Dynamic Spectral Guided Spatial Sparse Transformer for Hyperspectral Image Reconstruction

Junyang Wang, Xiang Yan, *Member, IEEE*, Hanlin Qin, *Member, IEEE*, Naveed Akhtar, *Member, IEEE*, Shuowen Yang, *Member, IEEE*, and Ajmal Mian, *Senior Member, IEEE*

***Abstract*—Hyperspectral image (HSI) reconstruction plays a crucial role in compressive spectral imaging with coded aperture snapshot spectrometry. Although HSI reconstruction has attracted much attention in recent years, it remains a challenging problem. Existing deep learning-based methods leverage all the spectral information to reconstruct the HSI images without considering the spectral redundancy of HSI images, leading to high computational costs. In this article, we present an efficient method named dynamic spectral guided spatial sparse transformer (DGST). Specifically, DGST consists of three core modules as follows. 1) spectral sparse multihead self-attention hybrid spatial feature enhancement (SSHE) module, which employs a top-*k* spectral sparsity method to filter noise and redundant spectral information while extracting spectral information from HSI. 2) Spatial information compensation module, which utilizes a multiscale approach to extract spatial information and compensates for the spatial information neglected by SSHE. 3) Mask-guided spatial sparse multihead self-attention hybrid spectral enhancement module, which dynamically generates masks to guide the filtering of irrelevant regions, reducing computational costs while focusing on spatial information reconstruction. Our DGST improves the quality of HSI reconstruction by integrating spatial–spectral details and global information. Extensive experiments on public HSI reconstruction benchmark datasets demonstrate that our approach achieves state-of-the-art performance in end-to-end hyperspectral reconstruction. The superior performance of the proposed DGST is showcased on real and simulated hyperspectral imaging datasets.**

***Index Terms*—Compressed imaging, dual attention, hyperspectral image (HSI) reconstruction, sparse transformer.**

Received 5 June 2024; revised 23 July 2024 and 9 August 2024; accepted 19 August 2024. Date of publication 29 August 2024; date of current version 10 September 2024. This work was supported in part by 111 Project under Grant B17035, in part by Shaanxi Province Key Research and Development Plan Project under Grant 2022JBGS2-09, in part by Shaanxi Province Science and Technology Plan Project under Grant 2023KXJ-170, in part by Xian City Science and Technology Plan Project under Grant 21JBGSZ-QCY9-0004, Grant 22JBGS-QCY4-0006, and Grant 23GBGS0001, and in part by the Aeronautical Science Foundation of China under Grant 20230024081027. *(Junyang Wang and Xiang Yan contributed equally to this work.) (Corresponding authors: Xiang Yan; Hanlin Qin.)*

Junyang Wang, Xiang Yan, Hanlin Qin, and Shuowen Yang are with the School of Optoelectronic Engineering, Xidian University, Xi'an 710071, China (e-mail: xyan@xidian.edu.cn; hlqin@mail.xidian.edu.cn).

Naveed Akhtar is with the School of Computing and Information Systems, Faculty of Engineering and IT, The University of Melbourne, Parkville, VIC 3052, Australia (e-mail: naveed.akhtar1@unimelb.edu.au).

Ajmal Mian is with the Department of Computer Science and Software Engineering, The University of Western Australia, Crawley, WA 6009, Australia (e-mail: ajmal.mian@uwa.edu.au).

The source code is released at https://github.com/WangJunYang2000/DGST.

Digital Object Identifier 10.1109/JSTARS.2024.3447729

I. Introduction

HYPERSPECTRAL imaging uses multichannel technology, where each channel can obtain the information at a specific spectral wavelength for an actual scene. Compared to regular RGB images, hyperspectral images (HSIs) contain richer information and reveal more details of the scene. Given this inherent superiority, HSIs are employed for many tasks, such as image classification [1], [2], [3], [4], [5], target detection [6], [7], [8], [9], target tracking [10], [11], remote sensing [12], [13], [14], [15], and medical image understanding [16], [17], [18]. However, capturing HSIs efficiently and effectively is a challenging problem.

At the advent, the HSI systems leveraged a spectrometer to scan along the spatial dimension and acquire the desired spectral information in an extended time. Such methods are particularly undesirable for dynamic scenes. As a remedy, researchers have proposed snapshot compressive imaging (SCI) systems to obtain the desired HSIs. It utilizes the principle of compressed sensing to map 3-D HSI into a single 2-D measurement [19], [20] and capture the HSI through snapshots in a single integration cycle. Numerous recent SCI systems are built of this new imaging paradigm [21], [22], [23], [24]. The coded aperture snapshot spectral imaging (CASSI) system [25] is widely used for SCI. CASSI generates a 2-D compressed measurement by modulating HSI signals at different wavelengths with coded aperture and dispersive elements. The measurement is then leveraged to recover its corresponding 3-D hyperspectral cube, i.e., the HSI. Efficiently recovering the 3-D HSI cube from a 2-D measurement is a key problem in this paradigm. It is generally formulated as an ill-posed inverse problem of image reconstruction.

To solve the above-mentioned problem, numerous 3-D HSI reconstruction approaches have been proposed. Earlier methods leverage handcrafted features and make assumptions, such as sparsity [26], [27], total variation [28], low rank [29], [30], and nonlocal similarity [31], [32]. However, these approaches rely on manual tweaking of key hyperparameters in the underlying models, leading to poor generalization. As with many other HSI processing tasks, its reconstruction has witnessed a huge success recently following deep learning approaches [33], [34], [35], [36], [37], [38]. The deep convolutional neural network (CNN) was first introduced to HSI reconstruction in Xiong et al.'s [39] work. Subsequently, a series of CNN-based approaches surfaced [40], [41], [42], [43]. However, due to inductive bias, CNN-based methods generally face limitations in modeling

the interspectra similarities and lack in leveraging long-range dependencies in the data. Moreover, CNN-based techniques generally also suffer from an irreversible spectral information loss in the feature extraction phase, which compromises the final reconstruction quality.

More recently, transformer architectures [44], [45], [46], [47] have become popular in computer vision. Due to their ability to exploit global dependencies between image regions and capture nonlocal similarities, transformers have achieved excellent results for many tasks in computer vision, e.g., object detection [48], semantic segmentation [49], and image restoration [50]. Inspired by their success in vision tasks, transformer-based methods have also been developed for HSI tasks, e.g., HSI denoising [51], [52], superresolution [53], [54], [55], [56], and classification [2], [57]. Transformers saw their early application in HSI reconstruction in Cai et al.'s [58] work. Subsequently, further efforts to tailor the technique for improved reconstruction quality appeared in [59] and [60]. To obtain the long-range interspectra similarity and dependencies, these methods design spectralwise multihead self-attention (S-MSA) or spectra-aware screening mechanism (SASM). The former regards each spectral feature as a token and computes multihead self-attention (MSA) along the spectral dimension. The latter divides the image into nonoverlapping patches and selects the ones containing information that can characterize HSI, to compute its corresponding MSA efficiently. To reduce computational cost, these methods apply spectral dimension self-attention or leverage spatial sparsity in the learning model.

Although the above-mentioned methods improve reconstruction to some extent, they are unable to fully leverage both spatial and spectral information simultaneously. Addressing that, S^2-transformer-based HSI reconstruction is proposed in Wang et al.'s [61] work, which performs spectral and spatial attention modeling to disentangle the blended information in a 2-D measurement. A customized deep unfolding transformer framework for HSI reconstruction is also proposed in [23], [62], and [63]. This framework breaks HSI reconstruction into a data subproblem and a prior subproblem. It converts traditional iterative optimization algorithms into a sequence of deep neural network blocks, addressing the two subproblems iteratively.

In the unfolding CASSI reconstruction [23], [62], [63], denoising networks are embedded in each stage of the unfolding network to optimize the reconstruction model step by step. In contrast, spectral and spatial collaborative attention transformer reconstruction framework [61] needs only one optimization step for the reconstruction. In this article, we also prefer this approach, which makes S^2-transformer-based reconstruction in Wang et al.'s [61] work highly relevant to our work. However, that approach faces two major challenges. 1) The use of raw spectra for spectral attention leads to an enormous computational load and lacks noise suppression. 2) S^2-transformer's utilization of spatial–spectral self-attention does not leverage the spatial sparsity of HSI data, which makes the model heavy in the number of parameters and computational load.

We propose a dynamic spectral guided spatial sparse transformer (DGST) framework for HSI reconstruction that addresses the above-mentioned challenges. Our DGST employs the spectral sparse multihead self-attention hybrid spatial feature enhancement (SSHE) module, which utilizes spectral sparse multihead self-attention (SS-MSA) to extract global spectral information. Simultaneously, a spatial feature enhancement (SPA-FE) module is integrated within SS-MSA to inject spatial information. Moreover, we design a spatial information compensation (SIC) module to compensate for the high-frequency spatial features of HSI. The features obtained by SSHE and SIC are fused to get coarse-grained features that generate dynamically varying masks, which guide filtering of irrelevant regions to reduce computational overhead. Eventually, the dynamic masks and the deep features obtained by the fusion of SSHE and SIC are input into a mask-guided spatially sparse multihead self-attention hybrid spectral feature enhancement (MSSHE) module that extracts high-level features from HSI that lead to extensive comprehensive features for favorable reconstruction performance. In the following, we summarize the specific contributions of our work.

1) We propose DGST for HSI reconstruction, which takes the advantage of the similarity and sparsity of HSI for image reconstruction by employing SSHE and MSSHE, enabling the reconstruction model to overcome noise and redundant information interference while reducing the computational and memory costs to facilitate spectral and spatial reconstructed capabilities.
2) We introduce the notion of SSHE that reduces reconstruction computations as well as suppresses noise. We also combine SSHE with an SIC module to extract high-frequency spatial information to reconstruct the spatial texture details of HSI.
3) We propose MSSHE that utilizes masks to guide spatial sparse MSA combined with spectral enhancement to improve reconstruction quality with low computation.

II. Related Work

Deep learning has profoundly accelerated the advancement of HSI reconstruction methods. In this instance, we primarily delineate the related deep HSI reconstruction models. Initially, we provide a brief overview of traditional model-based methods, followed by a discussion on CNN-based methods and transformer-based HSI reconstruction approaches.

A. CNN-Based HSI Reconstruction Methods

CNNs have been successfully applied to various low-level hyperspectral visual tasks, such as HSI denoising [64], [65], HSI resotoration [36], and HSI superresolution [54], [66]. Their success in these tasks also encouraged researchers to apply CNN architectures to HSI reconstruction. A large number of CNN-based techniques have been developed to learn mapping functions of hyperpectral image reconstruction [23], [41], [67], [68], [69], [70]. In these methods, λ-Net [67] developed a dual-stage model to reconstruct the desired HSI image using a hierarchical channel reconstruction to progressively reconstruct spectral channels leveraging the features extracted by the neural

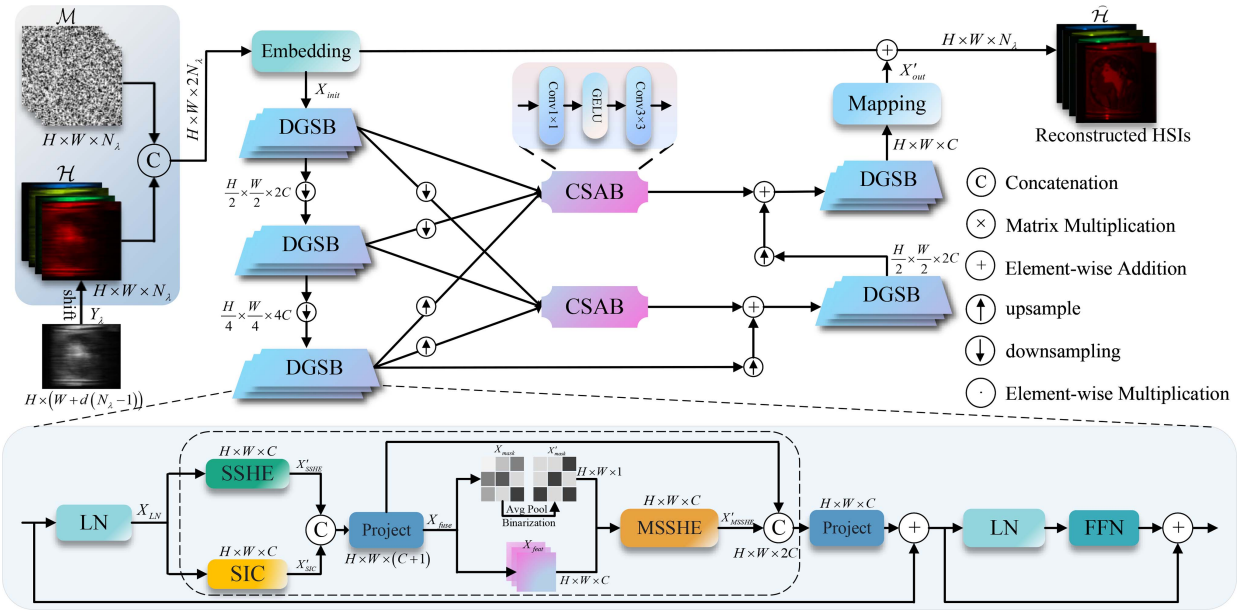

Fig. 1. Overview of DGST framework containing a three-level U-Net structure. The DGSB module consists of the SSHE module, SIC module, and MSSHE modules, along with an FFN. The DGSB dynamically generates an adaptive mask-guided MSSHE that utilizes MSS-MSA for spatial domain modeling. Subsequently, the spatial domain information extracted by MSSHE is fused with the spectral information and spatial context information extracted by SSHE and SIC, respectively. Subsequently, the fused features of MSSHE, SSHE, and SIC are fed into an FFN for further processing. Finally, the output of the last DGSB is processed through a mapping function and fused with the initial features to generate the reconstructed HSI.

network and previously reconstructed channels. Wang et al. [71] introduced a deep nonlocal unrolling HSI reconstruction approach that utilizes data-driven prior to adaptively exploit the local and nonlocal correlations in the spectral image. Similarly, Wang et al. [41] proposed a deep spatial–spectral prior-based HSI reconstruction method that replaces the conventional handcrafted prior with a data-driven alternative. TSA-Net [68] introduces spatial–spectral self-attention to sequentially reconstruct HSI. DGSMP [72] is a deep Gaussian scale mixture (GSM) prior for HSI reconstruction that learns the scale prior and estimates the local means of the GSM models by deep CNNs. HDNet [69] proposes a high-resolution dual-domain learning deep network for HSI reconstruction, which uses a high-resolution spatial–spectral attention module with feature fusion that provides fine continuous features. Meng et al. [73] presented a self-supervised CNN for spectral compressive imaging. They developed a new framework by integrating deep image priors into a plug-and-play regime to get the HSI reconstruction images. CNN-based methods demonstrate impressive performance, yet they exhibit limitations in capturing nonlocal similarities and correlations between spectra.

