ldots,i_N} b_{i_1,\ldots,i_N} \tag{3}$$

Definition 3 (Vector Outer Product): The outer product of two vectors $\mathbf{a} \in \mathbb{R}^I$, $\mathbf{b} \in \mathbb{R}^J$ is defined as:

$$\mathbf{C} = \mathbf{a} \circ \mathbf{b} \in \mathbb{R}^{I \times J} \tag{4}$$

Similarly, extending to multidimensional space, the outer product for vectors $\mathbf{a}_n \in \mathbb{R}^{I_n}$, $n = 1, \ldots, N$ is defined as:

$$\mathcal{C} = \mathbf{a}_1 \circ \mathbf{a}_2 \circ \cdots \circ \mathbf{a}_N \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N} \tag{5}$$

Definition 4 (Tensor ℓ_1-Norm): The ℓ_1-norm of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ is defined as:

$$\|\mathcal{A}\|_1 = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_N=1}^{I_N} |x_{i_1,i_2,\ldots,i_N}| \tag{6}$$

Definition 5 (Frobenius Norm): The Frobenius norm of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ is defined as:

$$\|\mathcal{A}\|_{\mathrm{F}} = \sqrt{\langle \mathcal{A}, \mathcal{A} \rangle} = \Big(\sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_N=1}^{I_N} |x_{i_1,i_2,\ldots,i_N}|^2 \Big)^{1/2} \tag{7}$$

C. Related Works

HSI-SR aims to recover an HR-HSI $\mathcal{Z} \in \mathbb{R}^{W \times H \times C}$ from an LR-HSI $\mathcal{X} \in \mathbb{R}^{W_{\mathrm{HSI}} \times H_{\mathrm{HSI}} \times C}$ and an HR-MSI $\mathcal{Y} \in \mathbb{R}^{W \times H \times C_{\mathrm{MSI}}}$, where C and C_{MSI} represent the number of spectral bands in HSI and MSI, respectively. The models for the observations LR-HSI and HR-MSI can be formulated as:

$$\begin{aligned} \mathcal{X} &= \mathcal{Z} \times_1 \mathbf{S}_1 \times_2 \mathbf{S}_2 \\ \mathcal{Y} &= \mathcal{Z} \times_3 \mathbf{P} \end{aligned} \tag{8}$$

where $\mathbf{S}_1 \in \mathbb{R}^{W_{\mathrm{HSI}} \times W}$ and $\mathbf{S}_2 \in \mathbb{R}^{H_{\mathrm{HSI}} \times H}$ are the blurring and downsampling matrices along horizontal axis and vertical axis, respectively, and $\mathbf{P} \in \mathbb{R}^{C_{\mathrm{MSI}} \times C}$ denotes the spectral response matrix associated with the imaging sensor.

In this section, we review three existing groups of HSI super-resolution methods.

1) Model-Based HSI Super-Resolution: Judiciously utilizing the prior information about the images under consideration is a common approach for solving inverse problems in image restoration [15], [22], [46], [47], [48]. Model-based methods heavily rely on spectral and spatial priors. The key idea behind model-based methods can be summarized as:

$$\begin{aligned} \min_{\mathcal{Z}} \; & \|\mathcal{X} - \mathcal{Z} \times_1 \mathbf{S}_1 \times_2 \mathbf{S}_2\|_F \\ & + \|\mathcal{Y} - \mathcal{Z} \times_3 \mathbf{P}\|_F + \lambda \phi(\mathcal{Z}) \end{aligned} \tag{9}$$

where λ is a regularization parameter and $\phi(\mathcal{Z})$ is a regularization term reflecting the prior knowledge of the HSI.

By reasonably constructing regularization terms and mining basic features from information-rich images, degraded images can be effectively restored. Many methods have been applied to HSI super-resolution, including TV-based [12], [13], [14], sparsity-based [15], [16], [17], [18] and low-rank-based [19], [20], [21], [22], [23], [24], [49] methods, etc. For example, He et al. [12] used 3D total variation to describe the local space and spectral smoothness of the highlight band; Yokoya et al. [50] proposed a coupled nonnegative matrix decomposition (CNMF) to alternatively estimate the abundance matrix and endmember of the HSI; Lin et al. [51] further developed a method called CO-CNMF by judiciously applying alternating direction method of multipliers (ADMM) and various inherent matrix structures, therefore much more efficient and effective than the CNMF, besides its better robustness against noise contamination; Zhao et al. [17] proposed a hyperspectral super-resolution method based on sparse representation and spectral mixing model; Xue et al. [18] considered the spatial/spectral subspace low-rank relationships and proposed a subspace clustering method based on structured sparse low-rank representation; Han et al. [22] explored the similar patch structure of images and proposed a super-resolution model based on non-local similarity; Xu et al. [19] proposed a model based on nonlocal and CP tensor decomposition; Xu et al. [49] quantized HSI tensor into higher-order tensor and proposed a coupled Tensor Ring (TR) representation model; He et al. [20] developed a coupled TR model with spectral kernel normalization to exploit global spectral low-rank attributes.

2) Deep Learning-Based HSI Super-Resolution: In recent years, deep learning-based methods have gradually become mainstream, benefiting from the powerful data-driven feature extraction capabilities of deep networks. The key idea behind the deep learning-based methods can be summarized as:

$$\min_{\boldsymbol{\theta}} \mathrm{E}(f_{\boldsymbol{\theta}}(\mathcal{X}, \mathcal{Y}), \widehat{\mathcal{Z}}) \tag{10}$$

where $\mathrm{E}(\cdot)$ is the loss function of data-fitting error between $f_{\boldsymbol{\theta}}(\mathcal{X}, \mathcal{Y})$ and the given data $\widehat{\mathcal{Z}}$, $f_{\boldsymbol{\theta}}(\mathcal{X}, \mathcal{Y})$ is the output of the CNN parameterized by $\boldsymbol{\theta}$.

Deep learning-based methods can mine implicit properties of data from large amounts of data. Wang et al. [25] proposed a deep residual convolution network and used it to improve the spatial resolution of hyperspectral images; Han et al. [26] proposed a CNN-based spatial-spectral fusion architecture for fusion of LR-HSI and HR-MSI; Hu et al. [27] designed a transformer-based network that can explore the internal relationship of features globally. Meanwhile, approaches combining model-based and deep learning have gained popularity, including Deep Plug-and-Play and Deep Unrolling. Lai et al. [28] used a trained CNN denoiser as a prior in the reconstruction model; Dong et al. [29] expanded the HSI-SR optimization model and designed a new deep convolution network; Ma et al. [30] expanded the HSI-SR model and designed a network structure of Transformer+3D-CNN to explore the global spatial interaction ability and spatial-spectral correlation of data at the same time.

3) Deep Image Prior: Recently, Ulyanov et al. [36] proposed Deep Image Prior (DIP), where an untrained deep network is initialized with random weights and optimized so that the generated image approximates the target. DIP is an unsupervised approach whose prior stems solely from the fixed convolutional structure of the generative network. The key idea behind DIP can be summarized as:

$$\begin{aligned} &\min_{\boldsymbol{\theta}} \ \mathrm{E}(f_{\boldsymbol{\theta}}(\mathcal{J}) \times_1 \mathbf{S}_1 \times_2 \mathbf{S}_2, \mathcal{X}) + \mathrm{E}(f_{\boldsymbol{\theta}}(\mathcal{J}) \times_3 \mathbf{P}, \mathcal{Y}) \\ &\text{s. t.} \quad \mathcal{Z}^* = f_{\boldsymbol{\theta}}(\mathcal{J}) \end{aligned} \tag{11}$$

where $\mathcal{J}$ is an initial random input with the same dimension of $\mathcal{Z}$. Problem (11) can be handled by gradient descent method for obtaining a local minimizer provided that $\mathrm{E}(\cdot)$ is differentiable.

The powerful learning capability and flexible framework of DIP enable excellent performance on many image restoration tasks. However, traditional DIP utilizes only the network structure as a prior, thus suffering many limitations. To upgrade the DIP's learning capability, some improved DIP schemes have been proposed for HSI. Zhang et al. [52] proposed an unsupervised DIP framework that uses degenerate estimation in the HSI-SR algorithm design; Zhang et al. [53] incorporated supervised learning into the unsupervised DIP, utilizing a priori information from supervision for preliminary image fusion.

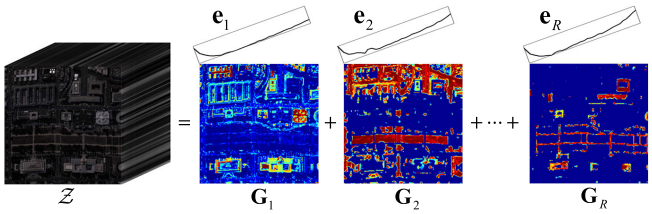

Fig. 1. The matrix-vector tensor factorization of HSI, where $\mathbf{e}_r$ and $\mathbf{G}_r$ denote the spectral signature and corresponding abundance map of *r-th* materials, respectively.

III. Proposed Method

In this section, we first present our optimization model. Then, we present the network structure and our loss function.

A. Optimization Model

1) MVTF Inspired DIP: Traditional DIP methods typically initialize the random input with the same size as the image, and optimize through iterative training [36]. However, this is difficult to jointly capture dense spatial and spectral characteristics, the results thus turn out to be suboptimal.

