

# Spectral-Enhanced Sparse Transformer Network for Hyperspectral Super-Resolution Reconstruction

Yuchao Yang<sup>ID</sup>, Student Member, IEEE, Yulei Wang<sup>ID</sup>, Member, IEEE, Hongzhou Wang,  
Lifu Zhang<sup>ID</sup>, Senior Member, IEEE, Enyu Zhao<sup>ID</sup>, Member, IEEE, Meiping Song<sup>ID</sup>,  
and Chunyan Yu<sup>ID</sup>, Senior Member, IEEE

**Abstract**—Hyperspectral image (HSI) has garnered increasing attention due to its capacity for capturing extensive spectral information. However, the acquisition of high spatial resolution HSIs is often restricted by current imaging hardware limitations. A cost-effective approach to enhance spatial resolution involves fusing HSIs with high spatial resolution multispectral images collected from the same scene. Traditional convolutional neural network-based models, although gained prominence in HSI super-resolution reconstruction, are typically limited by their small receptive field of the convolutional kernel, primarily emphasizing local information while neglecting nonlocal characteristics of the image. In light of these limitations, this article proposes a novel spectral-enhanced sparse transformer (SEST) network for HSI super-resolution reconstruction. Specifically, the proposed SEST employs a sparse transformer to capture nonlocal spatial similarities efficiently, along with a spectral enhancement module to learn and exploit spectral low-rank characteristics. Integrated within a multiwindow residual block, the abovementioned two components collaboratively extract and combine distinct fine-grained features through a weighted linear fusion process, facilitating the integration of spatial and spectral information to optimize the reconstruction result. Experimental results validate the superior performance of the proposed SEST model against current state-of-the-art methods in both visual and quantitative metrics, thus confirming the effectiveness of the proposed approach.

**Index Terms**—Hyperspectral image (HSI), multiwindow residual block, nonlocal information, sparse transformer, spectral enhancement, super-resolution.

## I. INTRODUCTION

HYPERSPECTRAL imaging systems have the capability to concurrently capture surface information across hundreds

Received 11 June 2024; revised 8 July 2024 and 1 September 2024; accepted 3 September 2024. Date of publication 10 September 2024; date of current version 2 October 2024. This work was supported in part by the National Key R&D Program of China under Grant 2022YFE0904400, in part by the National Nature Science Foundation of China under Grant 42271355 and Grant 61801075, in part by the Natural Science Foundation of Liaoning Province (2022-MS-160), and in part by the Fundamental Research Funds for the Central Universities under Grant 3132024234. (*Corresponding authors: Yulei Wang; Lifu Zhang.*)

Yuchao Yang, Yulei Wang, Hongzhou Wang, Enyu Zhao, Meiping Song, and Chunyan Yu are with the Center of Hyperspectral Imaging in Remote Sensing, Information Science and Technology College, Dalian Maritime University, Dalian 116026, China (e-mail: yyc@dlnu.edu.cn; wangyulei@dlnu.edu.cn; whz1579@163.com; zhaoenyu@dlnu.edu.cn; smping@163.com; yuchunyan1997@126.com).

Lifu Zhang is with the National Engineering Research Center for Satellite Remote Sensing Applications, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: zhanglf@aircas.ac.cn).

The code of the SEST model will be available at <https://github.com/YuleiWang1/SEST>.

Digital Object Identifier 10.1109/JSTARS.2024.3457814

of continuous bands, yielding a set of spectral images depicting the same scene [1], [2]. The primary advantage of hyperspectral images (HSIs) over traditional natural or multispectral images lies in their richer spectral information, which enables accurate distinction and identification of objects within the image scene, making them widely applicable in various fields such as image classification [3], [4], target detection [5], [6], and change detection [7]. However, due to inherent limitations in incident energy and imaging hardware, HSIs often suffer from low-spatial resolution (named as LR-HSI), which significantly restricts their practical applications [8]. Unlike hardware upgrades, hyperspectral super-resolution reconstruction (HSI-SR), as an image postprocessing technique, obtains high spatial resolution hyperspectral images (HR-HSIs) from an algorithmic perspective. Due to its low cost and high efficiency, HSI-SR has become a necessary and promising research direction.

Generally, HSI-SR techniques can be divided into two categories based on whether auxiliary information is required: single HSI super-resolution and fusion-based HSI super-resolution. The former is highly independent and easy to implement, but since HSI-SR is an ill-posed problem, relying solely on a single LR-HSI can result in reconstructed images lacking detailed information. The latter introduces HR-MSIs of the same scene as auxiliary data, leveraging the advantages of different data sources to achieve better reconstruction results [9].