B. Visual Transformer-Based HSI Reconstruction Methods

MST [58] and CST [60] were the first methods to employ transformers for HSI reconstruction while capturing the inter-spectra similarity and dependencies. CST [60] embeds spatial sparsity into the transformer structure to reduce reconstruction model parameters and computational complexity while improving reconstruction performance. Subsequently, the degradation-aware unfolding half-shuffle transformer (HST) [62] was designed for HSI reconstruction that simultaneously gets local contents and nonlocal dependencies. The HST was then plugged into a degradation-aware unfolding HSI reconstruction framework to improve the reconstruction performance. Wang et al. [61] proposed a spatial–spectral transformer with a masking-aware learning strategy for HSI reconstruction. In this method, the authors simultaneously leveraged spatial and spectral attention modeling to disentangle the blended information in a 2-D measurement along spectral and spatial dimensions. Liu et al. [63] proposed a pixel adaptive deep unfolding transformer of HSI reconstruction that leveraged pixel-level adaptive and nonlocal spectral transformer to recover the pixel and spectra for HSI. More recently, with the remarkable success of the diffusion model in high-fidelity image synthesis, the diffusion model was also introduced in spectral compressive imaging in Wu et al.'s [74] work. It applies a latent diffusion prior to generating degradation-free prior to improve the regression ability of the deep unfolding HSI reconstruction framework.

C. Sparse Representation

The computational complexity of transformers scales quadratically with the spatial dimensions that increase the computational complexity. To do this, various sparse transformers were proposed for image classification, image restoration, image superresolution, and image retraining [75], [76], [77], [78]. For example, BiFormer [75] introduces dynamic sparse attention to obtain flexible content-aware computation allocation by dual-layer routing. In [76] and [78], they design a sparse attention module that can make the sparse regions interact with each other. It will greatly enhance the representation ability

of transformers. SGSFormer [77] utilizes sparse self-attention to filter out redundant information and noise, enabling the model to focus on the features more relevant to the degraded regions. DynaST [79] leverages dynamic attention units to cover changes in the optimal token count. DRSformer [78] proposes an adaptable top-k selection operator to selectively retain the most important attention scores from each query key for better feature aggregation.

In the field of HSI reconstruction, CST [60] embedded HSI sparsity into deep learning for HSI reconstruction, where is the first study to apply the spare transformer to HSI reconstruction. They proposed an SASM for coarse-grained selection. The encoding mask used in CST [60] is generated by SASM that is fixed and lacks adaptability. By contrast, we employ a dynamic mask generation strategy that allows the encoding mask to evolve with the increasing depth of the model. Moreover, inspired by DRSformer [78], we introduce for the first time the incorporation of spectral sparsity into deep learning for HSI reconstruction.

III. Method

Our objective is to design an efficient transformer model for HSI reconstruction from 2-D compressed measurements. Considering a spectral image $\mathbf{X}_\lambda \in \mathbb{R}^{H\times W\times N_\lambda}$, where N_λ represents the total number of channels, the captured HSI from real scenes is first modulated by a coded aperture $\mathbf{C}_\lambda \in \mathbb{R}^{H\times W\times N_\lambda}$. Here, H and W represent the height and width of the HSI, respectively. The 2-D temporary measurement $\mathbf{Y}_\lambda$ can be represented as $\mathbf{Y}_\lambda = \sum_{n_\lambda=1}^{N_\lambda} \text{shear}(\mathbf{C}_\lambda \odot \mathbf{X}_\lambda) + \mathbf{G}$. Here, $\odot$ denotes elementwise multiplication, $\text{shear}(\cdot)$ is a shearing operation along the y-axis, and $\mathbf{G} \in \mathbb{R}^{H\times(W+d(N_\lambda-1))}$ signifies the measurement noise arising from the process.

The aim of HSI reconstruction is to utilize the $\mathbf{Y}_\lambda$ to reconstruct the desired high fidelity HSI image $\hat{\mathbf{X}}_\lambda$ that makes it as close to the real HSI image as possible. In general, this objective is translated into an ill-posed inverse problem. As a solution, we design a DGST. This structure does not rely on physical degradation mechanisms but requires sufficient sample data for learning an end-to-end reconstruction model. The overall framework of the proposed DGST is shown in Fig. 1. The DGST framework is an end-to-end reconstruction paradigm that consists of three core modules as follows.

1) SSHE for extracting global spectral information from HSIs.
2) SIC module for compensating high-frequency spatial information during the reconstruction process.
3) MSSHE module for capturing spatial details of HSIs.

We leverage these three modules to enhance the performance of HSI reconstruction and reduce the computational complexity.

A. Overall Pipeline

Given the initial 2-D measurements $\mathbf{Y}_\lambda$, we first shift it along the horizontal direction according to the dispersive function with a stride k to obtain a 3-D tensor $\mathcal{H} \in \mathbb{R}^{H\times W\times N_\lambda}$. Subsequently, $\mathcal{H}$ is concatenated with a 3-D mask $\mathcal{M} \in \mathbb{R}^{H\times W\times N_\lambda}$ to generate a new 3-D tensor $\mathbf{X} \in \mathbb{R}^{H\times W\times 2N_\lambda}$, which serves as the input of the DGST network. Then, the spectral dimension of $\mathbf{X}$ is halved by pointwise convolutions to generate the low-level shallow feature embedding $\mathbf{X}_{\text{init}} \in \mathbb{R}^{H\times W\times N_\lambda}$. To improve the training speed and accuracy of our proposed HSI reconstruction model, we make the $\mathbf{X}_{\text{init}}$ pass through layer normalization to obtain new feature $\mathbf{X}_{\text{LN}}$.

In this study, we adopted a three-layer U-Net as the primary framework for our DGST to enhance reconstruction performance since it can well capture detailed information in HSIs. As shown in Fig. 1, our proposed HSI reconstruction framework consisted of several dynamic spectral guided spatial sparse blocks (DGSBs). The U-Net encoder and decoder at each level contain multiple DGSB blocks. With the layers of the U-Net framework increasing from top to bottom, the number of DGSB blocks gradually increases to boost the capability of feature extraction, which will help extract abundant deep features of the HSI. Especially, the obtained shallow features $\mathbf{X}_{\text{init}}$ are sent to the encoder layers of the U-Net structure to encode the shallow information of the HSI. The spatial resolution decreases while the spectral channel dimension doubles with the network layers increasing. The decoder of this U-Net structure takes the low-resolution deep feature $\mathbf{X}_{LR}^{\frac{H}{4}\times\frac{W}{4}\times 4C}$ as input and gradually restores the HSI with increasing of its network layers.

In this symmetric structure, we use pixel shuffle and convolution to achieve upsampling and downsampling, respectively. In this case, the spatial details of the HSI could be lost during the encoding and decoding processes once the classical U-Net structure is adopted. Therefore, we introduce a cross-scale feature aggregation block (CSAB) to aggregate information from multiple scales to reduce semantic information loss and preserve structure and texture details in the reconstructed image. Using our proposed structure, we obtain the decoded feature $\mathbf{X}_{\text{out}}$. The decoder features $\mathbf{X}_{\text{out}}$ are a series of feature maps containing various image feature information, which cannot directly generate the final reconstructed HSI image. This is because the encoder encoded the shallow features and the corresponding decoder gets new high-level features that must be mapped into an HSI image.

To obtain the reconstructed HSI, we need to transform these feature maps through a mapping layer to obtain the reconstructed result $\mathbf{X}'_{\text{out}}$. Finally, we fuse the low-level feature $\mathbf{X}_{\text{init}}$ with $\mathbf{X}'_{\text{out}}$ with skip connections to obtain the final reconstructed HSI $\hat{\mathcal{H}}$ that can further alleviate the loss of certain fine-grained details during the encoding and decoding process. Based on this method, we compute the HSI image $\hat{\mathcal{H}}$ with heightened fidelity and superior visual quality. In the following, we provide a detailed description of our SSHE, SIC, and MSSHE modules.

1) SSHE Module: The S-MSA [58], [59] has shown promising results for HSI reconstruction. However, it does not fully consider the inherent redundancy of spectral information in HSIs. Thus, it has large computational complexity, which inadvertently also compromises spectral domain reconstruction performance. In the imaging process of the CASSI system, sensors may introduce a certain level of electronic noise and thermal noise, while compression algorithms may lead to information loss. In addition, inaccuracies are possible during the mask generation process, which can result in additional noise when generating a

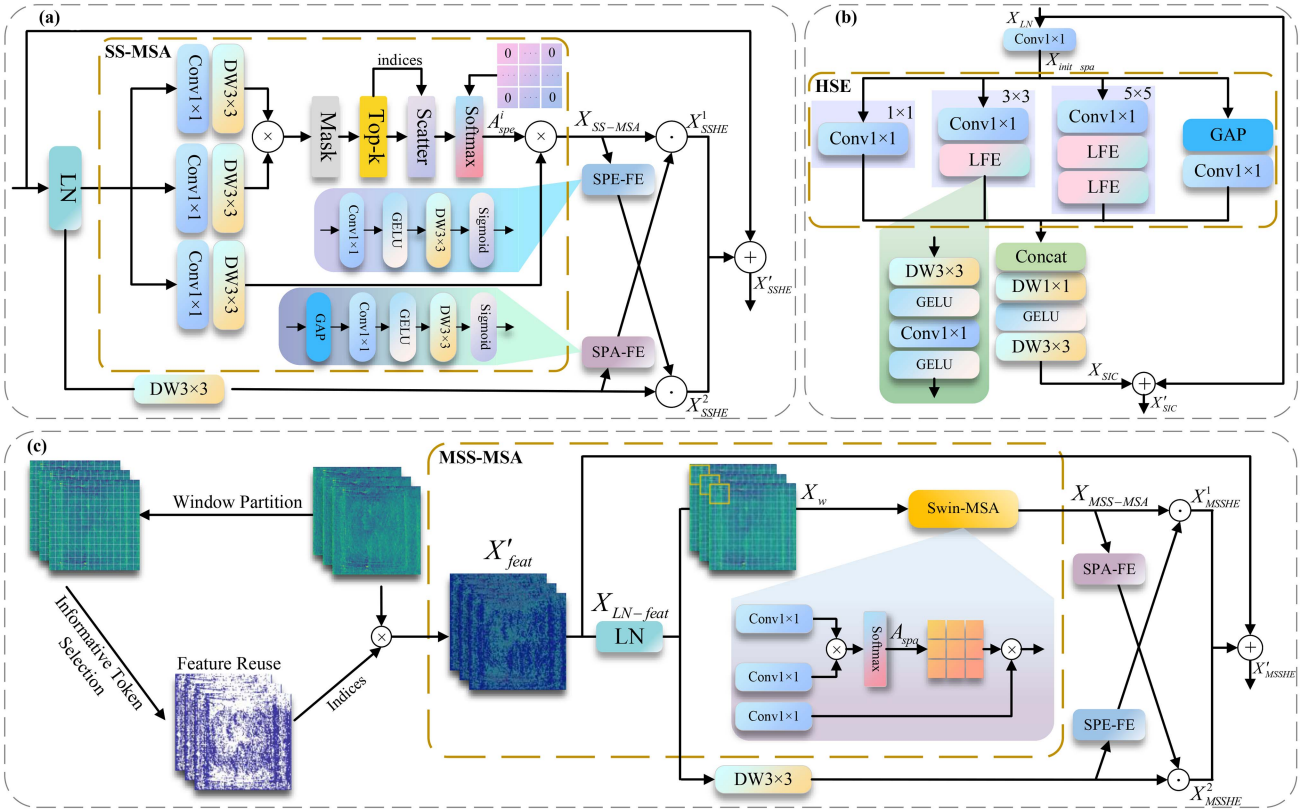

Fig. 2. (a) SSHE module consists of an SS-MSA module and an spatial feature enhancement (SPA-FE) module. (b) Illustration of the SIC module. (c) MSSHE module utilizes the dynamically generated mask from the coarse-grained feature extraction stage to guide the MSS-MSA to spatial sparsity representation. It also leverages the spectral feature enhancement (SPE-FE) to enhance spectral features.

3-D HSI cube from 2-D measurements. Ultimately, these noises will impact the reconstruction results of the HSI.

In recent years, the sparse transformers have achieved great success in image restoration task [18], [76], [78]. Inspired by this, we try to introduce it into HSI reconstruction. Specifically, we customize an SS-MSA architecture to solve noise signals and spectral information redundancy in HSI reconstruction. As shown in Fig. 2(a), we propose the SSHE module, which consists of an SS-MSA module, a spectral feature enhancement (SPE-FE) module, and a spatial feature enhancement (SPA-FE) module. Traditional S-MSA calculates the similarity matrix by performing a dot product between the query vector and key vectors, followed by a softmax operation. However, our proposed SS-MSA adopts a top-k calculation strategy, where the top-k values are first computed before applying the softmax operation to generate the similarity matrix. In our proposed framework, we adapt the multiple distinct top-k selections followed by summation that facilitates the capture of a broader spectrum of hierarchical information. It will provide a more comprehensive representation of interdata relationships. Based on this strategy, it can effectively capture spectral characteristics and enhance the ability of the model to characterize complex data structures. In this article, we optimize the calculation process of the similarity matrix with the aid of SS-MSA sparsity. The process only retains the relevant information related to the target, which reduces the model's computational cost while ensuring its performance. In addition, SS-MSA also helps mitigate the noise, further improving the reconstruction results' accuracy and reliability.