Methods based on MVTF capture the intrinsic structure of the tensor and represent the data in a sparse manner [45]. MVTF can decompose HSI into a series of spectral signatures (endmembers) and proportions in pixels (abundance matrices) of different materials (cf. Fig. 1), which can be defined by

$$\mathcal{Z} \approx \sum_{r=1}^{R} \mathbf{G}_r \circ \mathbf{e}_r \tag{12}$$

where $\mathbf{G}_r \in \mathbb{R}^{W \times H}$, $\mathbf{e}_r \in \mathbb{R}^{C}$ represent abundance matrix and spectral signature vector, respectively, and $r = 1, \ldots, R$, R is the number of materials in $\mathcal{Z}$.

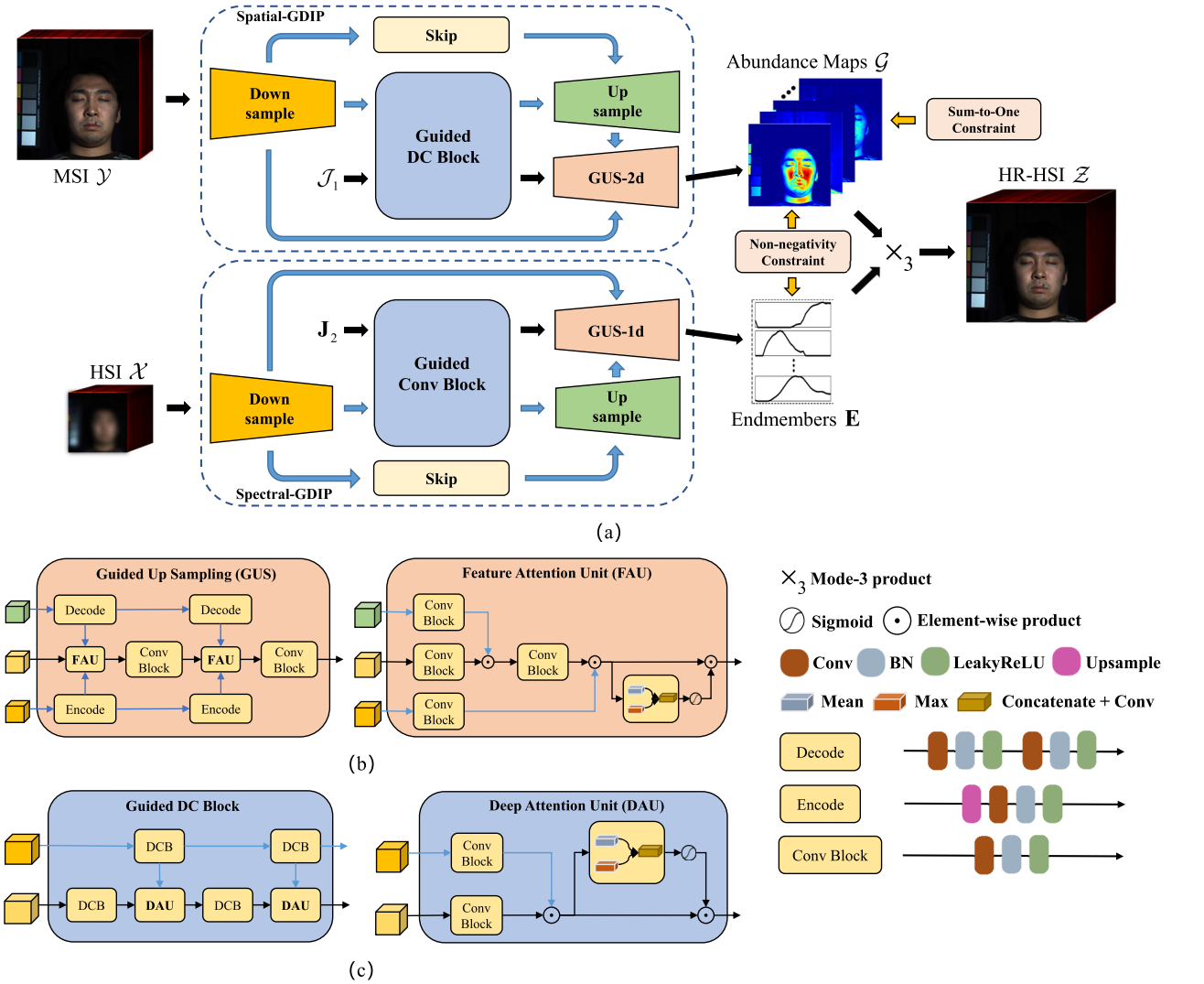

Fig. 2. (a) The proposed CS2DIPs model, which consists of two main modules with identical architecture: Spatial-GDIP and Spectral-GDIP, each including (b) a two-layer Guided Up Sampling (GUS) and the embedded Feature Attention Unit (FAU), and (c) a two-layer Guided Deformed Convolution (DC) Block (for the Spatial-GDIP) or Guided Conv Block (for the Spectral-GDIP) and the embedded Deep Attention Unit (DAU), except for 1D (2D) convolutions performed in Conv (DC) Block. Finally, after the normalization via non-negativity and sum-to-one, the outputs of Spatial-GDIP and Spectral-GDIP are fused into the desired HR-HSI by mode-3 tensor product.

To effectively utilize the MVTF, we adopt the following low-rank representation:

$$\mathcal{Z} \approx \sum_{r=1}^{R} \mathbf{G}_r \circ \mathbf{e}_r = \mathcal{G} \times_3 \mathbf{E} \tag{13}$$

where $\mathcal{G} = [\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_R] \in \mathbb{R}^{W \times H \times R}$ and $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_R] \in \mathbb{R}^{C \times R}$ represent abundance tensor and spectral signature matrix, respectively.

2) Inherent Constraints on $\mathcal{G}$ and $\mathbf{E}$: Traditional DIP training methods primarily consider the loss function used and its impact on the restored data, which should not violate physical constraints [37]. To this end, we incorporate nonnegativity and sum-to-one constraints, simply because the former is a common constraint for both $\mathcal{G}$ and $\mathbf{E}$ while the latter is necessary, particularly for $\mathcal{G}$.

3) The Proposed Optimization Model: To efficiently exploit the strength of NMVTF and constraints based on physical properties, the proposed optimization model for the HSI super-resolution is formulated as

$$\begin{aligned} \min_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2} \; & \|\mathcal{Y} - \mathcal{Z} \times_3 \mathbf{P}\|_{\mathrm{F}}^2 + \|\mathcal{X} - \mathcal{Z} \times_1 \mathbf{S}_1 \times_2 \mathbf{S}_2\|_{\mathrm{F}}^2 \\ \text{s. t.} \quad & \mathcal{Z} = \mathcal{G} \times_3 \mathbf{E}, \quad \mathcal{G} = f_{\boldsymbol{\theta}_1}(\mathcal{J}_1, \mathcal{Y}) \geq 0, \\ & \mathbf{E} = f_{\boldsymbol{\theta}_2}(\mathbf{J}_2, \mathcal{X}) \geq 0, \quad \sum_{r=1}^{R} \mathcal{G}_{i,j,r} = 1 \end{aligned} \tag{14}$$

where $\mathcal{J}_1 \in \mathbb{R}^{\widehat{W} \times \widehat{H} \times N_1}$, $\mathbf{J}_2 \in \mathbb{R}^{\widehat{B} \times N_2}$ denote random inputs to the CS2DIPs and $f_{\boldsymbol{\theta}_1}(\cdot, \mathcal{Y})$ and $f_{\boldsymbol{\theta}_2}(\cdot, \mathcal{X})$ are the corresponding outputs of GUS-2d and GUS-1d in CS2DIPs (cf. Fig. 2), and $\widehat{W} = \lceil W/2^M \rceil$, $\widehat{H} = \lceil H/2^M \rceil$, $\widehat{B} = \lceil C/2^M \rceil$ and M denotes network depth.

B. Network Architecture

Inspired by the spectral-spatial properties of HSI and the U-net structure, the proposed CS2DIPs method is shown in Fig. 2. Next, let us present its essential constituents.

1) Network Framework With Low-Rank Prior: To efficiently utilize the low-rank prior in (14), as shown in Fig. 2(a), designed CS2DIPs model consists of two subnetworks with identical architecture: Spatial-GDIP and Spectral-GDIP for learning $\mathcal{G}$ (abundance tensor) and $\mathbf{E}$ (spectral signature matrix) respectively. Spatial-GDIP and Spectral-GDIP adopt the same GDIP architecture, though they perform 2D convolutional and 1D convolutional operations of GDIP, respectively. They are coupled via the mode-3 tensor product, and their learning processes are guided by the HR-MSI and LR-HSI observations, respectively.

2) Guided Upsampling Generation Under-Parameterization Subframework: Inspired by [54] and [55], we propose a guided upsampling generation subframework called GDIP (cf. Fig. 2(a)), comprising a U-Net with skip connections and an upsampling network. Generally, the input (i.e., $\mathcal{J}_1$ or $\mathbf{J_2}$) of each GDIP passing through the guidance network plays a supervisory role in the generation of $\mathcal{G}$ and $\mathbf{E}$.

Existing research works have shown that DIP's effectiveness stems from spectral bias induced by upsampling [42], [43]. Hence, we use a generative network containing only upsampling layers, and an under-parameterized design.