Over the past few decades, the HSI-SR domain has rapidly developed, with many fusion-based methods emerging. These methods can be broadly categorized into model-based and deep learning (DL)-based approaches. Model-based methods typically involve manually constructing various priors (e.g., self-similarity [10], sparsity [11], [12], and low rank [13], [14]) as regularizers to achieve reconstruction. While these methods show commendable performance, they suffer from issues such as being time-consuming and having limited representational capacity due to their reliance on manually crafted priors. With the rapid development of DL techniques, especially convolutional neural networks (CNNs), DL-based methods have demonstrated impressive performance [15], [16], [17], [18], [19]. These methods are inherently data-driven, allowing networks to autonomously learn priors from the characteristics of the dataset itself, thus offering greater flexibility [20]. However, the fixed receptive field of CNNs due to convolution kernel size makes them inefficient in modeling long-range dependencies, which somewhat limits their fusion performance [21].

To address these issues, the self-attention mechanism from the natural language processing (NLP) field has garnered increasing attention for its outstanding global feature extraction capabilities. Especially, the vision transformer (ViT) [22] introduced self-attention mechanisms into the computer vision (CV) field for the first time by partitioning images into patches and adding positional embeddings. Its excellent global context modeling capability effectively addresses edge effects in HSI-SR tasks. However, the global self-attention mechanism of ViT exhibits quadratic computational complexity with respect to input image size, leading to increased GPU memory demands. To mitigate these limitations, the Swin Transformer [23] uses a hierarchical structure and shifted window mechanism to reduce the length of the input sequence, making the self-attention mechanism a more versatile backbone network. Nonetheless, it still faces challenges such as fixed window sizes and high computational and memory demands when processing high-resolution images. The Performer [24] approximates the softmax function using orthogonal random feature mapping, thereby altering the order of matrix computations in self-attention to achieve linear complexity. The sparse transformer [25] restricts each element to interact only with a subset of elements in the sequence, thus sparsifying the attention score matrix.

Motivated by these developments, especially the sparse transformer and channel attention mechanism [26], this article constructs the spectral-enhanced sparse transformer (SEST), a novel network architecture for HSI super-resolution reconstruction. Specifically, in the spatial domain, sparse self-attention is used to model long-range dependencies, complemented by a local enhanced feed-forward network (LeFF) to retain complex local details, thereby achieving more efficient local-global feature learning. In the spectral domain, a spectral enhancement module is designed to explore the correlations between adjacent bands of HSIs and integrate them into the transformer structure, effectively preserving the original spectral information while promoting the interaction between spatial and spectral information, thus enhancing the network's ability to learn critical features. Incorporating these two designed elements, a multiwindow residual block is employed to learn multiscale features from the input image, while long and short skip connections are added to the network, contributing to the flexibility of information flow and the robustness of the network. The primary contributions of this article are as follows.

- 1) A novel SEST-based super-resolution reconstruction network is proposed, effectively leveraging nonlocal spatial similarity and the spectral low-rankness inherent in HSIs.
- 2) A multiwindow residual block is specifically designed to extract features at different levels of granularity. The incorporation of a weighted linear combination facilitates the fusion of these features, contributing to an enhancement in the quality of the reconstructed image.
- 3) The implementation of a spectral enhancement module in the self-attention calculation stage to boost the network's capability of spectral information extraction, facilitating the recalibration of the self-attention map to activate more pixels. The incorporation of LeFF instead of the standard FFN further enhances the network's capacity to exploit local contextual details.

The rest of this article is organized as follows. Section II reviews related work in the domain of HSI-SR, Section III clarifies the proposed SEST architecture along with the loss function, Section IV presents the experimental validation of the proposed method, and Section V concludes this article.

## II. RELATED WORKS

This section reviews some of the most notable recent advancements in HSI-SR techniques, categorizing them into three types: model-based methods, CNN-based methods, and ViT-based methods. Additionally, the limitations of existing ViT-based methods are analyzed in detail.

### A. Model-Based Methods

Model-based methods, a classical approach to HSI-SR, can be divided into three primary categories. The first category includes techniques based on panchromatic sharpening, such as component substitution (CS) and multiresolution analysis. A widely employed method within this category is the adaptive Gram–Schmidt algorithm proposed by Aiazzi et al. [27], which incorporates the spectral response function into CS. Another notable method, proposed by Selva et al. [28], involves a super-resolution framework utilizing linear regression to represent each hyperspectral band as a linear combination of multispectral bands. While computationally efficient, these methods often yield unreliable quality of the reconstructed image. The second category consists of methods relying on prior information analysis of HSI, including sparse representation-based and Bayesian-based methods. Specifically, Akhtar et al. [29] proposed a Bayesian framework based on sparse representation, deriving the probability distribution of the spectral bases and computing sparse coding of the high-resolution image. Wei et al. [30] integrated the explicit solution of the Sylvester equation into the Bayesian-based method, named “fast fusion based on Sylvester equation” (FUSE), significantly reducing the algorithm complexity while ensuring the quality of the reconstructed image. Simões et al. [31] introduced total variation regularization for effective edge preservation in a convex optimization of subspace coefficients. The third category involves decomposition-based methods, with considerable attention of being explainable and understandable, where the most representative methods are matrix factorization-based methods. For instance, Yokoya et al. [32] proposed a coupled nonnegative matrix factorization (CNMF) method, employing CNMF to alternately estimate endmembers and abundances. However, these methods cannot effectively preserve the spectral structure of HSIs. Addressing that, tensor decomposition-based methods have been explored, such as the nonlocal sparse tensor decomposition-based method proposed by Dian et al. [33], estimating the sparse kernel tensor and dictionary for HSI-SR, showcasing potential in preserving spectral information for high-quality reconstruction.