Specifically, we use the feature $\mathbf{X}_{\mathrm{LN}}$ to encode the contextual information in the spectral domain with a combination of 1×1 convolution and 3×3 depthwise separable convolution (DW-Conv) to generate projections of queries $\mathbf{Q}_{\mathrm{spe}}$, keys $\mathbf{K}_{\mathrm{spe}}$, and values $\mathbf{V}_{\mathrm{spe}}$ to enrich local contextual information. They are described as follows:

$$\mathbf{Q}_{\mathrm{spe}} = \mathbf{X}_{LN}\mathbf{W}_p^Q\mathbf{W}_D^Q \tag{1}$$

$$\mathbf{K}_{\mathrm{spe}} = \mathbf{X}_{LN}\mathbf{W}_p^K\mathbf{W}_D^K \tag{2}$$

$$\mathbf{V}_{\mathrm{spe}} = \mathbf{X}_{LN}\mathbf{W}_p^V\mathbf{W}_D^V \tag{3}$$

where $\mathbf{W}_P^{(\cdot)}$ denotes 1×1 pointwise convolution, and $\mathbf{W}_D^{(\cdot)}$ represents 3×3 depthwise convolution. Next, we respectively divided $\mathbf{Q}_{\mathrm{spe}}$, $\mathbf{K}_{\mathrm{spe}}$, and $\mathbf{V}_{\mathrm{spe}}$ into N attention heads along the channel dimension: $\mathbf{Q}_{\mathrm{spe}} = [\mathbf{Q}_{\mathrm{spe}}^1, \ldots, \mathbf{Q}_{\mathrm{spe}}^N]$, $\mathbf{K}_{\mathrm{spe}} = [\mathbf{K}_{\mathrm{spe}}^1, \ldots, \mathbf{K}_{\mathrm{spe}}^N]$, $\mathbf{V}_{\mathrm{spe}} = [\mathbf{V}_{\mathrm{spe}}^1, \ldots, \mathbf{V}_{\mathrm{spe}}^N]$, where each head has a dimension of $\dim_{\mathrm{spe}} = \frac{C}{N}$, with C representing the number of channels and N representing the number of heads. Subsequently, we calculate the similarity matrix $\mathbf{A}_{\mathrm{spe}}^j \in \mathbb{R}^{\frac{C}{N}\times\frac{C}{N}}$ between $\mathbf{Q}_{\mathrm{spe}}^j$ and $\mathbf{K}_{\mathrm{spe}}^j$ via dot product operation in each head, where j denotes the jth head, $\mathbf{Q}_{\mathrm{spe}}^j$ represents the current band of interest from which we aim to gather information, and $\mathbf{K}_{\mathrm{spe}}^j$

represents all other bands in the sequence, serving as reference points for comparison with $\mathbf{Q}_{\text{spe}}^{j}$. This aids in determining the correlation and significance of each band relative to $\mathbf{Q}_{\text{spe}}^{j}$. Then, top-*k* screening is conducted for the spectral information in the attention matrix. To fully exploit the spectral information, SS-MSA performs multiple calculations of spectral attention maps with different sparsity levels and fills the nontop-*k* regions with zero to obtain multiple spectral attention maps with different sparsity levels. The similarity computation is expressed as follows:

$$\mathbf{A}_{\text{spe}}^{i} = \text{soft}\max\left(\frac{\text{top-}k\left(\mathbf{Q}_{\text{spe}}^{j}\left(\mathbf{K}_{\text{spe}}^{j}\right)^{T}\right)}{\boldsymbol{\alpha}} + B\right) \tag{4}$$

$$\mathbf{Y}_{\text{spe}}^{j} = \sum_{t=1}^{\text{n}} \mathbf{A}_{\text{spe}}^{i} \cdot \mathbf{V}_{\text{spe}}^{j} \tag{5}$$

where B represents the relative positional encoding, the scaling factor α is a learnable parameter that is used to alleviate the saturation problem that may occur during softmax function computation, which can lead to gradient vanishing, *j* denotes the index of the current attention head, *k* represents sparsity level, $\mathbf{A}_{\text{spe}}^{i}$ represents the *i*th similarity matrix with top-*k* extracted regions, and $\mathbf{V}_{\text{spe}}^{j}$ represents the information related to each band in the *j*th head, carrying the actual features of the input, *n* represents the number of different top-*k* values. $\mathbf{Y}_{\text{spe}}^{j}$ represents the attention weighted values of the *j*th head.

Based on the obtained multiple similarity matrices with different sparsity levels, we perform softmax computation to focus SS-MSA on the spectral-related information in HSI, rather than irrelevant content. We get a series of similarity matrices with different sparsity levels, which are then computed with $\mathbf{V}_{\text{spe}}$ to generate spectral attention maps. Finally, we can obtain the spectral feature $\mathbf{Y}_{\text{spe}}$ by concatenating all $\mathbf{Y}_{\text{spe}}^{j}$. They can be denoted as follows:

$$\mathbf{Y}_{\text{spe}} = \text{concat}\left(\mathbf{Y}_{\text{spe}}^{1}, \mathbf{Y}_{\text{spe}}^{2}, \ldots, \mathbf{Y}_{\text{spe}}^{N}\right). \tag{6}$$

Once the spectral feature $\mathbf{Y}_{\text{spe}}$ is obtained, the output of SS-MSA, $\mathbf{X}_{\text{SS-MSA}}$ can be described as follows:

$$\mathbf{X}_{\text{SS-MSA}} = \mathbf{Y}_{\text{spe}}\mathbf{W}_{\text{spe}} \tag{7}$$

where $\mathbf{W}_{\text{spe}} \in \mathbb{R}^{C\times C}$ denotes the learnable weight matrix that is used for linear transformation in the attention mechanism.

Next, we incorporate SPA-FE and spectral feature enhancement (SPE-FE) modules into the SSHE to improve the spatial–spectral information interaction within the SSHE, thereby enhancing the quality of HSI reconstruction. The SPA-FE module uses global pooling to extract global information from the input feature map, which will facilitate the learning of abstract and generalized features to enhance the model's generalization capability. Subsequently, the module applies pointwise convolution (1×1 convolution), GELU activation function, and DW-Conv to the output of the global pooling layer. These operations aim to optimize feature representation while preserving the spatial structure of the input feature space. Finally, a sigmoid function is applied to enhance the capability of the network's feature representation. The SPE-FE module and SPA-FE module share similarities, yet SPE-FE diverge in their approach by omitting the global pooling layer to focus on extracting information in the spectral domain. In the SPA-FE module, feature information processed through DW convolution is injected into SS-MSA, while in the SPE-FE module, spectral information derived from SS-MSA is integrated into feature information processed after DW convolution. This interplay between the two modules fosters interaction between spatial and spectral information, enhancing the overall feature extraction capability of SSHE and thereby strengthening the model's perception and understanding of the data. Ultimately, we obtained the $\mathbf{X}'_{\text{SSHE}}$ feature that contains rich spectral feature information extracted by the SSHE

$$\mathbf{X}_{\text{SSHE}}^{1} = \mathbf{X}_{\text{SS-MSA}} \odot \text{SPA-FE}\left(\text{DW}\left(\mathbf{X}_{\text{LN}}\right)\right) \tag{8}$$

$$\mathbf{X}_{\text{SSHE}}^{2} = \text{DW}\left(\mathbf{X}_{\text{LN}}\right) \odot \text{SPE-FE}\left(\mathbf{X}_{\text{SS-MSA}}\right) \tag{9}$$

$$\mathbf{X}_{\text{SSHE}} = \mathbf{X}_{\text{SSHE}}^{1} + \mathbf{X}_{\text{SSHE}}^{2} + \mathbf{X}_{\text{init}} \tag{10}$$

where SPA-FE$(\cdot)$ and SPE-FE$(\cdot)$ represents the spatial feature enhancement and spectral feature enhancement module, respectively. DW$(\cdot)$ denotes the DW-Conv and $\odot$ represents elementwise multiplication.

2) SIC Module: The proposed SSHE utilizes spectral attention to model the spectral domain and extract spectral information from the HSI images. However, the spectral information extracted by SSHE mainly contains global low-frequency spectral information, lacking high-frequency local spatial information. This makes the dynamically generated masks fail to effectively represent the relationships between different regions, leading to poor spatial attention sparsity that can increase computational complexity or degrade the reconstruction quality. To address this, we design an SIC module that compensates for the spatial details ignored by the SSHE, enabling the fusion of spatial and spectral information. This improves the high-fidelity reconstruction of HSIs while preserving more detailed features and texture characteristics. The specific design of SIC module is illustrated in Fig. 2(b).

Specifically, we first input the $\mathbf{X}_{\text{LN}}$ into a pointwise convolutional operation for preliminary feature extraction, obtaining an initialized spatial feature $\mathbf{X}_{\text{init\_spa}}$. Then, $\mathbf{X}_{\text{init\_spa}}$ is fed into a high-frequency spatial feature extractor (HSE). The HSE is an inception structure, extracting multiscale spatial information from the initialized feature by four different branches that will compensate for the missing high-frequency spatial information after spectral attention. To reduce the parameter and computational complexity of the SIC module, each branch is equipped with a pointwise convolutional operation at its corresponding head, compressing the spectral dimension to $\dim_{\text{SIC}} = \frac{C}{4}$.

The four branches of HSE are composed of pointwise convolution, lightweight feature extraction module (LFE), and global average pooling (GAP) operations. LFE consists of DW-Convs, GELU activation function, and pointwise convolution, which is stacked to achieve feature information extraction at different scales. The deep features obtained from the four different scales are concatenated along the spectral dimension, and processed with two DW-Convs and an activation function to generate deep features $\mathbf{X}_{\text{SIC}}$ containing multiple scale information. Finally, to

reduce the loss of spatial detail information during processing, we apply skip connections to fuse the initial features $\mathbf{X}_{\text{LN}}$ extracted by layer normalization with the $\mathbf{X}_{\text{SIC}}$ to generate new feature $\mathbf{X}'_{\text{SIC}}$ that containing more spatial textures and details. Then, we can fuse the spectral feature $\mathbf{X}_{\text{SSHE}}$ obtained by the SSHE module with this spatial feature $\mathbf{X}'_{\text{SIC}}$ to get the spatial–spectral fusion feature $\mathbf{X}_{\text{fuse}}$. It is denoted as follows:

$$\mathbf{X}'_{\text{SIC}} = \mathbf{X}_{\text{SIC}} + \mathbf{X}_{\text{LN}} \tag{11}$$

$$\mathbf{X}_{\text{fuse}} = \text{concat}\left(\mathbf{X}_{\text{SSHE}}, \mathbf{X}'_{\text{SIC}}\right) \tag{12}$$

where concat$(\cdot)$ denotes the operation of concatenating these feature maps together.

3) MSSHE Module: In Cai et al.'s [60] work, a spatial sparse multihead attention mechanism is proposed that typically involves designing a small mask generator to create a fixed mask for extracting the top-*k* relevant spatial elements. For that method, the feature maps of the HSI undergo changes with the features encoding and decoding of the three-level U-Net structure. However, the masks generated by this strategy do not change, and cannot adaptively filter out irrelevant regions at different scales in each layer, leading to excessive sparsity or poor sparsity effects in the spatial sparse multihead attention. In contrast, our MSSHE module applies dynamically changing masks to extract the top-*k* relevant spatial elements. This allows the model to learn the relevance of the spatial elements at each stage, efficiently extract the top-*k* relevant spatial elements, and reduce the number of parameters. It enhances the model's ability to utilize spatial domain fine-grained features and global features. Furthermore, we employ a local attention mechanism and a sliding window strategy to facilitate information interaction between the windows. By doing so, we can significantly enhance the modeling capability for the spatial domain and improve the reconstruction performance of the proposed model. As illustrated in Fig. 2(c), our MSSHE module is composed of three major components, namely mask-guided spatial sparse multihead self-attention (MSS-MSA), SPE-FE, and SPA-FE.

We apply a projection mapping function to the fusion feature $\mathbf{X}_{\text{fuse}}$ with a $\text{Conv}_{3\times3}$ to get a new feature $\mathbf{F} \in \mathbb{R}^{H\times W\times(N_\lambda+1)}$. Then, we partition the feature $\mathbf{F}$ along the channel dimension into two parts: $\mathbf{X}_{\text{mask}} \in \mathbb{R}^{H\times W\times 1}$ and $\mathbf{X}_{\text{feat}} \in \mathbb{R}^{H\times W\times N_\lambda}$. Next, we apply GAP and subsequent binarization to the $\mathbf{X}_{\text{mask}}$ to get the new mask feature $\mathbf{X}'_{\text{mask}}$. Afterward, we also utilize it to guide the spatial feature $\mathbf{X}_{\text{feat}}$ to obtain sparse features. Subsequently, the top-*k* algorithm is performed on this sparse feature to select the relevant tokens as new spatial–spectral sparse feature $\mathbf{X}'_{\text{feat}}$. Subsequently, $\mathbf{X}'_{\text{feat}}$ passed through layer normalization to obtain $\mathbf{X}_{\text{LN-feat}}$. Finally, the $\mathbf{X}_{\text{LN-feat}}$ is divided into nonoverlapping local windows $\mathbf{X}_w \in \mathbb{R}^{\frac{H}{M}\times\frac{W}{M}\times N_\lambda}$, and spatial attention is calculated for each window in the spatial domain. In the spatial attention branch, $\mathbf{X}_w$ is linearly transformed into $\mathbf{Q}_{\text{spa}}, \mathbf{K}_{\text{spa}}, \mathbf{V}_{\text{spa}}$

$$\mathbf{Q}_{\text{spa}} = \mathbf{X}_w\mathbf{W}^Q, \mathbf{K}_{\text{spa}} = \mathbf{X}_w\mathbf{W}^K, \mathbf{V}_{\text{spa}} = \mathbf{X}_w\mathbf{W}^V \tag{13}$$

where $\mathbf{W}^Q$, $\mathbf{W}^K$, and $\mathbf{W}^V$ are learnable projection matrices. Following the strategy of SSHE with multihead attention, we divide the $\mathbf{Q}_{\text{spa}}$, $\mathbf{K}_{\text{spa}}$, and $\mathbf{V}_{\text{spa}}$ along the spectral dimension into multiple heads. In this section, we set the number of heads to 1 for simplification purposes. Subsequently, we use dot product interaction for $\mathbf{Q}_{\text{spa}}$ and $\mathbf{K}_{\text{spa}}$ to generate a similarity matrix $\mathbf{A}_{\text{spa}} \in \mathbb{R}^{M^2\times M^2}$

$$\mathbf{A}_{\text{spa}} = \text{soft}\max\left(\frac{\mathbf{Q}_{\text{spa}}\mathbf{K}^T_{\text{spa}}}{\beta} + B\right) \tag{14}$$

$$\mathbf{Y}_{\text{spa}} = \mathbf{A}_{\text{spa}} \cdot \mathbf{V}_{\text{spa}} \tag{15}$$

where β is a learnable parameter that adjusts the inner product before the softmax function, while $\mathbf{Y}_{\text{spa}}$ represents the attention weighted values and $\mathbf{W}_{\text{spa}} \in \mathbb{R}^{C\times C}$ denotes the learnable weight matrix that is used for linear transformation in the attention mechanism. Once the spatial feature $\mathbf{Y}_{\text{spa}}$ is obtained, the output of MSS-MSA, i.e., $\mathbf{X}_{\text{MSS-MSA}}$ can be described as follows:

$$\mathbf{X}_{\text{MSS-MSA}} = \mathbf{Y}_{\text{spa}}\mathbf{W}_{\text{spa}}. \tag{16}$$