Based on (8), HR-MSI undergoes spatial blurring while LR-HSI experiences spectral downsampling. We incorporate HR-MSI and LR-HSI in GDIP to guide the generation of $\mathcal{G}$ and $\mathbf{E}$, respectively. This enables more rapid and accurate learning of dense spatial features and spectral signatures. The upsampling-based generation for $\mathcal{G}$ and $\mathbf{E}$ are given by:

$$\begin{aligned} \mathcal{G} &= f_{\boldsymbol{\theta}_1}(\mathcal{J}_1, \mathcal{Y}) \\ \mathbf{E} &= f_{\boldsymbol{\theta}_2}(\mathbf{J}_2, \mathcal{X}) \end{aligned} \tag{15}$$

where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ represent the up-to-date CNN parameters of Spatial-GDIP and Spectral-GDIP, respectively.

3) Guided Up Sampling: The designed guided upsampling (GUS) module enables surface-level guidance from the observations for generating $\mathcal{G}$ and $\mathbf{E}$. The downsampling and upsampling of the bootstrap network simultaneously guide the upsampling to reconstruct corresponding regions. Unlike supervised learning, our bootstrap data rely solely on either HR-MSI observations or LR-HSI ones.

The feature attention unit (FAU) is a key component of GUS. To ensure validity of the top and bottom samples, we first multiply the generated data with the top sample, and then the bottom sample after a Conv+BN+LeakyReLU layer. FAU introduces an attention mechanism to deepen the learning of spatial and spectral features of $\mathcal{G}$ and $\mathbf{E}$, respectively. Fig. 2(b) shows the details of the two-layer GUS and FAU structure.

4) Guided Deformable Convolution Block (GDC): Deformable Convolution (DC) expands the receptive field and enhances model transformation capability through adaptive kernel shapes [56]. As shown in Fig. 2(c), the designed GDC structure with stacked deformable convolutions enables free deformation of the sampling network. Study in [57] has shown that deeper stacking can yield better results. We also use a Deep Attention Unit (DAU) to enhance deep feature learning from the guidance data, and reduce DIP's burden. DAU uses an attention mechanism to selectively emphasize important features from the guided data while yielding the generated data. Since deformable convolutions only apply to 2D, we simply use conv block in Spectral-GDIP instead of DC block.

5) Normalization on GDIPs' Outputs: The constraints on $\mathcal{G}$ and $\mathbf{E}$ in (14) are incorporated at the network level. Specifically, we add ReLU operations on the generated $\mathcal{G}$ and $\mathbf{E}$ to enforce non-negativity. Meanwhile, $\mathcal{G}$ undertakes a sum-to-one normalization, i.e.,

$$\mathcal{G} := \mathcal{G} / \Big(\sum_{r=1}^{R} \mathcal{G}_{i,j,r} + \epsilon \Big) \tag{16}$$

where $\epsilon > 0$ is a small constant to avoid division by zero.

C. Loss Function

To train the model by (14), we use the Huber loss function (HLF) [58], [59], which is more robust against outliers than Frobenius norm, while less sensitive to discrete anomalies than ℓ_1 norm. The HLF can be defined by

$$\mathrm{h}_\delta(t) = \begin{cases} \dfrac{1}{2}t^2 & |t| \le \delta \\ \delta|t| - \dfrac{1}{2}\delta^2 & |t| > \delta \end{cases} \tag{17}$$

where δ is a user-defined threshold. The overall loss function used can be expressed as

$$\mathcal{L} = \mathrm{H}_\delta(\mathcal{Y} - \mathcal{Z} \times_3 \mathbf{P}) + \mathrm{H}_\delta(\mathcal{X} - \mathcal{Z} \times_1 \mathbf{S}_1 \times_2 \mathbf{S}_2), \tag{18}$$

where $\mathrm{H}_\delta(\mathcal{Y} - \mathcal{Z} \times_3 \mathbf{P}) \triangleq \sum_{ijr} \mathrm{h}_\delta(y_{ijr} - \hat{y}_{ijr})$ (i.e., sum of HLFs defined by (17) for all $y_{ijr} \in \mathcal{Y}$ and $\hat{y}_{ijr} \in \mathcal{Z} \times_3 \mathbf{P}$), and $\mathrm{H}_\delta(\mathcal{X} - \mathcal{Z} \times_1 \mathbf{S}_1 \times_2 \mathbf{S}_2)$ is defined similarly.

D. Differences Between CS2DIPs and Closely Related Approaches

Most existing DIP-based methods utilize the classic U-Net to obtain some prior information without focusing on the fundamental spatial-spectral coupling structure, which therefore may not comprehensively cover most essential spatial and spectral characteristics of hyperspectral images [37], [39], [55]. In contrast, the proposed CS2DIPs equipped with two coupled DIPs in parallel are capable of capturing more useful spatial and spectral features from the observed HR-MSI and LR-HSI, respectively, which are then used to guide the generation of the crucial abundance tensor and spectral signature matrix for the fusion of the super-resolution HSI, meanwhile taking some inherent physical constraints into account.

IV. Experimental Results

In this section, we present extensive experimental results with both simulation and real datasets to demonstrate the efficacy of the proposed CS2DIPs.

A. Datasets

We selected five simulation datasets and one real dataset for experimental evaluation. The simulation datasets include: CAVE[1], Harvard[2], Pavia Centre (PaviaC)[3], KSC[3], Washington

[1] https://www1.cs.columbia.edu/CAVE/databases/multispectral/
[2] http://vision.seas.harvard.edu/hyperspec/explore.html
[3] https://www.ehu.eus/ccwintco/index.php/

DC Mall (WDC)[4]; the real dataset is University of Houston (UH)[5].

The CAVE dataset contains 32 HSIs, each with 512 × 512 pixels and 31 spectral bands, covering the visible spectrum from 400 nm to 700 nm. We use the 'face_ms' image as our simulation dataset and extract 256 × 256 pixels sub-image through downsampling.

The Harvard dataset contains 50 HSIs, each with 1392 × 1040 pixels and 31 spectral bands, covering the visible spectrum from 420 nm to 720 nm. We use 'img1' image as our simulation dataset and extract 256 × 256 sub-images through downsampling and cropping.

The PaviaC dataset was obtained by the imaging spectrometer of the catoptrics system, with 1096 × 1096 pixels, 115 spectral bands, and 102 residual spectral bands. We selected a portion of 192 × 192 pixels and 102 spectral bands as the reference image.

The KSC dataset was acquired by the NASA AVIRIS instrument over the Kennedy Space Center, with 512 × 614 pixels and 176 spectral bands. We selected a 256 × 256 pixels sub-image covering the first 103 spectral bands.

The WDC dataset contains a total of 191 bands ranging from 0.4 to 2.4 um visible and near-infrared bands, with a data size of 1208 × 307. We extracted a 256 × 256 pixels sub-image by cropping the first 103 spectral bands.

The UH dataset was acquired by the National Center for Airborne Laser Mapping (NCALM) over the University of Houston campus and its neighborhood. The dataset provides HR-MSI of 83440 × 24040 pixels and LR-HSI of 4172 × 1202 pixels with 48 bands. We selected a sub-image LR-HSI of 32 × 32 × 48 and obtained an HR-MSI of 256 × 256 × 3 through cropping and downsampling.

B. Comparison Methods

The peer methods for performance comparison with the proposed CS2DIPs include four model-based methods: Hysure [60], CSTF [61], NLSTF [62], SC-LL1 [63] and six deep learning-based methods: DIP-2D [37], DIP-3D [37], DeepTensor [39], GDD [55], uSDN [31] and MIAE [34]. However, MIAE is an unsupervised blind estimation method developed without the prior information of parameters $\mathbf{P}$, $\mathbf{S}_1$ and $\mathbf{S}_2$. For a fair comparison, we further considered a non-blind counterpart, referred to as MIAE*, by replacing all the estimated values of these parameters with true values in the original MIAE, and both of them were included in the performance comparison in our experiment.

In the experiment, all datasets were normalized on the interval [0,1]. The quality of the generated HR-HSI images is evaluated using four performance indexes, including peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), spectral angle mapper (SAM), and erreur relative globale adimensionnelle de synthèse (ERGAS).

[4]https://engineering.purdue.edu/ biehl/MultiSpec/hyperspectral.html

[5]https://hyperspectral.ee.uh.edu/?page_id=1075

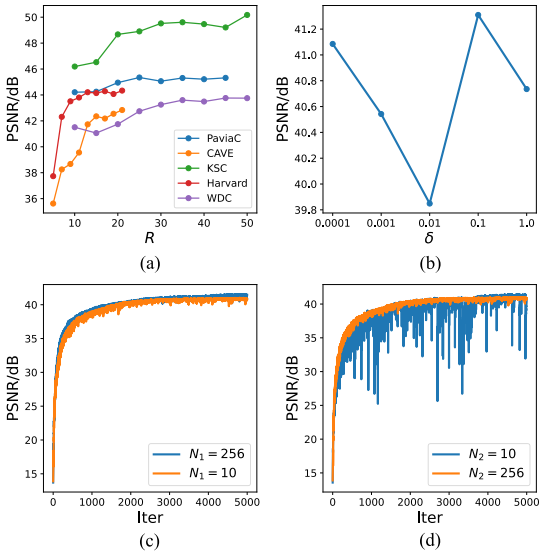

Fig. 3. Performance of the proposed CS2DIPs on five datasets for $K = 8$, in terms of PSNR versus hyperparameters (R, $\delta = 0.1$, $N_1 = 10$, $N_2 = 256$) (a), while the top right plot, bottom left plot, and bottom right plot show the performance on the CAVE dataset, specifically ($R = 10$, δ, $N_1 = 10$, $N_2 = 256$) (b), PSNR versus iteration number for ($R = 10$, $\delta = 0.1$, $N_1 \in \{10, 256\}$, $N_2 = 256$) (c), and ($R = 10$, $\delta = 0.1$, $N_1 = 10$, $N_2 \in \{10, 256\}$) (d), respectively.