Next, we integrate spatial feature and spectral feature enhancement modules within the MSSHE to improve the spatial–spectral information interaction within the MSSHE, thereby enhancing the quality of HSI reconstruction. In the case of the SPA-FE and SPE-FE modules within MSSHE, their inputs are different. Specifically, SPA-FE injects deep spatial information computed by MSS-MSA into shallow feature information obtained from DW convolution. In contrast, SPE-FE focuses on extracting spectral information from the shallow feature information processed by DW convolution, and then integrates the extracted spectral information into MSS-MSA. This strategic interaction between spatial and spectral information enhances MSSHE's capability in extracting spatial and spectral features. As a result, we obtained a feature $\mathbf{X}'_{\text{MSSHE}}$ extracted by the MSSHE to enrich spatial characteristic information

$$\mathbf{X}^1_{\text{MSSHE}} = \mathbf{X}_{\text{MSS-MSA}} \odot \text{SPE-FE}\left(\text{DW}\left(\mathbf{X}_{\text{LN-feat}}\right)\right) \tag{17}$$

$$\mathbf{X}^2_{\text{MSSHE}} = \text{DW}\left(\mathbf{X}_{\text{LN-feat}}\right) \odot \text{SPA-FE}\left(\mathbf{X}_{\text{MSS-MSA}}\right) \tag{18}$$

$$\mathbf{X}_{\text{MSSHE}} = \mathbf{X}^1_{\text{MSSHE}} + \mathbf{X}^2_{\text{MSSHE}} + \mathbf{X}'_{\text{feat}} \tag{19}$$

where SPA-FE$(\cdot)$ and SPE-FE$(\cdot)$ denote the spatial and spectral feature enhancement modules, respectively.

The MSSHE module focuses on capturing spatial details when reconstructing HSIs. However, the use of a sliding window approach may disrupt the global dependence of HSIs. Hence, we integrate the spectral interdomain feature information extracted by the SSHE module with the high-frequency local spatial information extracted by SIC. The fused feature information is then connected with the spatial details and texture features extracted by the MSSHE module and collectively input into a feedforward neural network (FFN) to provide global information to the spectral domain to alleviate the potentially disruptive effects. In this case, we separately learn the spatial and spectral information of HSIs by their corresponding spatial attention and spectral attention and fuse spatial and spectral attention features to gain the fused feature $\mathbf{Y}$, which helps improve reconstruction

$$\mathbf{Y} = \mathcal{F}\left(\text{Project}\left(\text{concat}\left(\mathbf{X}_{\text{feat}}, \mathbf{X}_{\text{MSSHE}}\right) + X_{\text{init}}\right)\right). \tag{20}$$

In the above-mentioned equation, concat$(\cdot)$ denotes the operation of concatenating the feature maps, $\mathcal{F}(\cdot)$ refers to a FFN, and the $\mathrm{Project}(\cdot)$ is a mapping function composed of $\mathrm{Conv}_{3\times3}$ that reduces the number of channels from $2N_\lambda$ to N_λ .

In our model, we adopted an FFN structure similar to Restormer [50]. Our FFN can control the flow of information at each hierarchical level, allowing each level to focus on fine details that complement those at other levels, and to better utilize contextual information to enrich features.

IV. Experiments

We conduct our experiments on both real HSI and simulation datasets. Following the existing literature [23], [58], [63], we also select a set of 28 wavelengths ranging from 450 to 650 nm by employing spectral interpolation techniques applied to the HSI data.

A. Experimental Settings

1) Simulated HSI Data: To ensure the fairness of the experiments, we adopt a similar experimental strategy to that of [23], [58], and [63], we also use CAVE [80] as the training dataset and KAIST [81] as the testing dataset. We selected ten scenarios from the KAIST dataset to assess the efficacy of our method and contrast it with alternative approaches. In addition, to validate the effectiveness of the proposed algorithm in the remote sensing domain and assess its generalization capability, we conducted tests on existing remote sensing datasets, such as Pavia Centre, Pavia University [82], and Urban. These datasets are cropped and resized to 256×256 pixels, with spectral dimensions uniformly selected from bands 11 to 38, resulting in a test set of size $256 \times 256 \times 28$. For the CAVE dataset, we selected 28 wavelengths from 450 to 650 nm and configured the fundamental channel with $\mathrm{C} = \mathrm{N}_\lambda = 28$ to retain the HSI information and obtained the HSI data by spectral interpolation. Then, we randomly cropped patches of size $256 \times 256 \times 28$ from the dataset for training.

2) Real HSI Data: For the real-world experiment, we utilized the real HSI data provided by TSA-Net [68] to verify the superiority of our method in real scenes. However, since the widely used datasets captured by the CASSI system consist mainly of indoor and outdoor close-range data and lack remote sensing datasets, we only conducted real experiments on the dataset provided by TSA-Net [68]. To this end, we randomly cropped the 3-D real HSI dataset to generate patches of size $660 \times 660 \times 28$, which matches the physical mask size, and set the displacement k in dispersion to 2.

3) Evaluation Metrics: For the simulation experiment, we conducted the quantitative comparison with the full-reference image quality assessment metrics, peak signal-to-noise ratio (PSNR), structure similarity index measure (SSIM), and spectral angle mapping (SAM), which are commonly employed for assessing the performance of HSI reconstruction in previous works [23], [58], [63]. Specifically, PSNR is used to assess the quality of reconstructed images by comparing differences between the reconstructed results and original images. A higher PSNR value indicates greater similarity between the reconstructed and original images. SSIM is based on similarities of local luminance, contrast, and structure between a reference image and a distorted image. It is applied to measure the similarity between restored images and references. The SSIM value ranges between [0, 1], where values closer to 1 indicate higher similarity. On the contrary, it reveals that there are larger differences between the reconstructed results and references. SAM is used to measure the similarity between two spectral vectors. It evaluates spectral similarity by comparing the angle between two vectors; a smaller angle indicates higher similarity, while a larger angle indicates lower similarity.

4) Implementation Details: Our DGST framework is implemented with PyTorch. We construct two variants of DGST with different complexities according to the difference in the number of DGSBs in the three-level U-net structure, called DGST-S and DGST-L, with the composition of the DGSB module block numbers DGST-S (112), and DGST-L (246), respectively. The DGSB is mainly composed of SSHE, SIC, and MSSHE. All DGST model variants are trained with Adam [83] optimizer ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) on a single RTX 3090 GPU for 300 epochs using a cosine annealing [84] scheme. The learning rate is set to 4×10^{-4}, and the model is optimized by minimizing the Charbonnier loss. The batch size is set to 4. In the shallow feature embedding module, the embedding dimension is set to 28, aiming to restore the spectral dimension of the concatenated HSI cube obtained during network input, and back it to its original size. In addition, the quantity of distinct top-k value for SS-MSA is set to 4, while the sparsity levels are set to $\frac{1}{2}C$, $\frac{2}{3}C$, $\frac{3}{4}C$, and $\frac{4}{5}C$, respectively. MSS-MSA utilizes a window size of 8, with a spatial sparsity rate set at 0.5. During the training phase, we use random horizontal flips and rotations to augment the training data, which makes the proposed model training more effective.

B. Comparison Results

1) Qualitative Results: We benchmark our proposed DGST method against 17 SOTA HSI reconstruction techniques, which encompass one model-based approaches (DeSCI [31]), five CNN-based approaches (λ-net [67], TSA-Net [68], DGSMP [72], GAP-Net [85], HDNet [69]), five transformer-based approaches (MST [58], MST++ [59], CST [60], S^2-transformer [61], D2PL-Net [86]), five deep unfolding-based approaches (ADMM-Net [87], DAUHST [62], EDUNet [88], LRSDN [89], and PADUT [63]) and one diffusion-based method (DiffSCI [90]). The main comparison metrics include params, floating point operations per second (FLOPS), PSNR, SSIM, and SAM, which are widely used in previous works, such as [58], [60], and [91]. In addition, to ensure fair experiments, we adopt the same settings as [58], [60], and [63], where all methods are trained on the same dataset and tested on ten simulated scene data. To validate the generalization performance of our method, we also compare it with nine SOTA HSI reconstruction methods on the popular remote sensing datasets of Pavia Centre, Pavia University, and Urban. The methods include four CNN-based techniques (λ -net [67], GAP-Net [85], ADMM-Net [87],

HDNet [69]) and five transformer-based methods (MST [58], MST++ [59], CST [60], DAUHST [62], and PADUT [63]). To ensure fairness, we trained all methods using the same settings. For the D2PL-Net, EDUNet, LRSDN, and DiffSCI, we used the values of PSNR and SSIM from the original papers as compared parameters since the codes of these three methods are not released. It should be noted that their dataset selection and processing approaches align with ours. However, SAM metrics were not disclosed in their papers, thus we cannot get them. Next, we will conduct detailed quantitative analyses on the KAIST, Pavia Centre, Pavia University, and Urban datasets as follows.

1) *KAIST Dataset:* Table I displays the objective assessment outcomes of various methods on the ten simulated scene datasets. In Table I, it can be seen that our smallest model DGST-S has fewer parameters than the compared methods, except MST-S. However, our DGST-S has higher SSIM, PSNR, and SAM compared to MST-S. Compared with all other methods, our DGST-S model has the lowest computational complexity. Our larger model DGST-L*, outperforms CST-L* for all the evaluation metrics. Specifically, compared to CST-L*, our approach is more efficient in terms of parameter count and computational complexity, yet it achieves a significant performance improvement. Our method shows an increase of 0.61 in PSNR, 0.08 in SSIM, and a decrease of 0.425 in SAM compared with CST-L*. This achievement is mainly attributed to the SS-MSA strategy. In comparison to CST-L*, our DGST with SS-MSA can more effectively extract spectral information and focus more precisely on the reconstruction of spectral dimension details. Moreover, compared to the most relevant methods D2PL-Net and S^2-transformer, our DGST-L* achieves better performance on all evaluation metrics. In particular, our DGST demonstrates a PSNR improvement of 0.64 dB and an SSIM improvement of 0.053 when compared to D2PL-Net. Compared to the S^2-transformer method, our method shows a PSNR improvement of 0.25 dB, an SSIM improvement of 0.007, and a reduction in SAM by 1.494. Compared with the recently introduced deep unfolding-based approaches DAUHST-2stg, EDUNet, LRSDN, and PADUT-3stg, our larger model DGST-L* is superior to the DAUHST-2stg and EDUNet in terms of SSIM, PSNR, and SAM. Our DGST-L* is only 0.22 lower in PSNR but with higher SSIM compared to the PADUT-3stg. Notably, in comparison to the recently proposed diffusion-based method DiffSCI, our method exhibits lower PSNR values only in scenarios 3 and 5. However, across all scenarios, our SSIM values consistently outperform those of DiffSCI. Moreover, from Table I, we can also find that our larger model DGST-L* has the highest SSIM value compared with all the other methods. Compared with all other SOTA methods, our DGST approach exhibits the second best performance in terms of spectral angle mapper (SAM) in the field of HSI reconstruction, apart from PADUT-3stg. This underscores the advantage of deep unfolding-based models in this domain.

2) *Pavia Centre, Pavia University, and Urban Dataset:* According to the results in Table II, the proposed DGST demonstrates significant advantages in terms of PSNR and SSIM compared to the four CNN-based methods. Furthermore, when compared to MST-L, MST++, and DAUHST-2stg; our DGST also leads in terms of PSNR and SSIM. In comparison to CST-L, although DGST has a slightly lower PSNR on the Urban dataset, it exhibits superior performance in terms of SSIM. On the Pavia Centre and Pavia University datasets, our DGST shows improvements over CST-L. Compared to PADUT-3stg, DGST stands out more on the Pavia Centre and Urban datasets; while on the Pavia Centre dataset, although the PSNR and SSIM are 0.04 and 0.001 dB lower than PADUT-3stg, it still achieves strong results. To validate the real-time performance of our method, we conducted inference time calculations across the Pavia Centre, Pavia University, and Urban datasets. As indicated in the table, our approach exhibits commendable performance in terms of real-time image reconstruction, with only slight delays compared to HDNet and MST++. Notably, our method achieves superior reconstruction quality compared to HDNet and MST++ while sacrificing real-time constraints, evidenced by higher PSNR, SSIM, and SAM metrics. In comparison to DAUHST-2stg and PADUT-3stg, our method outperforms DAUHST-2stg notably. However, it does not exhibit significant advantages over PADUT-3stg in terms of the Pavia University dataset. This underscores the advantages of deep unfolding-based models in the reconstruction domain. However, for a comprehensive evaluation of various methods' generalization capability and effectiveness on remote sensing datasets, we conducted a overall assessment across the entire dataset. From Table II, it can be clear seen that our DGST method achieves the best performance in terms of PSNR and SSIM metrics, highlighting its superior effectiveness in spatial and pixel-level reconstruction. This underscores DGST's significant advantage in the field of remote sensing image reconstruction, demonstrating excellent generalization performance. Overall, CST-L demonstrates good generalization but lacks specialization, whereas PADUT-3stg excels in specialization but lacks generalization. Our method also obtained acceptable performance for SAM that ranked second among above-mentioned methods. In comparison, our DGST method not only demonstrates strong specialization capabilities but also exhibits excellent generalization, along with good real-time performance.