C. Implementation Details

For each reference image, i.e., HR-HSI ($\mathcal{Z}$) used as the ground truth (GT), the observation image LR-HSI ($\mathcal{X}$) is generated through Gaussian blur processing on the GT in the spatial domain (Gaussian kernel with the size of 8, mean of 0, and standard deviation of $\sqrt{3}$), followed by downsampling on the resulting image, with downsampling ratio K (the ratio $|\mathcal{Z}|/|\mathcal{X}|$) equal to 8 or 16. The observation image HR-MSI ($\mathcal{Y}$) is generated by the convolution integral of the simulated spectral response (based on Nikon cameras) with the GT for each spectral band in the 3-band (4-band) HR-MSI in the CAVE and Harvard (PaviaC, KSC and WDC) dataset.

In the CS2DIPs network proposed above, we used a four-layer Spatial-GDIP and a three-layer Spectral-GDIP. In the Guided DC Block and Guided Conv Block (cf. Fig. 2(c)), we used four-layer structures. We employed the Adam optimizer with learning rate $l_r = 0.001$. The maximum number of iterations I_{max} was set to 5000.

In the experiment, the proposed CS2DIPs method was implemented by the Python framework. All experiments were run on a computer with an Intel i5-11400F CPU, 16GB RAM, and an NVIDIA RTX 3060Ti GPU.

D. Hyperparameter Setting

The hyperparameters δ, N_1, and N_2 for the proposed CS2DIPs are advisably chosen through experiments on the CAVE dataset for $K = 8$, while R chosen from the results over all five different datasets. Let us consider its PSNR performance versus i) R and δ, respectively, and ii) iteration number for $N_1, N_2 \in \{10, 256\}$. These experimental results in terms of PSNR are shown in Fig. 3, which suggests the values for parameters $\delta = 0.1$, $N_1 = 10$, $N_2 = 256$, and $R = 15(35)$ in CAVE and Harvard (PaviaC, KSC and WDC) datasets in our experiment.

TABLE I

QUANTITATIVE RESULTS OF VARIOUS METHODS IN TERMS OF PSNR AND SSIM (SAM AND ERGAS) FOR WHICH "↑" ("↓") INDICATES THAT THE LARGER (SMALLER) THE NUMERICAL VALUES, THE BETTER THE CORRESPONDING RESULTS. THE BEST RESULTS ARE SHOWN IN BOLDFACE AND THE SECOND-BEST RESULTS ARE UNDERLINED

Dataset	K		Model-based				Deep learning-based							
			Hysure	CSTF	NLSTF	SC-LL1	DIP-2D	DIP-3D	DeepTensor	GDD	uSDN	MIAE	MIAE*	CS2DIPs
PaviaC	8	PSNR↑	41.8564	44.4161	42.5509	43.7674	40.6688	31.8720	42.1538	36.4880	38.2184	44.8197	44.3061	**45.4553**
		SSIM↑	0.9772	0.9820	0.9808	0.9821	0.9749	0.9064	0.9798	0.9724	0.9744	0.9862	0.9861	**0.9870**
		SAM↓	5.2172	4.2677	4.3206	4.2138	4.9628	10.0787	4.3635	6.1030	4.8501	**3.5845**	4.0184	3.6370
		ERGAS↓	0.9977	0.8551	0.9046	0.8281	1.0252	2.4077	0.8936	1.3892	1.2591	0.7447	0.8299	**0.7302**
	16	PSNR	42.0278	43.2510	41.7238	42.5921	36.2528	23.4993	41.0583	31.8996	36.7558	40.8712	42.4956	**43.7723**
		SSIM	0.9771	0.9809	0.9800	0.9790	0.9506	0.7414	0.9744	0.9550	0.9707	0.9790	0.9837	**0.9865**
		SAM	5.2463	4.7193	**4.4215**	4.7975	7.3062	30.2313	5.1937	9.6372	5.0644	5.8260	4.5367	4.4507
		ERGAS	0.5045	0.4688	0.4817	0.4718	0.7849	6.4200	0.5126	1.1300	0.6959	0.5360	0.4690	**0.4675**
KSC	8	PSNR	44.3267	46.8281	46.5175	45.7773	44.0290	35.2691	44.8934	37.0482	41.3123	48.0106	48.5049	**49.2031**
		SSIM	0.9902	0.9912	0.9919	0.9905	0.9858	0.9362	0.9889	0.9801	0.9866	0.9945	0.9954	**0.9957**
		SAM	3.2741	2.2649	2.0807	2.2847	2.5719	5.7862	2.1700	3.7014	3.4984	1.7486	1.7074	**1.5531**
		ERGAS	0.7823	0.6134	0.5682	0.6510	0.7068	1.7036	0.6634	1.0752	1.1475	0.5531	1.0577	**0.4738**
	16	PSNR	43.7015	44.1690	46.2307	42.8605	40.2387	29.5175	42.2755	33.8124	38.4487	45.3775	44.6371	**47.8394**
		SSIM	0.9891	0.9895	0.9918	0.9854	0.9712	0.8616	0.9825	0.9761	0.9870	0.9933	0.9929	**0.9949**
		SAM	3.6288	3.3031	2.1671	3.1441	3.8842	12.7708	3.1838	5.1056	3.1398	2.2033	2.5797	**1.8494**
		ERGAS	0.4384	0.4380	0.2957	0.4815	0.5285	2.1551	0.4550	0.7637	0.7599	0.3631	0.4260	**0.2833**
face_ms (in CAVE)	8	PSNR	40.8127	41.5749	41.6309	**41.6518**	39.2696	35.4453	41.1079	40.7374	40.0221	41.5412	41.0155	**42.4317**
		SSIM	0.9821	0.9803	0.9880	0.9904	0.9709	0.9800	0.9851	0.9856	0.9836	**0.9957**	0.9920	0.9932
		SAM	8.1228	7.5396	5.2579	4.5668	7.4794	6.2131	5.3395	5.3598	7.1135	**2.8866**	2.9952	3.5676
		ERGAS	1.1091	1.0656	1.0154	1.0172	1.2446	1.7885	0.9680	1.0704	1.2076	0.9989	1.6750	**0.9210**
	16	PSNR	39.8758	39.2641	40.6773	40.5941	35.1134	33.5948	39.4047	38.2755	37.3119	38.6762	39.4252	**40.6890**
		SSIM	0.9801	0.9657	0.9861	0.9880	0.9441	0.9734	0.9800	0.9789	0.9051	**0.9930**	0.9903	0.9867
		SAM	15.2860	20.4102	5.9205	5.8135	10.2970	7.6966	6.2931	6.7279	13.2707	**3.7664**	4.0084	5.0414
		ERGAS	0.6817	0.7607	0.5794	0.5757	0.9901	1.0807	0.6442	0.7219	0.8590	0.6810	0.6512	**0.5648**
img1 (in Harvard)	8	PSNR	42.5584	43.8037	43.7421	43.7758	38.7044	36.1467	42.3843	39.8425	38.3692	44.1371	44.1757	**44.1874**
		SSIM	0.9733	0.9783	0.9775	0.9787	0.9811	0.9669	0.9909	0.9746	0.9602	0.9937	0.9820	**0.9946**
		SAM	2.8892	2.3205	2.4091	2.3446	4.4211	5.6776	2.4872	2.6060	4.9993	1.9774	**1.9686**	2.0439
		ERGAS	0.5866	0.5251	0.5059	0.5021	1.2482	1.1533	0.5185	0.8085	1.0315	0.4650	**0.4582**	0.5165
	16	PSNR	42.3327	43.2603	43.5474	43.3981	36.2908	32.2827	41.9525	39.5549	39.0854	43.7429	43.7558	**43.7818**
		SSIM	0.9732	0.9784	0.9774	0.9779	0.9697	0.9462	0.9905	0.9698	0.9644	**0.9935**	0.9818	0.9799
		SAM	3.0143	2.4327	2.4365	2.4549	5.9409	7.2683	2.6339	3.3126	4.0790	2.0885	**2.0762**	2.1435
		ERGAS	0.2953	0.2761	0.2593	0.2765	0.9204	1.7134	0.2686	0.5012	0.4246	0.2392	0.2383	**0.2165**
WDC	8	PSNR	38.9554	40.7166	41.2060	40.2992	39.1008	34.6670	37.9136	33.4236	36.9547	42.6309	42.9682	**43.5299**
		SSIM	0.9814	0.9836	0.9884	0.9825	0.9898	0.9454	0.9837	0.9754	0.9834	0.9914	0.9922	**0.9935**
		SAM	4.4154	3.4420	3.6815	3.7524	3.9122	9.8385	4.6560	6.1538	3.9249	2.9397	2.8119	**2.6241**
		ERGAS	0.9261	0.8063	0.7263	0.8354	0.8207	2.2276	1.0072	1.2988	1.0522	0.6374	0.6057	**0.5360**
	16	PSNR	38.2839	39.1586	40.4444	37.4912	35.4277	33.3222	35.7710	29.1583	34.8646	36.6861	38.9156	**42.4423**
		SSIM	0.9799	0.9814	0.9869	0.9744	0.9806	0.9366	0.9735	0.9516	0.9755	0.9795	0.9888	**0.9922**
		SAM	4.4933	4.2819	3.8795	4.6568	5.8583	11.4644	6.1513	9.8073	5.5464	4.7639	3.7111	**2.8559**
		ERGAS	0.5080	0.4947	0.3983	0.6108	0.6230	1.3741	0.6861	1.0641	0.6670	0.5957	0.4445	**0.3124**

E. Experimental Results on Simulated Data

The obtained simulation results with five simulation datasets are numerically listed in Table I, where the best results are given in boldface and the second best results are underlined for clarity. One can observe from this table, that CS2DIPs perform the best, MIAE* the second best, and both of them outperform the other 10 methods under test overall. However, the running time for obtaining the simulation results for the datasets CAVE and PaviaC are listed in Table II for the case of K equal to 8, indicating that the proposed CS2DIPs expends more running time than most of the methods under test, around 4.9 to 7.5 times running time of MIAE.