As shown in Tables I and II, the results validate the superiority of our proposed DGST method in HSI reconstruction. This is mainly attributed to the spatial high-frequency information compensation provided by our designed SIC module and the low-frequency information extraction enabled by the spectral attention module. Furthermore, the dynamic generation of masks through the spectral attention module serves as sparse guidance for spatial attention that effectively extracts sparse spatial information from the high-dimensional HSI. This strategy not only

TABLE I
COMPARISON OF DGST'S RESULTS FOR 10 SIMULATED SCENARIOS ON THE KAIST DATASET WITH PARAMS, FLOPS, PSNR (TOP), SSIM (MIDDLE), AND SAM (BOTTOM) FOR DIFFERENT APPROACHES

Methods	Params	GFLOPs	Metrics	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg
DeSCI [31]	–	–	PSNR	27.13	23.04	26.62	34.95	23.94	22.38	24.45	22.03	24.56	23.59	25.27
			SSIM	0.748	0.62	0.818	0.897	0.706	0.683	0.743	0.673	0.732	0.587	0.721
			SAM	11.263	16.049	8.366	6.439	11.605	16.693	11.297	21.620	9.668	17.381	13.038
λ-Net [67]	62.64M	117.98	PSNR	30.1	28.49	27.73	37.01	26.19	28.64	26.47	26.09	27.5	27.13	28.53
			SSIM	0.849	0.805	0.87	0.934	0.817	0.853	0.806	0.831	0.826	0.816	0.841
			SAM	14.127	17.409	15.605	24.039	16.351	26.036	14.078	27.574	15.862	26.020	19.710
TSA-Net [68]	44.25M	110.06	PSNR	32.03	31	32.25	39.19	29.39	31.44	30.32	29.35	30.01	29.59	31.46
			SSIM	0.892	0.858	0.915	0.953	0.884	0.908	0.878	0.888	0.89	0.874	0.894
			SAM	8.736	10.350	7.391	8.369	6.721	9.700	7.659	11.377	7.668	9.558	8.753
DGSMP [72]	3.76M	646.65	PSNR	33.26	32.09	33.06	40.54	28.86	33.08	30.74	31.55	31.66	31.44	32.63
			SSIM	0.915	0.898	0.925	0.964	0.882	0.937	0.886	0.923	0.911	0.925	0.917
			SAM	9.225	11.920	7.740	7.741	9.878	8.275	8.081	10.905	8.340	7.366	8.947
GAP-Net [85]	4.27M	78.58	PSNR	33.74	33.26	34.28	41.03	31.44	32.4	32.27	30.46	33.51	30.24	33.26
			SSIM	0.911	0.9	0.929	0.967	0.919	0.925	0.902	0.905	0.915	0.895	0.917
			SAM	9.114	13.071	8.605	9.549	7.874	12.611	8.404	16.084	8.757	13.234	10.730
ADMM-Net [87]	4.27M	78.58	PSNR	34.12	33.62	35.04	41.15	31.82	32.54	32.42	30.74	33.75	30.68	30.68
			SSIM	0.918	0.902	0.931	0.966	0.922	0.924	0.896	0.907	0.915	0.895	0.895
			SAM	8.596	13.048	8.133	9.372	7.900	12.758	8.340	16.889	8.291	12.831	10.616
HDNet [69]	2.37M	154.76	PSNR	35.14	35.67	36.03	42.3	32.69	34.46	33.67	32.48	34.89	32.38	34.97
			SSIM	0.935	0.94	0.943	0.969	0.946	0.952	0.926	0.941	0.942	0.937	0.943
			SAM	7.376	8.155	5.653	5.817	5.070	6.856	6.550	8.147	6.352	6.808	6.679
MST-S [58]	0.93M	12.96	PSNR	34.71	34.45	35.32	41.5	31.9	33.85	32.69	31.69	34.67	31.82	34.26
			SSIM	0.93	0.925	0.943	0.967	0.933	0.943	0.911	0.933	0.939	0.926	0.935
			SAM	7.299	10.655	6.809	8.687	7.136	10.335	6.932	12.609	8.012	10.274	8.875
MST-L [58]	2.03M	28.15	PSNR	35.4	35.87	36.51	42.27	32.77	34.8	33.66	32.67	35.39	32.5	35.18
			SSIM	0.941	0.944	0.953	0.973	0.947	0.955	0.925	0.948	0.949	0.941	0.948
			SAM	7.028	8.122	6.085	7.425	5.800	7.848	6.270	10.338	7.467	8.357	7.474
MST++ [59]	1.33M	19.42	PSNR	35.77	36.21	37.36	43.81	33.37	35.39	34.32	33.67	36.63	33.35	35.99
			SSIM	0.948	0.949	0.961	0.981	0.956	0.962	0.937	0.959	0.958	0.952	0.956
			SAM	6.740	7.477	4.595	6.957	4.641	6.613	5.927	8.295	5.725	6.621	6.359
CST-S [60]	1.20M	11.67	PSNR	34.78	34.81	35.42	41.84	32.29	34.49	33.47	32.89	34.96	32.14	34.71
			SSIM	0.93	0.931	0.944	0.967	0.939	0.949	0.922	0.945	0.944	0.932	0.94
			SAM	7.508	8.142	5.682	6.576	5.428	7.203	6.224	8.601	6.291	6.879	6.854
CST-L* [60]	3.00M	40.1	PSNR	35.96	36.84	38.16	42.44	33.25	35.72	34.86	34.34	36.51	33.09	36.12
			SSIM	0.949	0.955	0.962	0.975	0.955	0.963	0.944	0.961	0.957	0.945	0.957
			SAM	6.404	6.467	4.269	5.784	4.690	5.918	5.437	6.708	5.465	5.960	5.710
D2PL-Net [86]	7.65M	84.71G	PSNR	35.83	36.97	37.72	43.93	33.44	35.59	34.54	33.44	36.43	33.04	36.09
			SSIM	0.926	0.950	0.938	0.931	0.894	0.895	**0.954**	0.896	0.901	0.837	0.912
			SAM	–	–	–	–	–	–	–	–	–	–	–
EDUNet [88]	1.51M	24.24	PSNR	**36.48**	37.65	37.19	42.85	34.29	35.70	35.37	34.18	36.81	33.46	36.40
			SSIM	0.951	0.961	0.963	0.981	0.962	0.966	0.949	0.962	0.960	0.951	0.961
			SAM	–	–	–	–	–	–	–	–	–	–	–
DiffSCI [90]	–	–	PSNR	34.96	34.60	**39.83**	42.65	**35.21**	33.12	36.29	30.42	37.27	28.49	35.28
			SSIM	0.907	0.905	0.949	0.951	0.946	0.917	0.944	0.887	0.931	0.821	0.916
			SAM	–	–	–	–	–	–	–	–	–	–	–
S^2-Transformer [61]	3.01M	199.65	PSNR	36.17	37.57	37.29	42.96	34.40	**36.44**	35.41	**34.50**	36.54	33.57	36.48
			SSIM	0.949	0.958	0.957	0.975	0.960	0.965	0.946	0.963	0.959	0.952	0.958
			SAM	7.151	8.353	5.224	6.108	5.359	7.136	6.425	8.855	6.569	6.605	6.779
LRSDN [89]	–	–	PSNR	35.44	34.89	38.90	**45.29**	34.71	33.18	**37.76**	30.57	**39.49**	30.62	36.08
			SSIM	0.923	0.909	0.961	0.985	0.949	0.930	0.964	0.901	0.963	0.889	0.938
			SAM	–	–	–	–	–	–	–	–	–	–	–
DAUHST-2stg [62]	1.40M	18.44	PSNR	35.93	36.70	37.96	44.38	34.13	35.43	34.78	33.65	37.42	33.07	36.34
			SSIM	0.943	0.946	0.959	0.978	0.954	0.957	0.940	0.950	0.955	0.941	0.952
			SAM	6.136	7.124	4.340	5.868	4.374	6.619	5.417	8.566	5.200	6.338	5.998
PADUT-3stg [63]	1.35M	22.91	PSNR	36.25	**37.92**	39.63	44.55	34.59	35.58	35.69	33.76	**38.26**	33.24	**36.95**
			SSIM	0.951	0.963	**0.970**	0.985	0.964	0.965	0.950	0.960	0.963	0.947	0.962
			SAM	**5.798**	**5.692**	**3.557**	**4.036**	**3.486**	**5.046**	**4.915**	**6.594**	**4.263**	**5.092**	**4.848**
DGST-S	1.03M	10.1	PSNR	35.00	35.47	37.17	43.78	32.73	34.39	33.58	32.53	35.49	31.97	35.21
			SSIM	0.940	0.945	0.962	0.984	0.947	0.958	0.934	0.955	0.952	0.938	0.951
			SAM	6.673	6.857	4.547	4.998	4.652	5.887	5.795	7.011	6.241	5.506	5.817
DGST-L	2.74M	23.48	PSNR	36.14	37.38	38.08	44.02	33.87	35.62	34.80	33.86	37.30	33.10	36.41
			SSIM	0.956	0.960	0.963	0.984	0.961	0.964	0.943	0.962	0.962	0.953	0.961
			SAM	5.865	6.277	4.059	4.763	4.351	5.603	5.440	6.816	5.196	5.317	5.369
DGST-L*	2.74M	29.72	PSNR	36.43	37.75	38.47	44.56	33.96	35.83	35.10	34.07	37.57	**33.58**	36.73
			SSIM	**0.959**	**0.964**	0.966	**0.986**	**0.965**	**0.969**	0.947	**0.967**	**0.967**	**0.958**	**0.965**
			SAM	5.786	6.224	4.060	4.444	4.300	5.484	5.399	6.725	5.161	5.262	5.285

Where "*" indicates that no sparse approach is taken. Results shown in bold are the best.

TABLE II
COMPARISON ON THE PAVIA CENTRE, PAVIA UNIVERSITY, AND URBAN DATASETS WITH PARAMETERS, FLOPS, PSNR, AND SSIM FOR DIFFERENT APPROACHES

	λ-Net	GAP-Net	ADMM-Net	HDNet	MST-L	MST++	CST-L*	DAUHST-2stg	PADUT-3stg	DGST-L*
Params	62.64M	4.27M	4.27M	2.37M	2.03M	**1.33M**	3.00M	1.40M	1.35M	2.74M
GFLOPs	117.98	78.58	78.58	154.76	28.15	19.42	40.1	**18.44**	22.91	29.72
Time	0.45	0.19	0.17	**0.09**	0.23	0.1	0.26	0.3	0.19	0.16
Pavia Centre										
PSNR	27.17	31.02	31.23	32.41	32.87	33.04	33.23	33.14	33.27	**33.31**
SSIM	0.660	0.841	0.847	0.881	0.896	0.896	0.902	0.900	0.905	**0.906**
SAM	10.365	7.503	7.569	5.610	5.828	5.507	5.607	7.503	**4.660**	4.799
Pavia University										
PSNR	25.50	29.14	29.21	30.34	30.99	31.28	31.39	31.28	**31.43**	31.41
SSIM	0.672	0.847	0.849	0.886	0.902	0.904	0.910	0.905	**0.913**	0.912
SAM	9.971	7.529	7.474	5.486	5.891	5.361	5.307	7.529	**4.668**	4.808
Urban										
PSNR	27.40	31.13	31.33	32.38	32.75	32.81	**33.10**	33.09	33.07	33.09
SSIM	0.671	0.845	0.851	0.889	0.899	0.899	0.904	0.902	0.906	**0.908**
SAM	8.813	6.632	6.679	5.042	5.056	4.957	4.556	6.632	**3.787**	3.900
Average of Pavia Centre, Pavia University, and Urban Datasets										
PSNR	26.69	30.43	30.59	31.71	32.20	32.38	32.57	32.50	32.59	**32.60**
SSIM	0.667	0.844	0.849	0.885	0.899	0.900	0.905	0.902	0.908	**0.909**
SAM	9.716	7.221	7.241	5.379	5.592	5.275	5.157	7.221	**4.372**	4.502

Here, * indicates that no sparse approach is taken. The best results are bold-faced.

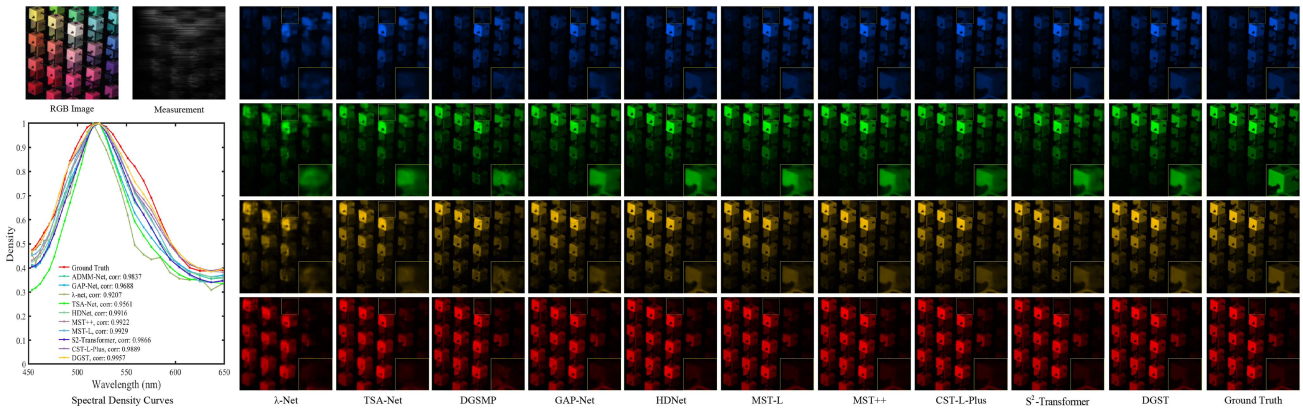

Fig. 3. Reconstruction results on Scene 2. Nine SOTA methods and the proposed method are presented on four out of 28 spectral channels. A comparison plot of the spectral density curve profiles for the selected regions is displayed in the bottom left position. For a more detailed analysis, we recommend zooming in for better visualization.

reduces Params and GFLOPs but also achieves better reconstruction performance compared to the most relevant methods. Overall, our method slightly lags behind the PADUT-3stg, but our method delves into the potential of the end-to-end deep learning approach, offering fresh insights into HSI reconstruction and yielding satisfactory outcomes in this domain.

2) Simulated HSI Reconstruction Results:

1) *KAIST Dataset:* In this experiment, we select four channels (466.0, 536.5, 584.5, and 648.0 nm) out of the 28 spectral channels to visualize Scene 2 from the KAIST dataset. We visually show the reconstruction of our method and other nine advanced methods with representative examples in Fig. 3. The figure illustrates that conventional model-based approaches struggle to completely recover all characteristics of the original HSI. Previous deep learning-based methods struggle to balance the preservation of high-frequency texture information and low-frequency structural information in HSI, leading to oversmoothing, loss of detailed information, or generation of speckle textures and color artifacts when preserving details. In contrast, our DGST method extracts high-frequency texture information through the SIC module and models the spatial and spectral dimensions separately using spatial attention and spectral attention to capture both the high-frequency texture information and low-frequency structural information of HSI. Fig. 3 indicates that our method performs exceptionally well in maintaining spatial details. We also generate spectral density curves for the reconstructed regions to validate the superiority of our method in terms of spectral consistency using ground truth. The spectral correlation coefficients displayed in the bottom left corner in Fig. 3 indicate that our method achieves the highest spectral correlation, affirming its effectiveness and superiority in HSI reconstruction.

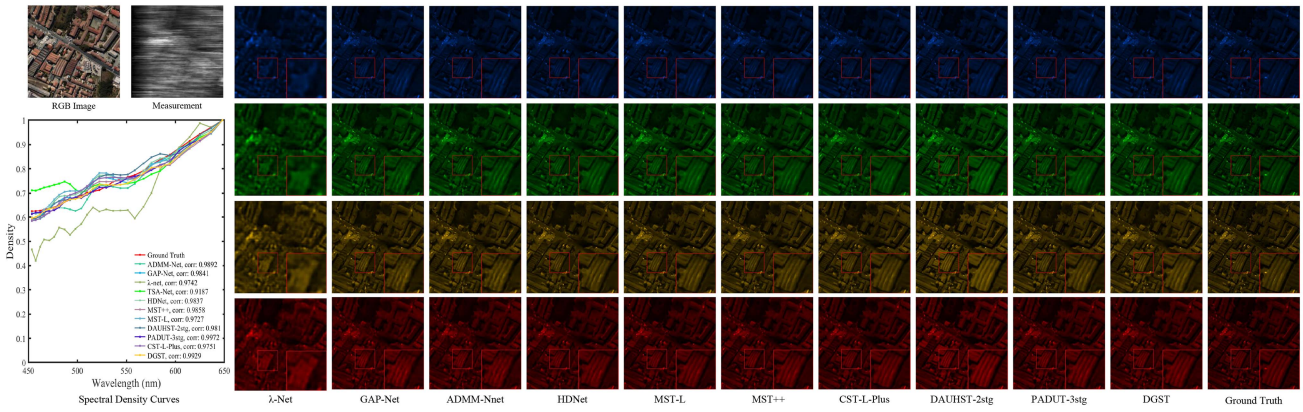

Fig. 4. Reconstruction results on Urban datasets. Nine SOTA methods and the proposed method are presented on four out of 28 spectral channels. A comparison plot of the spectral density curve profiles for the selected regions is displayed in the bottom left position. For a more detailed analysis, we recommend zooming in for better visualization.

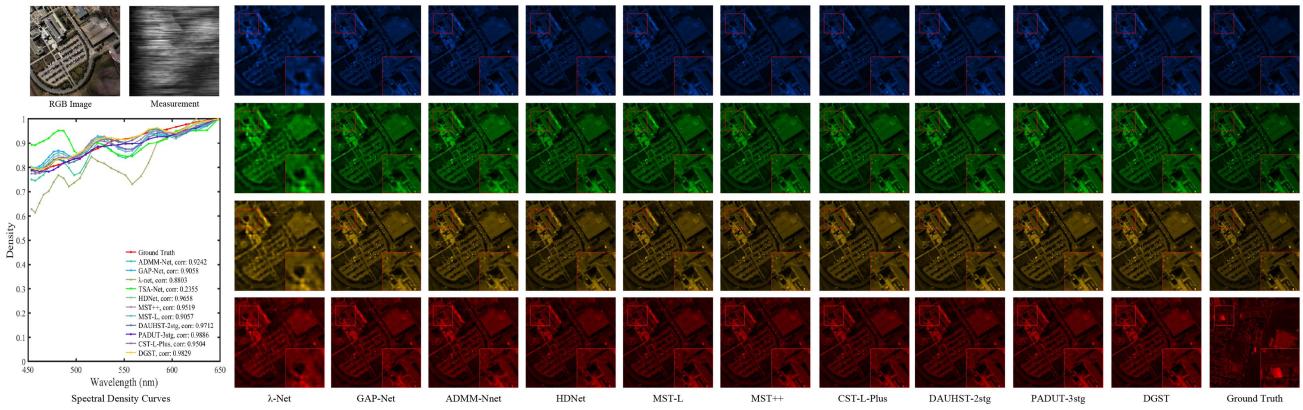

Fig. 5. Reconstruction results on Pavia University. Nine SOTA methods and the proposed method are presented on four out of 28 spectral channels. A comparison plot of the spectral density curve profiles for the selected regions is displayed in the bottom left position. For a more detailed analysis, we recommend zooming in for better visualization.

2) *Pavia Centre Dataset:* In the Pavia Centre dataset, as illustrated in Fig. 4, the variety of architectural shapes present in the scenes results in an overwhelming level of detail, leading to less than optimal reconstruction outcomes. Among CNN-based hyperspectral reconstruction methods, λ-net, GAP-Net, ADMM-Net, and HDNet display varying degrees of blurring and information loss in the reconstructed images. Despite HDNet achieving relatively superior reconstruction results, some blurring persists along building edges, failing to delineate clear boundaries. Transformer-based reconstruction methods excel in preserving contour information but struggle to effectively reconstruct edge details. Among these methods, MST-L, MST++, and CST-L-Plus outperform CNN-based methods in reconstructing more intricate details, with some enhancement in addressing building boundary issues, albeit with tendencies toward excessive smoothing and blurred reconstruction of small structures. Similar challenges are observed with the DAUHST-2stg method. Conversely, PADUT-3stg, combined with our approach, further addresses these issues, demonstrating outstanding performance in reconstructing edge details of buildings of varying sizes and achieving seamless transitions. The spectral correlation coefficients depicted in the bottom left corner of Fig. 4 attest that our method achieves the highest spectral correlation except for PADUT-3stg, thereby affirming its efficacy and superiority in hyperspectral image reconstruction.

3) *Pavia University Dataset:* In the Pavia University dataset, as illustrated in the Fig. 5, the presence of small-scale buildings in the scene leads to suboptimal reconstruction outcomes. Despite employing CNN-based hyperspectral reconstruction methods, such as λ-net, GAP-Net, and ADMM-Net, varying degrees of blurriness and information loss persist. Moreover, HDNet suffers from blurriness in reconstructing small targets, resulting in a significant loss of architectural details. In contrast, all transformer-based methods, depth-expanded techniques, and our proposed approach demonstrate notably superior reconstruction performance. While transformer-based methods and DAUHST-2stg exhibit some distortion in reconstructing lower edge lines, PADUT-3stg and our

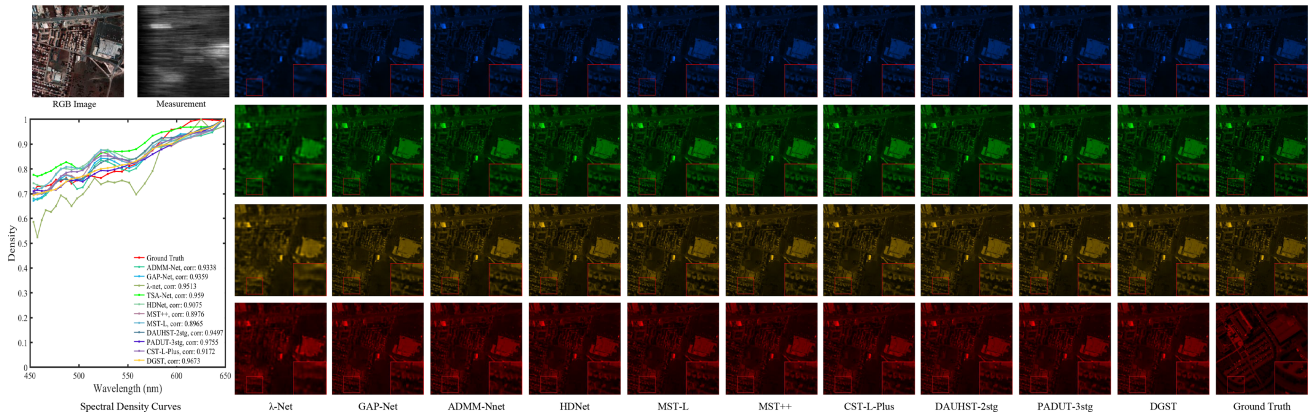

Fig. 6. Reconstruction results on Pavia Centre. Nine SOTA methods and the proposed method are presented on four out of 28 spectral channels. A comparison plot of the spectral density curve profiles for the selected regions is displayed in the bottom left position. For a more detailed analysis, we recommend zooming in for better visualization.

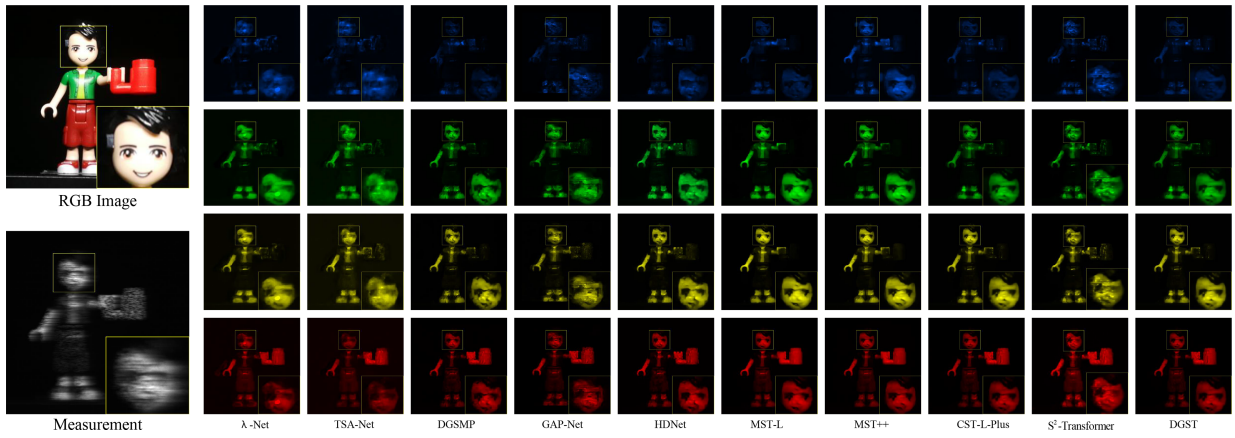

Fig. 7. Reconstructed real HSI comparisons of Scene 3 with four out of 28 spectral channels, including nine SOTA methods and our proposed method DGST. Zoom in for a better view.

method excel in this aspect. The spectral correlation coefficient in the bottom-left corner of Fig. 5 confirms that our method achieves the highest spectral correlation except for PADUT-3stg, thus validating its effectiveness and superiority in HSI reconstruction.

4) *Urban Dataset:* In the Urban dataset, as illustrated in Fig. 6, the presence of smaller scale buildings poses challenges for CNN-based hyperspectral reconstruction methods, such as λ -net, GAP-Net, and ADMM-Net, resulting in varying degrees of blurriness and information loss. In addition, HDNet exhibits subpar reconstruction performance for densely populated areas, characterized by significant blurriness and loss of detail. Conversely, transformer-based approaches show improvements in addressing these issues, although methods such as MST-L, MST++, and CST-L-Plus suffer from some degree of oversmoothing. In contrast, the results obtained by PADUT-3stg and our proposed method are notably superior. The spectral correlation coefficient in the bottom-left corner of Fig. 6 demonstrates that our method achieves the highest spectral correlation except for PADUT-3stg, confirming its effectiveness and superiority in HSI reconstruction.

3) Real HSI Reconstruction Results: To validate the effectiveness of the proposed DGST method on real-world scenes, we conducted additional experiments on a real-scene dataset with nine comparative methods. A qualitative comparison of results is provided in Fig. 5. This dataset was captured using an HSI system designed by TSA-Net [68], with each HSI containing 28 spectral channels. For the real-world scenes experiment, we followed the same setup as [58], [60], and [68] and retrained the models using the CAVE [80] and KAIST [81] datasets for training. To simulate more realistic conditions for scene reconstruction in the training set, we injected 11-bit random noise into the 2-D compressed measurement images, which were used as inputs to the model during training. From Fig. 5, we can see that our DGST method exhibits significant improvements in reconstruction compared to the previous model-based algorithms and CNN-based methods. As compared to the transformer-based methods [58], [60], [61], our method can reconstruct more spatial detail and texture information, resulting in more realistic image details. These results demonstrate the advantages of our DGST method in balancing high-frequency texture information and low-frequency structural information, particularly in the zoomed region.

TABLE III
ABLATION STUDIES OF INDIVIDUAL MODULES OF DGST

S-MSA	SS-MSA	Spa-FE	CSAB	SIC	MSS-MSA	Spe-FE	Params (M)	FLOPs (G)	PSNR (dB)	SSIM	SAM
✓							0.66	6.55	33.43	0.928	7.146
	✓						0.63	6.44	33.92	0.931	6.724
	✓	✓					0.65	6.57	33.04	0.932	6.605
	✓	✓	✓				0.71	8.03	34.16	0.935	6.519
	✓	✓	✓	✓			0.78	8.74	34.43	0.941	6.357
	✓	✓	✓	✓	✓		0.98	10.02	35.12	0.950	5.883
	✓	✓	✓	✓	✓	✓	1.03	10.10	35.21	0.951	5.817

TABLE IV
ABLATION STUDIES OF DGST CORE COMPONENTS

Method	Params (M)	FLOPs (G)	PSNR (dB)	SSIM	SAM
SSHE	0.71	8.03	34.16	0.935	6.519
SSHE+SIC	0.78	8.74	34.43	0.941	6.357
SSHE+MSSHE	0.93	9.38	34.77	0.947	6.174
SSHE+SIC+MSSHE (DGST)	1.03	10.10	35.21	0.951	5.817

TABLE V
ABLATION STUDY OF DIFFERENT SPATIAL AND SPECTRAL MULTIPLE SELF-ATTENTION MECHANISMS

Model	Params (M)	FLOPs (G)	PSNR (dB)	SSIM	SAM
Spectral MSA					
S-MSA	0.66	6.55	33.43	0.928	7.146
SS-MSA	0.63	6.44	33.92	0.931	6.724
Spectral MSA					
G-MSA	1.03	19.38	35.12	0.949	5.889
W-MSA	1.03	10.10	35.09	0.949	5.967
Swin-MSA	1.03	10.10	35.21	0.951	5.817

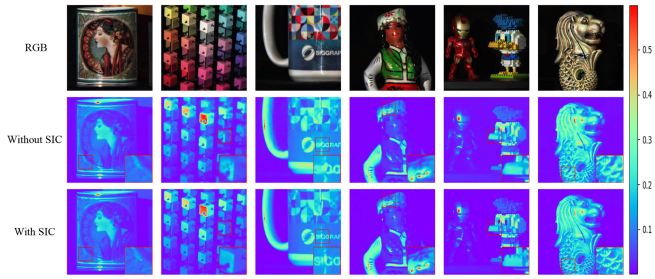

Fig. 8. Reconstruction effect of DGST-S. At the top is the original RGB image, while the middle and bottom rows show the reconstructed results without and with SCI, respectively. SCI is employed in combination with SSHE to generate a mask that guides MSSHE to pay attention to details during the reconstruction process.