For visual quality assessment, Figs. 4 through 7 display some results obtained by all the methods under test for WDC, CAVE, Harvard, and PaviaC datasets, respectively. For each figure, the reconstructed HR-HSI for three spectral bands ([30,15,10] for CAVE, [30,12,8] for Harvard, [85,55,35] for PaviaC, KSC and WDC) are shown in the top two rows and the absolute image errors (scaled up by 10 for clarity) for the 8th band are shown in the bottom two rows, respectively. One can observe from these figures: i) DIP-2D and DIP-3D perform the worst (cf. Figs. 6 and 7), ii) CS2DIPs perform the best for $K = 8$ and $K = 16$, iii) the performances of DeepTensor, GDD, and uSDN are in the middle among all the unsupervised deep learning based methods; iv) all the 4 model-based methods also perform comparably with each other but worse than CS2DIPs for $K = 8$ and 16. Therefore, the results shown in Figs. 4 through 7 are also consistent with those listed in Table I.

Fig. 8 shows the simulation results in terms of PSNR versus spectral band for K equal to 8 (16) for the top (bottom) row, where the performances of only 6 tested methods (i.e., CSTF, NLSTF, DeepTensor, MIAE, MIAE*, CS2DIPs) are illustrated for clarity, while the other 6 methods are omitted due to their inferior overall performances. Again, one can also observe that CS2DIPs show superior overall performance, while MIAE comes behind as the runner-up.

Finally, let us conclude this subsection by showing some convergence results for a comparison of the proposed CS2DIPs with 4 widely used DIP based methods (DIP-2D, DIP-3D, DeepTensor, GDD). Fig. 9 shows some results (obtained on the KSC dataset) in terms of PSNR versus iteration number for K equal to 8. The proposed CS2DIPs method exhibits a faster convergence rate and better PSNR performance than all the other methods.

F. Ablation Studies

In this section, we perform ablation experiments on the proposed CS2DIPs using the PaviaC dataset. Without loss of

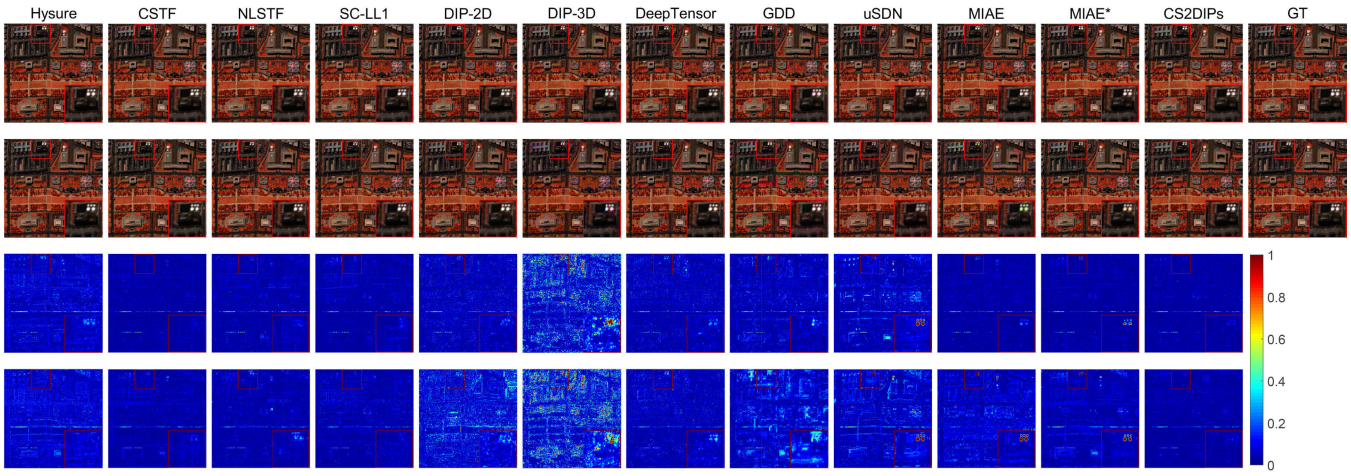

Fig. 4. Reconstructed images of various methods for the WDC dataset are illustrated by the false color image of [80,55,35] bands for $K = 8$ (the first row) and $K = 16$ (the second row), respectively, and the error images of the 8th band for $K = 8\,(16)$ are shown in the third (fourth) row.

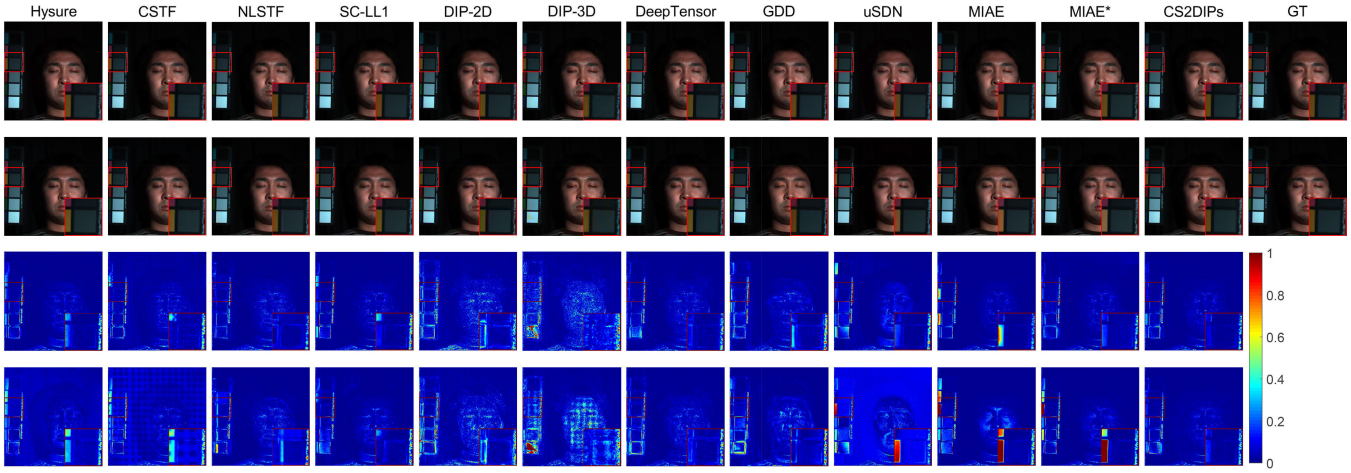

Fig. 5. Reconstructed images of various methods for the Face (CAVE) dataset are illustrated by the false color image of [30,15,10] bands for $K = 8$ (the first row) and $K = 16$ (the second row), respectively, and the error images of the 8th band for $K = 8\,(16)$ are shown in the third (fourth) row.

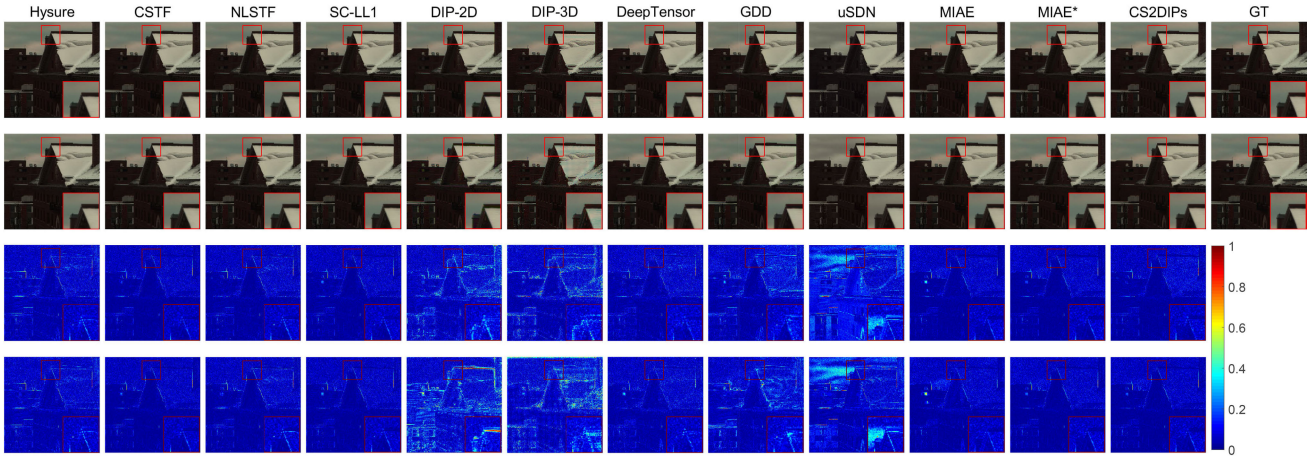

Fig. 6. Reconstructed images of various methods for the Img1 (Harvard) dataset are illustrated by the false color image of [30,12,8] bands for $K = 8$ (the first row) and $K = 16$ (the second row), respectively, and the error images of the 8th band for $K = 8\,(16)$ are shown in the third (fourth) row.

TABLE II

RUNNING TIME (IN SEC) OF ALL THE METHODS UNDER TEST EXPENDED ON DATASETS CAVE AND PAVIAC FOR THE CASE OF $K = 8$

	Hysure	CSTF	NLSTF	SC-LL1	DIP-2D	DIP-3D	DeepTensor	GDD	uSDN	MIAE	MIAE*	CS2DIPs
face_ms	23.85	20.53	9.84	70.46	403.37	1458.31	461.73	253.66	392.69	128.18	123.83	972.13
PaviaC	13.75	11.91	9.90	106.83	384.12	4442.27	482.28	256.64	810.25	136.84	127.54	762.03

generality, we fix $K = 8$ and use an 8×8 Gaussian blur kernel (zero mean and standard deviation equal to 3).

1) Effectiveness of CS2DIPs Components: The proposed CS2DIPs model consists of three components:

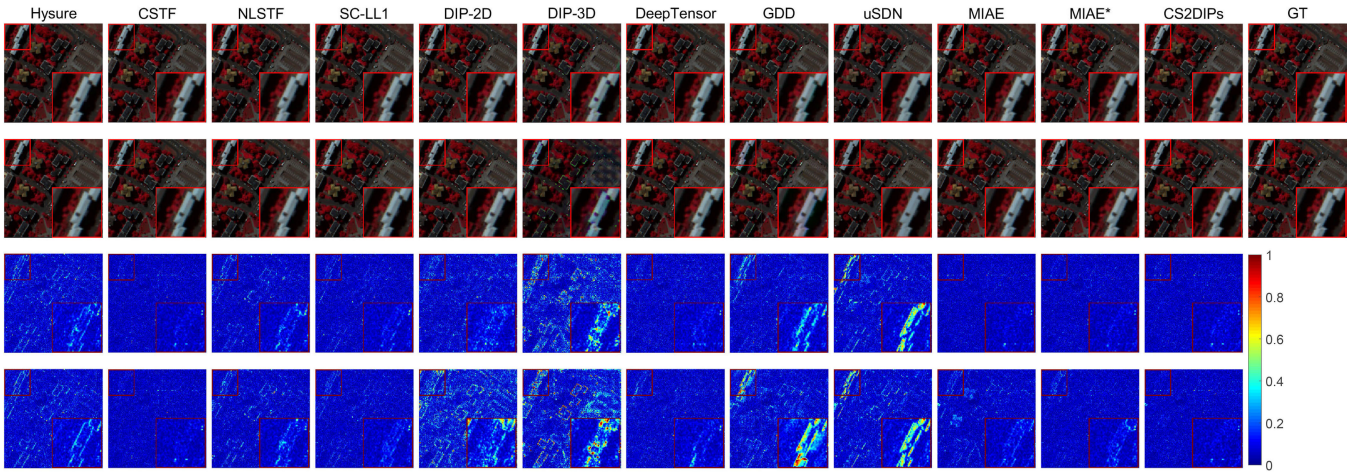

Fig. 7. Reconstructed images of various methods for the PaivaC dataset are illustrated by the false color image of [80,55,35] bands for $K = 8$ (the first row) and $K = 16$ (the second row), respectively, and the error images of the 8th band for $K = 8\,(16)$ are shown in the third (fourth) row.

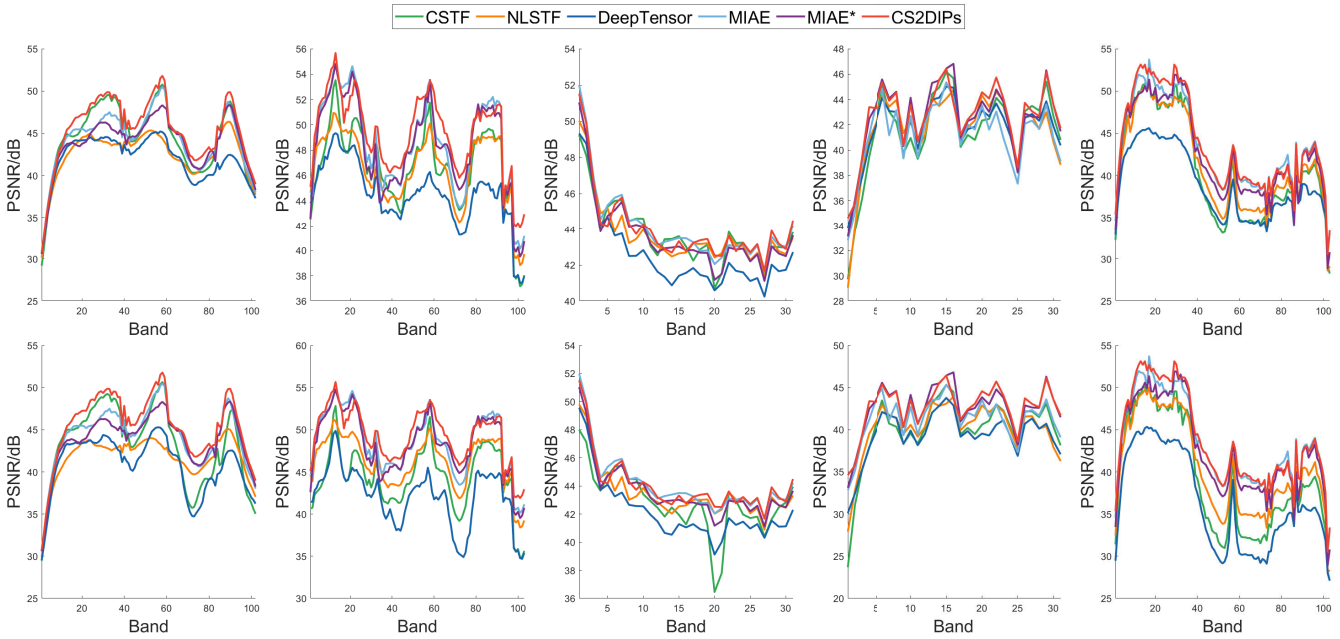

Fig. 8. Simulation results in terms of PSNR versus spectral band for K equal to 8 (top row) and 16 (bottom row) and the datasets PaviaC, KSC, CAVE, Harvard, and WDC (from left to right).

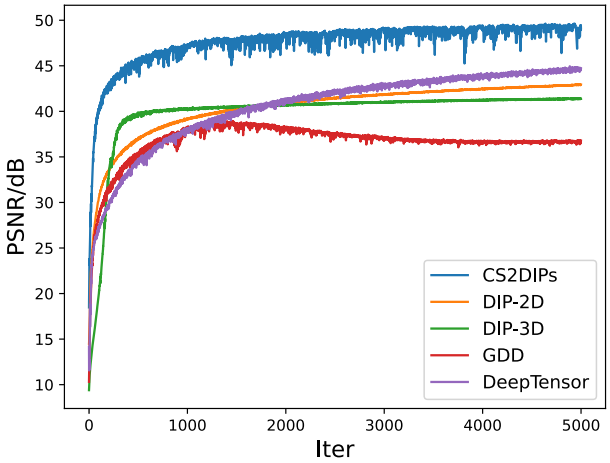

Fig. 9. Convergence of 5 DIP-based methods on the KSC dataset for $K = 8$.

1) decomposition based on NMVTF; 2) GDIP network structure; 3) spatial and spectral constraints based on physical properties. To demonstrate the importance of these three components, we designed three ablation experiments. The first replaces NMVTF with NMF, the second removes the GDIP and the third removes all equality and inequality constraints. These experiments are referred to as 'CS2DIPs w/ NMF', 'CS2DIPs w/o GDIP', and 'CS2DIPs w/o constraints'. The results are listed in Table III. The performance of the CS2DIPs version with GDIP removed is seriously degraded, with the PSNR drop of 2.83 dB, while the PSNR of the CS2DIPs with NMF replaced and constraints removed drops by 1.03 and 2.18 dB, respectively.

2) Effectiveness of GDIP Network Structure Components: The GDIP structure proposed in this paper consists of several modules. We designed four ablation experiments. The first removes the guided DCB, the second removes the GUS, the third removes the attention mechanism used in DAU and FAU, and the fourth removes the skip-connect branch. These experiments are referred to as 'CS2DIPs w/o GDC', 'CS2DIPs

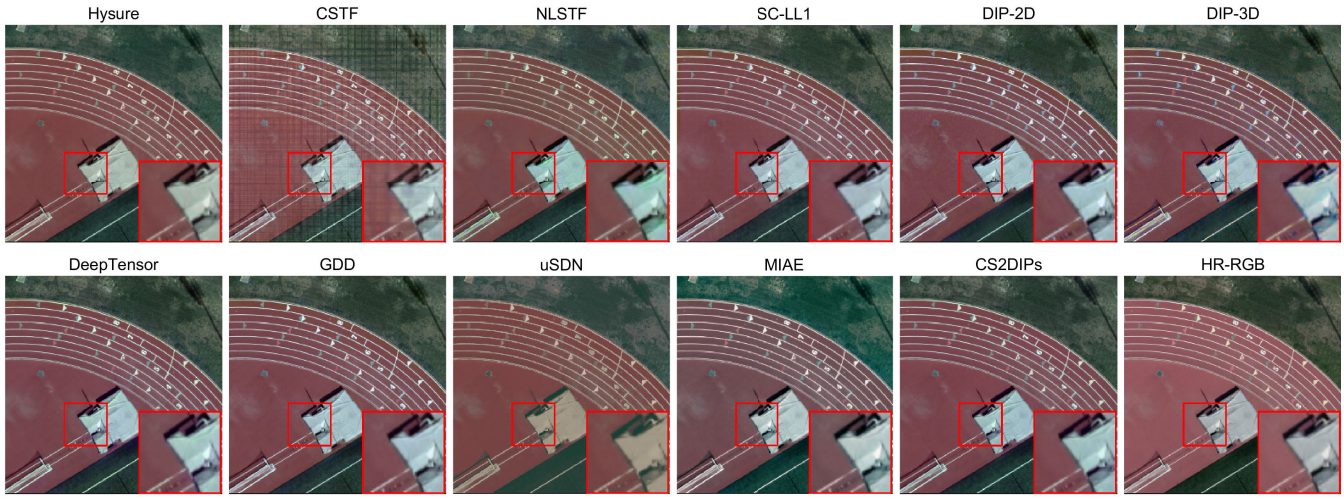

Fig. 10. Reconstructed HR-HSI images by all the methods under test on the real UH dataset (without GT $\mathcal{Z}$). The images shown are false colour images consisting of [17, 12, 10] bands with $K = 8$, except for the image HR-RGB (treated as the "GT" for visual assessment).