C. Ablation Study

This section includes ablation experiments for our DGST model, where we eliminate different modules and use S-MSA as our benchmark model.

1) Break-Down Ablation: To explore the influence of various modules on the reconstruction performance (PSNR and SSIM), we decomposed the DGST model and conducted experimental evaluations. The experimental results are summarized in Table III. Our baseline model achieves a PSNR of 33.43 dB and an SSIM value of 0.928. As we gradually incorporate our proposed modules, the reconstruction performance of our model improves, and the metrics show significant improvements. Specifically, when we replaced S-MSA with SS-MSA, the PSNR performance was improved by 0.49 dB. In addition, there was a reduction in parameter count and computational complexity. Subsequently, the PSNR performance further increases by 0.12 dB by introducing the CSAB module. Moreover, the inclusion of the SIC module also makes a PSNR improvement of 0.27 dB and an SSIM improvement of 0.006, and a reduction of SAM by 0.162. Finally, by embedding the MSS-MSA module into our proposed model, our overall model achieved a unified improvement of 0.69 dB in PSNR and 0.009 in SSIM, with a reduction of SAM by 0.474. These results validate that the effectiveness of our proposed improvement modules and the efficacy of the interactions between these modules.

2) SIC Modular Ablation: We conducted ablation experiments to validate the impact of the proposed SIC module on the reconstruction performance. We compare the results with and without SIC in Table IV. Only utilizing SIC, the recorded PSNR is 34.43 dB. When only MSSHE is used, the PSNR is 34.77 dB. However, when both SIC and MSSHE are employed simultaneously, the PSNR improves by 0.44 dB. To visually demonstrate the structural and textural changes in image reconstruction after introducing the SIC module, we visualize the reconstructed results of DGST in Fig 6. From the figure, we can see that the introduction of the SIC module can capture more spatial details and textural information. This improvement is attributed to the SIC capability of the SIC module, which focuses on capturing high-frequency spatial details within the HSI images. It dynamically generates masks by integrating the captured spatial details with the globally captured spectral information using the MSS-MSA attention mechanism. These masks effectively represent the spatial sparsity relationships within HSI, guiding the MSS-MSA to focus on reconstructing the details and structures. Our results convincingly verify the effectiveness of our SIC module.

3) Comparing Self-Attention Mechanisms: We also compared the standard MSA with other variations of MSA techniques. In terms of spectral MSA, we employed two methods, S-MSA and SS-MSA. The results presented in Table V indicate an enhancement of 0.49 dB in PSNR and a decrease of 0.422 in SAM with SS-MSA, accompanied by a reduction in both parameters and computational requirements. This is attributed to the SS-MSA which can reduce interspectral redundancy and mask inaccuracies caused by measurement errors. To validate the effectiveness of dynamic spectral-guided spatial sparsity, we conducted simulations for spatial MSA using three methods:

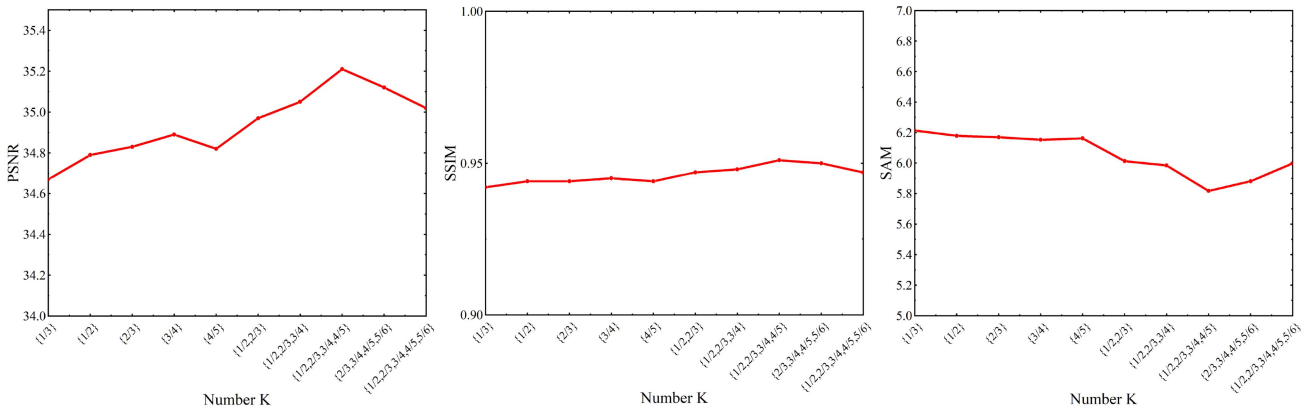

Fig. 9. Reconstruction effect of DGST-S. At the top is the original RGB image, while the middle and bottom rows show the reconstructed results without and with SCI, respectively. SCI is employed in combination with SSHE to generate a mask that guides MSSHE to pay attention to details during the reconstruction process.

Global MSA (G-MSA), local window MSA (W-MSA), and swin-MSA. From Table V, it can be observed that swin-MSA outperformed the other two methods with PSNR improvements of 0.09 and 0.12 dB, and SAM decreased by 0.072 and 0.15, respectively. This gain is attributed to swin-MSA's advantage in interwindow information interaction, enhancing the network's ability to capture long-range dependencies.

4) Sparsity Ablation Study of SS-MSA: Due to the high similarity between neighboring pixels, the top-k selection operation helps to reduce irrelevant contextual information from distant pixels. The choice of different K (top-k) values significantly impacts model performance. To thoroughly investigate this impact, we conducted multiple experiments evaluating different K values. From Fig. 9, it can be observed that the model generally performs poorly when K is small. And the performance will be improved as K increases because larger K values introduce richer spectral information. However, performance starts to decline as K continues to increase. When we aggregate computations using two different K values, we observe a performance improvement. Furthermore, we observe further performance enhancement by aggregating computations with multiple different K values, especially peaking at K values of $\frac{1}{2}$, $\frac{2}{3}$, $\frac{3}{4}$, and $\frac{4}{5}$. This improvement mainly benefits from the effective integration of more global information through the aggregation of multiple K values. However, performance declines again when too many K values are aggregated, as excessive K values may introduce irrelevant or unhelpful features. Therefore, based on experimental results, we ultimately determine to aggregate four K values, setting them to $\frac{1}{2}$, $\frac{2}{3}$, $\frac{3}{4}$, and $\frac{4}{5}$, to strike a balance between introducing beneficial information and avoiding the introduction of unintended information.

5) Dynamic Mask-Guided Ablation: In order to verify the effectiveness of our proposed dynamic mask, we conducted ablation experiments comparing different mask generation strategies. As shown in Table VI, our dynamic mask strategy only increased FLOPs by 0.07 G compared to the fixed mask strategy. However, it improved PSNR and SSIM by 0.16 and 0.004 dB, respectively. While the SAM is reduced by 0.204. This clearly demonstrates the superiority of our dynamic mask strategy.

TABLE VI
ABLATION OF MASK GENERATION STRATEGY

Mask strategy	Params (M)	FLOPs (G)	PSNR (dB)	SSIM	SAM
Dynamic	1.03	10.1	35.21	0.951	5.817
Nondynamic	1.03	10.03	35.05	0.947	6.021

TABLE VII
ABLATION STUDY OF DIFFERENT SPARSITY RATE

Sparsity rate	0	0.3	0.4	0.5	0.6	0.7
Params (M)	1.03	1.03	1.03	1.03	1.03	1.03
FLOPs (G)	11.21	10.62	10.79	10.1	9.53	9.21
PSNR (dB)	35.38	35.32	35.28	35.21	35.15	35.07
SSIM	0.953	0.952	0.951	0.951	0.95	0.948
SAM	5.743	5.764	5.809	5.817	5.876	5.988

The dynamic mask generation strategy is capable of adaptive changes. Thus, our dynamic mask generation strategy can efficiently solve this issue.

6) Sparse Rate Selection Ablation: The spatial sparsity rate plays a crucial role in balancing the computational cost and reconstruction performance of the HSI reconstruction models. Excessive sparsity will result in the loss of significant image details, while insufficient sparsity will cause an increase in computational costs. To explore the optimal sparsity rate, we tried a variety of solutions. Specially, we set the sparsity rates to 0, 0.3, 0.4, 0.5, 0.6, and 0.7, respectively. Table VII gives the HSI reconstruction performance with different sparse rates. We can see that the computational cost of the model decreases with an increase in sparsity rate. In this case, the PSNR and SSIM are generally decreasing. And the SAM metric is also reducing. Based on the above-mentioned observations, we set the sparsity rate to 0.5 which can achieve a relatively low computational cost while maintaining relatively high PSNR and SSIM metrics.

7) Patch Size Selection Ablation: In the field of HSI reconstruction, the window size of attention significantly influences the performance of transformer models. Generally speaking, the smaller windows may constrain the model's ability to capture a sufficiently broad range of contextual information. While the larger windows increase the model's parameter count. Thus, to

TABLE VIII
ABLATION STUDY OF DIFFERENT PATCH SIZES

Patch size	4	8	16	32
Params (M)	1.03	1.03	1.03	1.03
FLOPs (G)	10.07	10.1	10.23	10.78
PSNR (dB)	35.04	35.21	35.12	35.27
SSIM	0.948	0.951	0.95	0.951
SAM	5.991	5.817	5.881	5.812

choose a suitable window size, we tested multiple experiments of the patch size with 4, 8, 16, and 32. And the experiment results are presented in Table VIII. From Table VIII, it can be seen that the model's computational costs will rise with the increase of the patch size, while PSNR and SSIM are also improved. In particular, when the patch size increased from 4 to 8, PSNR increased by 0.07 dB, while FLOPs only increased by 0.03 G. However, the PSNR and SSIM decreased with further increases in patch size. Therefore, we decided to set the patch size to 8 in our HSI reconstruction model.

V. CONCLUSION

In this article, we proposed a new HSI reconstruction model called DGST. Our DGST model dynamically guides spatial sparsity through spectral information and aggregates spatial and spectral information to boost the HSI reconstruction. Specifically, we propose a three-level U-Net structure composed of DGSB modules, which utilizes SSHE and SIC as guided priors to enable MSS-MSA well represent the spatial information of HSI. Moreover, we model the spectral and spatial information by aggregating this 2-D information to enhance the spatial–spectral representation ability. Furthermore, the FFN in our model utilizes dual pathways to filter out irrelevant information that can improve the feature extraction ability. Extensive experiments demonstrate that our proposed method outperforms other comparable techniques.

REFERENCES

[1] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

[2] G. Sun et al., "Large kernel spectral and spatial attention networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.

[3] C. Yu et al., "Hyperspectral image classification method based on CNN architecture embedding with hashing semantic feature," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1866–1881, Jun. 2019.

[4] S. Ding, X. Ruan, J. Yang, J. Sun, S. Li, and J. Hu, "LSSMA: Lightweight spectral-spatial neural architecture with multi-attention feature extraction for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 6394–6413, 2024.

[5] R. Xu, X.-M. Dong, W. Li, J. Peng, W. Sun, and Y. Xu, "DBCTnet: Double branch convolution-transformer network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–15, 2024.

[6] P. Addabbo, N. Fiscante, G. Giunta, D. Orlando, G. Ricci, and S. L. Ullo, "Multiple sub-pixel target detection for hyperspectral imaging systems," *IEEE Trans. Signal Process.*, vol. 71, pp. 1599–1611, 2023.

[7] H. Gao, Y. Zhang, Z. Chen, F. Xu, D. Hong, and B. Zhang, "Hyperspectral target detection via spectral aggregation and separation network with target band random mask," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023.

[8] W. Xie, X. Zhang, Y. Li, K. Wang, and Q. Du, "Background learning based on target suppression constraint for hyperspectral target detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5887–5897, 2020.

[9] C. He, Y. Xu, Z. Wu, and Z. Wei, "Connecting low-level and high-level visions: A joint optimization for hyperspectral image super-resolution and target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–16, 2024.

[10] Z. Li, F. Xiong, J. Zhou, J. Lu, and Y. Qian, "Learning a deep ensemble network with band importance for hyperspectral object tracking," *IEEE Trans. Image Process.*, vol. 32, pp. 2901–2914, 2023.

[11] W. Li, Z. Hou, J. Zhou, and R. Tao, "SiamBAG: Band attention grouping-based siamese object tracking network for hyperspectral videos," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, 2023.

[12] Y. Wan, C. Chen, A. Ma, L. Zhang, X. Gong, and Y. Zhong, "Adaptive multi-strategy particle swarm optimization for hyperspectral remote sensing image band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.

[13] P. Duan, X. Kang, P. Ghamisi, and S. Li, "Hyperspectral remote sensing benchmark database for oil spill detection with an isolation forest-guided unsupervised detector," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–11, 2023.

[14] L. Liu, S. Lei, Z. Shi, N. Zhang, and X. Zhu, "Hyperspectral remote sensing imagery generation from RGB images based on joint discrimination," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7624–7636, 2021.

[15] L. Han et al., "Central cohesion gradual hashing for remote sensing image retrieval," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[16] H. Mangotra, S. Srivastava, G. Jaiswal, R. Rani, and A. Sharma, "Hyperspectral imaging for early diagnosis of diseases: A review," *Expert Syst.*, vol. 20, 2023, Art. no. e13311.

[17] S. Karim, A. Qadir, U. Farooq, M. Shakir, and A. A. Laghari, "Hyperspectral imaging: A review and trends towards medical imaging," *Curr. Med. Imag.*, vol. 19, no. 5, pp. 417–427, 2023.

[18] H. Yin and H. Chen, "Multi-branch separable 3D convolutional neural network for hyperspectral image denoising," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 8034–8048, 2023.

[19] X. Cao, H. Du, X. Tong, Q. Dai, and S. Lin, "A prism-mask system for multispectral video acquisition," *IEEE Ttrans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2423–2435, Dec. 2011.

[20] X. Yuan, D. J. Brady, and A. K. Katsaggelos, "Snapshot compressive imaging: Theory, algorithms, and applications," *IEEE Signal Process. Mag.*, vol. 38, no. 2, pp. 65–88, Mar. 2021.

[21] Y. Chen, Y. Wang, and H. Zhang, "Prior image guided snapshot compressive spectral imaging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 11096–11107, Sep. 2023.

[22] X. Zhang, Y. Zhang, R. Xiong, Q. Sun, and J. Zhang, "HerosNet: Hyperspectral explicable reconstruction and optimal sampling deep network for snapshot compressive imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17511–17520.

[23] Y. Dong, D. Gao, T. Qiu, Y. Li, M. Yang, and G. Shi, "Residual degradation learning unfolding framework with mixing priors across spectral and spatial for compressive spectral imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 22262–22271.

[24] L. Li, L. Wang, W. Song, L. Zhang, Z. Xiong, and H. Huang, "Quantization-aware deep optics for diffractive snapshot hyperspectral imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19748–19757.

[25] A. Wagadarikar, R. John, R. Willett, and D. Brady, "Single disperser design for coded aperture snapshot spectral imaging," *Appl. Opt.*, vol. 47, no. 10, pp. B44–B51, 2008.

[26] D. Kittle, K. Choi, A. Wagadarikar, and D. J. Brady, "Multiframe image estimation for coded aperture snapshot spectral imagers," *Appl. Opt.*, vol. 49, no. 36, pp. 6824–6833, 2010.

[27] N. Akhtar and A. Mian, "Hyperspectral recovery from RGB images using Gaussian processes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 100–113, Jan. 2020.

[28] X. Yuan, "Generalized alternating projection based total variation minimization for compressive sensing," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 2539–2543.

[29] S. Zhang, L. Wang, L. Zhang, and H. Huang, "Learning tensor low-rank prior for hyperspectral image reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12001–12010.

[30] 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/CVF Int. Conf. Comput. Vis.*, 2019, pp. 10182–10191.

[31] Y. Liu, X. Yuan, J. Suo, D. J. Brady, and Q. Dai, "Rank minimization for snapshot compressive imaging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 2990–3006, Dec. 2019.

[32] L. Wang, Z. Xiong, G. Shi, F. Wu, and W. Zeng, "Adaptive nonlocal sparse representation for dual-camera compressive hyperspectral imaging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 10, pp. 2104–2111, Oct. 2017.

[33] S. Chen, L. Zhang, and L. Zhang, "MSDformer: Multi-scale deformable transformer for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, 2023.

[34] F. Xiong, J. Zhou, J. Zhou, J. Lu, and Y. Qian, "Multitask sparse representation model inspired network for hyperspectral image denoising," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.

[35] Á. Pérez-García, M. E. Paoletti, J. M. Haut, and J. F. López, "Novel spectral loss function for unsupervised hyperspectral image segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[36] Y. Chang, L. Yan, H. Fang, S. Zhong, and W. Liao, "HSI-DeNet: Hyperspectral image restoration via convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 667–682, Feb. 2019.

[37] Y. Qian, H. Zhu, L. Chen, and J. Zhou, "Hyperspectral image restoration with self-supervised learning: A two-stage training approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022.

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

[39] Z. Xiong, Z. Shi, H. Li, L. Wang, D. Liu, and F. Wu, "HSCNN: CNN-based hyperspectral image recovery from spectrally undersampled projections," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 518–525.

[40] Z. Shi, C. Chen, Z. Xiong, D. Liu, and F. Wu, "HSCNN++: Advanced CNN-based hyperspectral recovery from RGB images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 1052–10528.

[41] L. Wang, C. Sun, Y. Fu, M. H. Kim, and H. Huang, "Hyperspectral image reconstruction using a deep spatial-spectral prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8024–8033.

[42] Y. Fu, T. Zhang, L. Wang, and H. Huang, "Coded hyperspectral image reconstruction using deep external and internal learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3404–3420, Jul. 2022.

[43] T. Li and Y. Gu, "Progressive spatial–spectral joint network for hyperspectral image reconstruction," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

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

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

[46] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 15908–15919.

[47] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.

[48] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 213–229.

[49] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.

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

[51] C. Wang et al., "Translution-SNet: A semisupervised hyperspectral image stripe noise removal based on transformer and CNN," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[52] F. Wang, J. Li, Q. Yuan, and L. Zhang, "Local–global feature-aware transformer based residual network for hyperspectral image denoising," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–19, 2022.

[53] L. Gao, J. Li, K. Zheng, and X. Jia, "Enhanced autoencoders with attention-embedded degradation learning for unsupervised hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–17, 2023.

[54] J. Hu, X. Jia, Y. Li, G. He, and M. Zhao, "Hyperspectral image super-resolution via intrafusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7459–7471, Oct. 2020.

[55] Y. Tang et al., "A CNN-transformer embedded unfolding network for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–16, 2024.

[56] W. Dong, Y. Xu, J. Qu, and S. Hou, "Learning multi-modal cross-scale deformable transformer network for unregistered hyperspectral image super-resolution," in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 1573–1581.

[57] F. I. Alam, J. Zhou, A. W. -C. Liew, X. Jia, J. Chanussot, and Y. Gao, "Conditional random field and deep feature learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1612–1628, Mar. 2019.

[58] Y. Cai et al., "Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17481–17490.

[59] Y. Cai et al., "MST++: Multi-stage spectral-wise transformer for efficient spectral reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 744–754.

[60] Y. Cai et al., "Coarse-to-fine sparse transformer for hyperspectral image reconstruction," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2022, pp. 686–704.

[61] J. Wang, K. Li, Y. Zhang, X. Yuan, and Z. Tao, "Ŝ2-transformer for mask-aware hyperspectral image reconstruction," 2022, *arXiv:2209.12075*.

[62] Y. Cai et al., "Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 37749–37761.

[63] M. Li, Y. Fu, J. Liu, and Y. Zhang, "Pixel adaptive deep unfolding transformer for hyperspectral image reconstruction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 12913–12922.

[64] 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.

[65] L. Zhuang, M. K. Ng, L. Gao, and Z. Wang, "Eigen-CNN: Eigenimages plus eigennoise level maps guided network for hyperspectral image denoising," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–18, 2024.

[66] Y. Xu et al., "Hyperspectral image super-resolution with ConvLSTM skip-connections," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–16, 2024.

[67] X. Miao, X. Yuan, Y. Pu, and V. Athitsos, "λ-net: Reconstruct hyperspectral images from a snapshot measurement," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4058–4068.

[68] Z. Meng, J. Ma, and X. Yuan, "End-to-end low cost compressive spectral imaging with spatial-spectral self-attention," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 187–204.

[69] X. Hu et al., "HDNet: High-resolution dual-domain learning for spectral compressive imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17521–17530.

[70] Y. Chen, X. Gui, J. Zeng, X.-L. Zhao, and W. He, "Combining low-rank and deep plug-and-play priors for snapshot compressive imaging," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2023.3294262.

[71] L. Wang, C. Sun, M. Zhang, Y. Fu, and H. Huang, "DNU: Deep non-local unrolling for computational spectral imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1658–1668.

[72] T. Huang, W. Dong, X. Yuan, J. Wu, and G. Shi, "Deep Gaussian scale mixture prior for spectral compressive imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16211–16220.

[73] Z. Meng, Z. Yu, K. Xu, and X. Yuan, "Self-supervised neural networks for spectral snapshot compressive imaging," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2602–2611.

[74] Z. Wu, R. Lu, Y. Fu, and X. Yuan, "Latent diffusion prior enhanced deep unfolding for spectral image reconstruction," 2023, *arXiv:2311.14280*.

[75] L. Zhu, X. Wang, Z. Ke, W. Zhang, and R. W. Lau, "BiFormer: Vision transformer with bi-level routing attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 10323–10333.

[76] J. Zhang, Y. Zhang, J. Gu, Y. Zhang, L. Kong, and X. Yuan, "Accurate image restoration with attention retractable transformer," 2022, *arXiv:2210.01427*.

[77] J. Xue et al., "Segmentation guided sparse transformer for under-display camera image restoration," 2024, *arXiv:2403.05906*.

[78] X. Chen, H. Li, M. Li, and J. Pan, "Learning a sparse transformer network for effective image deraining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 5896–5905.

[79] S. Liu, J. Ye, S. Ren, and X. Wang, "DynaST: Dynamic sparse transformer for exemplar-guided image generation," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2022, pp. 72–90.
[80] J.-I. Park, M.-H. Lee, M. D. Grossberg, and S. K. Nayar, "Multispectral imaging using multiplexed illumination," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, IEEE, 2007, pp. 1–8.
[81] I. Choi, D. S. Jeon, G. Nam, D. Gutierrez, and M. H. Kim, "High-quality hyperspectral reconstruction using a spectral prior," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 218:1–218:13, 2017.
[82] F. Dell'Acqua, P. Gamba, A. Ferrari, J. A. Palmason, J. A. Benediktsson, and K. Árnason, "Exploiting spectral and spatial information in hyperspectral urban data with high resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 1, no. 4, pp. 322–326, Oct. 2004.
[83] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[84] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts" 2016, *arXiv:1608.03983*.
[85] Z. Meng, S. Jalali, and X. Yuan, "GAP-net for snapshot compressive imaging," 2020, *arXiv:2012.08364*.
[86] J. Yang, T. Lin, F. Liu, and L. Xiao, "Learning degradation-aware deep prior for hyperspectral image reconstruction," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.
[87] J. Ma, X.-Y. Liu, Z. Shou, and X. Yuan, "Deep tensor ADMM-Net for snapshot compressive imaging," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 10222–10231.
[88] X. Qin, Y. Quan, and H. Ji, "Enhanced deep unrolling networks for snapshot compressive hyperspectral imaging," *Neural Netw.*, vol. 174, 2024, Art. no. 106250.
[89] Y. Chen, W. Lai, W. He, X.-L. Zhao, and J. Zeng, "Hyperspectral compressive snapshot reconstruction via coupled low-rank subspace representation and self-supervised deep network," *IEEE Trans. Image Process.*, vol. 33, pp. 926–941, 2024.
[90] Z. Pan, H. Zeng, J. Cao, K. Zhang, and Y. Chen, "DiffSCI: Zero-shot snapshot compressive imaging via iterative spectral diffusion model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 25297–25306.
[91] Y. Cai, Y. Zheng, J. Lin, X. Yuan, Y. Zhang, and H. Wang, "Binarized spectral compressive imaging," in *Proc. 37th Conf. Neural Inf. Process. Syst.*, 2023, pp. 38335–38346.

Junyang Wang received the B.S. degree in software engineering from the Henan University of Science and Technology, Luoyang, China, in 2018. He is currently working toward the master's degree in electronic information with the Xidian University, Xi'an, China.

His research interests include deep learning and image processing.

Xiang Yan (Member, IEEE) received the B.S and Ph.D. degrees in electronic science and technology and physical electronics from Xidian University, Xi'an, China, in 2012 and 2018, respectively.

He is currently an Associate Professor with Xidian University, Xi'an, China. From 2016 to 2018, he was a Visiting Ph.D. Student with the School of Computer Science and Software Engineering, University of Western Australia, Crawley, WA, Australia, working closely with Prof. Ajmal Mian. His research interests include image processing, computer vision, and deep learning.

Hanlin Qin (Member, IEEE) received the B.S and Ph.D. degrees in electronic information engineering and electronic science and technology from Xidian University, Xi'an, China, in 2004 and 2010, respectively.

He is currently a Full Professor with the School of Optoelectronic Engineering, Xidian University. He has authored or coauthored more than 100 scientific articles. His research interests include electro-optical cognition, advanced intelligent computing, and autonomous collaboration.

Naveed Akhtar (Member, IEEE) received the master's degree in computer science from Hochschule Bonn-Rhein-Sieg, Sankt Augustin, Germany, and the Ph.D. degree in computer science from the University of Western Australia, Crawley, WA, Australia.

He is a Senior Lecturer with the University of Melbourne, Melbourne, VIC, Australia.

Dr. Akhtar was the recipient of the Discovery Early Career Researcher Award from the Australian Research Council. He is a Universal Scientific Education and Research Network Laureate in formal sciences. He was a finalist of the Western Australia's Early Career Scientist of the Year 2021. He is an ACM Distinguished Speaker and an Associate Editor for IEEE TRANSACTIONS NEURAL NETWORKS AND LEARNING SYSTEMS.

Shuowen Yang (Member, IEEE) received the B.S. degree in electronic science and technology and the Ph.D. degree in physical electronics from Xidian University, Xi'an, China, in 2016 and 2023, respectively.

He is currently a Postdoctoral Researcher with the Department of Photo-Electronic Information, Xidian University. From 2021 to 2022, during his Ph.D. study, he visited the University of Granada, Granada, Spain, working closely with Prof. Rafael Molina. His research interests include computational spectral imaging and compressive sensing reconstruction.

Ajmal Mian (Senior Member, IEEE) received the B.S. degree in avionics from NED University, Karachi, Pakistan, in 1993; the M.S. degree in information security from NUST, Karachi, Pakistan, in 2003; and the Ph.D. degree in computer science from UWA, Perth, Austrian, in 2006.

He is a Professor of computer science with The University of Western Australia, Crawley, WA, Australia. His research interests include computer vision, machine learning, remote sensing, and 3-D point cloud analysis.

Dr. Mian was the recipient of several awards including the West Australian Early Career Scientist of the Year Award 2012, the HBF Mid-Career Scientist of the Year Award 2022, Excellence in Research Supervision Award, EH Thompson Award, ASPIRE Professional Development Award, Vice-chancellors Mid-career Research Award, Outstanding Young Investigator Award, and the Australasian Distinguished Doctoral Dissertation Award, and three esteemed national fellowships from the Australian Research Council (ARC) including the recent Future Fellowship Award 2022. He has secured research funding from the ARC, NHMRC, DARPA, and the Australian Department of Defence. He was a Senior Editor for IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS and Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING AND THE PATTERN RECOGNITION JOURNAL. He is a Fellow of the International Association for Pattern Recognition.