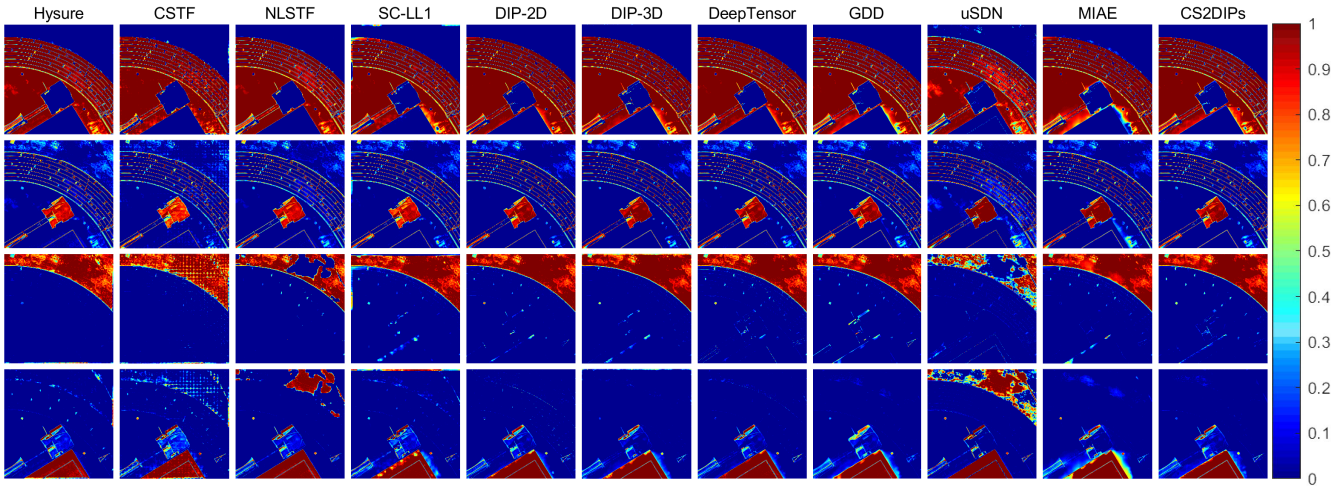

Fig. 11. Abundance maps of all the reconstructed HR-HSI images (for real UH dataset) shown in Fig. 10, consisting of 4 distinct material distributions (one for each row), including the brown-color material (first row), white-color material (second row), light green-color material (third row) and dark green-color material (fourth row) material, respectively.

TABLE III
CS2DIPs ABLATION EXPERIMENT METRICS ON THE PAVIAC DATASET

Methods	PSNR↑	SSIM↑	SAM↓	ERGAS↓
CS2DIPs	**45.4553**	**0.9870**	**3.6370**	**0.7302**
CS2DIPs w/ NMF	44.4243	0.9846	3.9591	0.8652
CS2DIPs w/o GDIP	42.6212	0.9805	4.2602	0.8698
CS2DIPs w/o constraints	43.2751	0.9810	4.5303	1.2904
CS2DIPs w/o GDC	44.7494	0.9866	3.7236	0.7739
CS2DIPs w/o GUS	32.9389	0.9054	6.5395	2.0245
CS2DIPs w/o attention	43.2376	0.9840	4.1614	0.8526
CS2DIPs w/o skip-connect	44.6695	0.9865	3.6992	0.7678

w/o GUS', 'CS2DIPs w/o attention' and 'CS2DIPs w/o skip-connect'. The results are also listed in Table III. One can observe that removing each of these components leads to some performance degradation; serious performance loss happens in the case of w/o GUS (e.g., PSNR loss of 12.51 dB), implying that the component GUS is most sensitive to the performance loss; and the component GDC is most insensitive to the performance loss.

G. Experimental Results on Real Data

We used the real UH dataset released by the 2018 IEEE GRSS Data Fusion Contest to evaluate the effectiveness of CS2DIPs. The resolution is set to $K = 8$. $\mathbf{P}$, $\mathbf{S}_1$ and $\mathbf{S}_2$ are estimated using the method in [64] since they are unknown for this real dataset. The experimental results are shown in Fig. 10, colorblackfrom which one can see that the proposed CS2DIPs method yields good visual-quality fusion results.

In order to further assess the quality of the HR-HSI recovered by all the methods under test, we used the SCM method [65] for spectral unmixing of the recovered HR-HSI. From the real image of the HR-RGB shown in Fig. 10, one can classify the image into four objects: (1) red plastic runway; (2) white runway scale, markers inside the soccer field, and rain shed; (3) light green grass outside the runway and light green triangular icons inside the runway; and (4) dark green grass inside the soccer field. We set the maximum number of iterations to 100 and the number of materials to 4 (i.e., $R = 4$) in the HSI under consideration, for all the algorithms under test. The 4 abundance maps of HR-HSI recovered are shown in Fig. 11. From this figure, one can see that the

abundance maps yielded by applying SCM to unmixing the HR-HSI data reconstructed by CSTF, NLSTF, SC-LL1, and uSDN have some bias and detail loss, and the abundance maps decomposed by Hysure, DeepTensor, and CS2DIPs have more accurate details.

V. Conclusion

We have presented the CS2DIPs for HSI-SR (shown in Fig. 2), that, without the need for pretraining, can effectively learn the abundance tensor and spectral signature matrix of the desired HR-HSI from the given HR-MSI and LR-HSI in a coupling-guided fashion with physical constraints of (non-negativity and sum-to-one) incorporated in the meantime. The proposed CS2DIPs is also an unsupervised DIP-based method by minimizing the differentiable convex HLF (cf. (18)), which can effectively exploit intrinsic statistical spatial-spectral correlations, and various prior characteristics embedded in HR-MSI and LR-HSI. To the best of our knowledge, it is applied to DIP-based HSI-SR for the first time, in addition to its recent application to DIP-based HSI denoising and inpainting [59], [66]. Extensive simulated experiments and real-data experiments have been provided to demonstrate the CS2DIPs' superior overall performance over state-of-the-art methods.

References

[1] R. O. Green et al., "Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS)," *Remote Sens. Environ.*, vol. 65, no. 3, pp. 227–248, Sep. 1998.

[2] Y. Xie, Y. Qu, D. Tao, W. Wu, Q. Yuan, and W. Zhang, "Hyperspectral image restoration via iteratively regularized weighted schatten p-norm minimization," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4642–4659, Aug. 2016.

[3] S. Prasad and L. M. Bruce, "Limitations of principal components analysis for hyperspectral target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 4, pp. 625–629, Oct. 2008.

[4] H. Ren and C.-I. Chang, "Automatic spectral target recognition in hyperspectral imagery," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 39, no. 4, pp. 1232–1249, Oct. 2003.

[5] S. Chabrillat, A. F. H. Goetz, L. Krosley, and H. W. Olsen, "Use of hyperspectral images in the identification and mapping of expansive clay soils and the role of spatial resolution," *Remote Sens. Environ.*, vol. 82, nos. 2–3, pp. 431–445, Oct. 2002.

[6] Z. Ting-ting and L. Fei, "Application of hyperspectral remote sensing in mineral identification and mapping," in *Proc. 2nd Int. Conf. Comput. Sci. Netw. Technol.*, Dec. 2012, pp. 103–106.

[7] D. Moshou et al., "Plant disease detection based on data fusion of hyperspectral and multi-spectral fluorescence imaging using Kohonen maps," *Real-Time Imag.*, vol. 11, no. 2, pp. 75–83, Apr. 2005.

[8] K. Golhani, S. K. Balasundram, G. Vadamalai, and B. Pradhan, "A review of neural networks in plant disease detection using hyperspectral data," *Inf. Process. Agricult.*, vol. 5, no. 3, pp. 354–371, Sep. 2018.

[9] R. Dian, L. Fang, and S. Li, "Hyperspectral image super-resolution via non-local sparse tensor factorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5344–5353.

[10] Q. Wei, J. Bioucas-Dias, N. Dobigeon, J. Tourneret, M. Chen, and S. Godsill, "Multiband image fusion based on spectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7236–7249, Dec. 2016.

[11] C. Lanaras, E. Baltsavias, and K. Schindler, "Hyperspectral super-resolution by coupled spectral unmixing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3586–3594.

[12] S. He, H. Zhou, Y. Wang, W. Cao, and Z. Han, "Super-resolution reconstruction of hyperspectral images via low rank tensor modeling and total variation regularization," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 6962–6965.

[13] Y. Wang, X. Chen, Z. Han, and S. He, "Hyperspectral image super-resolution via nonlocal low-rank tensor approximation and total variation regularization," *Remote Sens.*, vol. 9, no. 12, p. 1286, Dec. 2017.

[14] Q. Yuan, L. Zhang, and H. Shen, "Regional spatially adaptive total variation super-resolution with spatial information filtering and clustering," *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2327–2342, Jun. 2013.

[15] N. Akhtar, F. Shafait, and A. Mian, "Sparse spatio-spectral representation for hyperspectral image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2014, pp. 63–78.

[16] N. Akhtar, F. Shafait, and A. Mian, "Bayesian sparse representation for hyperspectral image super resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3631–3640.

[17] Y. Zhao, J. Yang, Q. Zhang, L. Song, Y. Cheng, and Q. Pan, "Hyperspectral imagery super-resolution by sparse representation and spectral regularization," *EURASIP J. Adv. Signal Process.*, vol. 2011, no. 1, pp. 1–10, Dec. 2011.

[18] J. Xue, Y.-Q. Zhao, Y. Bu, W. Liao, J. C.-W. Chan, and W. Philips, "Spatial-spectral structured sparse low-rank representation for hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 30, pp. 3084–3097, 2021.

[19] Y. Xu, Z. Wu, J. Chanussot, P. Comon, and Z. Wei, "Nonlocal coupled tensor CP decomposition for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 348–362, Mar. 2019.

[20] W. He, Y. Chen, N. Yokoya, C. Li, and Q. Zhao, "Hyperspectral super-resolution via coupled tensor ring factorization," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108280.

[21] H. Liu, W. Jiang, Y. Zha, and Z. Wei, "Coupled tensor block term decomposition with superpixel-based graph Laplacian regularization for hyperspectral super-resolution," *Remote Sens.*, vol. 14, no. 18, p. 4520, Sep. 2022.

[22] X. Han, B. Shi, and Y. Zheng, "Self-similarity constrained sparse representation for hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5625–5637, Nov. 2018.

[23] J. Liu, Z. Wu, L. Xiao, J. Sun, and H. Yan, "A truncated matrix decomposition for hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 29, pp. 8028–8042, 2020.

[24] Y. Fu, T. Zhang, Y. Zheng, D. Zhang, and H. Huang, "Hyperspectral image super-resolution with optimized RGB guidance," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11661–11670.

[25] C. Wang, Y. Liu, X. Bai, W. Tang, P. Lei, and J. Zhou, "Deep residual convolutional neural network for hyperspectral image super-resolution," in *Image and Graphics*. Shanghai, China: Springer, Sep. 2017, pp. 370–380.

[26] X.-H. Han, B. Shi, and Y. Zheng, "SSF-CNN: Spatial and spectral fusion with CNN for hyperspectral image super-resolution," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 2506–2510.

[27] J.-F. Hu, T.-Z. Huang, L.-J. Deng, H.-X. Dou, D. Hong, and G. Vivone, "Fusformer: A transformer-based fusion network for hyperspectral image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[28] Z. Lai, K. Wei, and Y. Fu, "Deep plug-and-play prior for hyperspectral image restoration," *Neurocomputing*, vol. 481, pp. 281–293, Apr. 2022.

[29] W. Dong, C. Zhou, F. Wu, J. Wu, G. Shi, and X. Li, "Model-guided deep hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 30, pp. 5754–5768, 2021.

[30] Q. Ma, J. Jiang, X. Liu, and J. Ma, "Learning a 3D-CNN and transformer prior for hyperspectral image super-resolution," *Inf. Fusion*, vol. 100, Dec. 2023, Art. no. 101907.

[31] Y. Qu, H. Qi, and C. Kwan, "Unsupervised sparse Dirichlet-Net for hyperspectral image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2511–2520.

[32] J. Yao, D. Hong, J. Chanussot, D. Meng, X. Zhu, and Z. Xu, "Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, Aug. 2020, pp. 208–224.

[33] Y. Qu, H. Qi, C. Kwan, N. Yokoya, and J. Chanussot, "Unsupervised and unregistered hyperspectral image super-resolution with mutual Dirichlet-Net," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022.

[34] J. Liu, Z. Wu, L. Xiao, and X.-J. Wu, "Model inspired autoencoder for unsupervised hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.

[35] Z. Wang, B. Chen, R. Lu, H. Zhang, H. Liu, and P. K. Varshney, "FusionNet: An unsupervised convolutional variational network for hyperspectral and multispectral image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 7565–7577, 2020.

[36] V. Lempitsky, A. Vedaldi, and D. Ulyanov, "Deep image prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9446–9454.

[37] 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.

[38] J. Liu, Y. Sun, X. Xu, and U. S. Kamilov, "Image restoration using total variation regularized deep image prior," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 7715–7719.

[39] V. Saragadam, R. Balestriero, A. Veeraraghavan, and R. G. Baraniuk, "DeepTensor: Low-rank tensor decomposition with deep network priors," 2022, *arXiv:2204.03145*.

[40] Y.-S. Luo, X.-L. Zhao, T.-X. Jiang, Y.-B. Zheng, and Y. Chang, "Hyperspectral mixed noise removal via spatial–spectral constrained unsupervised deep image prior," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 9435–9449, 2021.

[41] 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.

[42] N. Rahaman et al., "On the spectral bias of neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5301–5310.

[43] Z.-Q. John Xu, Y. Zhang, T. Luo, Y. Xiao, and Z. Ma, "Frequency principle: Fourier analysis sheds light on deep neural networks," 2019, *arXiv:1901.06523*.

[44] 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, Sep. 2020.

[45] Y. Qian, F. Xiong, S. Zeng, J. Zhou, and Y. Y. Tang, "Matrix-vector nonnegative tensor factorization for blind unmixing of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 3, pp. 1776–1792, Mar. 2017.

[46] Z. Long, C. Zhu, J. Liu, and Y. Liu, "Bayesian low rank tensor ring for image recovery," *IEEE Trans. Image Process.*, vol. 30, pp. 3568–3580, 2021.

[47] Z. Long, C. Zhu, J. Liu, P. Comon, and Y. Liu, "Trainable subspaces for low rank tensor completion: Model and analysis," *IEEE Trans. Signal Process.*, vol. 70, pp. 2502–2517, 2022.

[48] Y.-Q. Zhao and J. Yang, "Hyperspectral image denoising via sparse representation and low-rank constraint," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 296–308, Jan. 2014.

[49] Y. Xu, Z. Wu, J. Chanussot, and Z. Wei, "Hyperspectral images super-resolution via learning high-order coupled tensor ring representation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4747–4760, Nov. 2020.

[50] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 2, pp. 528–537, Feb. 2011.

[51] C.-H. Lin, F. Ma, C.-Y. Chi, and C.-H. Hsieh, "A convex optimization-based coupled nonnegative matrix factorization algorithm for hyperspectral and multispectral data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1652–1667, Mar. 2018.

[52] L. Zhang, J. Nie, W. Wei, Y. Li, and Y. Zhang, "Deep blind hyperspectral image super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2388–2400, Jun. 2020.

[53] L. Zhang, J. Nie, W. Wei, Y. Zhang, S. Liao, and L. Shao, "Unsupervised adaptation learning for hyperspectral imagery super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3073–3082.

[54] R. Heckel and P. Hand, "Deep decoder: Concise image representations from untrained non-convolutional networks," 2018, *arXiv:1810.03982*.

[55] T. Uezato, D. Hong, N. Yokoya, and W. He, "Guided deep decoder: Unsupervised image pair fusion," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 87–102.

[56] R. Wang et al., "DCN v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems," in *Proc. Web Conf.*, Apr. 2021, pp. 1785–1797.

[57] H. Wu et al., "Contrastive learning for compact single image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 10551–10560.

[58] P. J. Huber, *Robust Statistics*, vol. 523. Hoboken, NJ, USA: Wiley, 2004.

[59] K. F. Niresi and C.-Y. Chi, "Unsupervised hyperspectral denoising based on deep image prior and least favorable distribution," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5967–5983, 2022.

[60] M. Simoes, J. Bioucas-Dias, L. B. Almeida, and J. Chanussot, "A convex formulation for hyperspectral image superresolution via subspace-based regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3373–3388, Jun. 2015.

[61] S. Li, R. Dian, L. Fang, and J. M. Bioucas-Dias, "Fusing hyperspectral and multispectral images via coupled sparse tensor factorization," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4118–4130, Aug. 2018.

[62] R. Dian, S. Li, L. Fang, T. Lu, and J. M. Bioucas-Dias, "Nonlocal sparse tensor factorization for semiblind hyperspectral and multispectral image fusion," *IEEE Trans. Cybern.*, vol. 50, no. 10, pp. 4469–4480, Oct. 2020.

[63] M. Ding, X. Fu, T.-Z. Huang, J. Wang, and X.-L. Zhao, "Hyperspectral super-resolution via interpretable block-term tensor modeling," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 3, pp. 641–656, Apr. 2021.

[64] R. Dian, A. Guo, and S. Li, "Zero-shot hyperspectral sharpening," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12650–12666, Jun. 2023.

[65] Y. Zhou, A. Rangarajan, and P. D. Gader, "A spatial compositional model for linear unmixing and endmember uncertainty estimation," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5987–6002, Dec. 2016.

[66] K. F. Niresi and C.-Y. Chi, "Robust hyperspectral inpainting via low-rank regularized untrained convolutional neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.