### B. CNN-Based Methods

The burgeoning interest in DL has led to the rapid development of CNN-based methods for HSI-SR. For example, Li et al. [34] proposed an X-shaped interactive autoencoder

network, integrating the concept of matrix factorization into DL to facilitate cross-modal learning between hyperspectral and multispectral data. To inject more texture details into HSI, the IFMSR [35] integrates an RGB-induced detail enhancement and a deep cross-modal feature modulation module. While demonstrating efficacy in various scenarios, it is noteworthy that these models are primarily data-driven, with a lack of interpretability. In response, Xie et al. [16] proposed a model-based DL method using a deep unfolding network inspired by the traditional alternating iterative algorithm model, employing CNN to learn the proximal operator and model parameters. The aforementioned methods are all based on 2D convolution, failing to effectively model the spectral structure of HSIs with 3D characteristics. Mei et al. [36] introduced 3D fully convolutional networks into HSI-SR tasks, allowing the network to better learn both spatial and spectral information in HSI. However, due to the high spectral resolution of HSI, methods based on 3D CNNs suffer from challenges of large parameter size and high computational complexity. In response to these issues, Li et al. [37] considered the spectral similarity of HSI and utilized the similarity between bands to achieve grouping, thereby reducing the computational cost of the network. Li et al. [38], on the other hand, addressed computational efficiency from a network structure perspective by designing separable 3D convolutions, aiming to mitigate the computational burden while preserving the spatial and spectral separability. Furthermore, the ill-posed nature of the HSI-SR task poses significant challenges for single-stage learning. To address this, Li et al. [39] designed a coarse-to-fine dual-stage learning framework. In the coarse stage, a symmetric feature propagation model is utilized for broader feature extraction. In the fine stage, a back-projection refinement network is introduced to learn specific features of the image.

### C. Visual Transformer-Based Methods

The transformer network, first proposed by Vaswani et al. [40], quickly gained widespread attention due to its core component, the self-attention mechanism, which has powerful global context modeling capabilities. Dosovitskiy et al. [22] are the first to introduce transformer into the CV field, dividing input image into nonoverlapping patches to generate sequence elements, which are then fed into the transformer model for image recognition. Following the success of this work, transformers have been widely applied in various advanced visual tasks, such as object detection [41], [42], image classification [43], and image segmentation [44]. Among these, the Swin Transformer proposed by Liu et al. [23] is particularly noteworthy. It restricts attention computation to local windows and enhances the network's ability to capture contextual information effectively through shift operations and a hierarchical architecture, significantly reducing the computational cost of self-attention. Inspired by this, numerous transformer-based image reconstruction methods have emerged. For instance, SwinIR [45] employs Swin Transformer-based residual blocks to extract deep features from images, showcasing the immense potential of transformers in image reconstruction. In the HSI-SR domain, Fusformer [46] made the first attempt to use ViT encoders as the main body of

the network to explore the spatial and spectral information of HSIs, achieving promising results. However, utilizing a single module to simultaneously model spatial and spectral features increases the complexity of network learning. To address this issue, Long et al. [47] employed Swin Transformer to design spatial and spectral self-attention blocks, cascading them to obtain global spatial features and spectral sequence information, contributing to a more efficient and effective HSI super-resolution reconstruction process. Although the design of the shifted window can reduce computational costs, it also weakens the interaction of image boundary information. To address this, Deng et al. [48] proposed a Pyramid Shuffle-and-Reshuffle Transformer (PSRT) method, which employs shuffle techniques to achieve long-range interaction between patches. Despite the remarkable results achieved by these methods, their approaches to acquiring global information remain inefficient, particularly when dealing with hyperspectral data with high redundancy. Additionally, existing methods often overlook the interaction between spatial and spectral information. Moreover, relying solely on single-type feature modeling is not conducive to the fine reconstruction of images.

## III. PROPOSED METHOD

Drawing inspiration from sparse transformer and channel attention mechanisms, this section proposed a novel SEST network, specifically tailored for HSI-SR tasks. The overall network architecture of the proposed SEST, along with its hierarchical structure, is presented. Additionally, a comprehensive explanation of the key component of SEST, the spectral-enhanced sparse transformer residual layer (SSRL) is provided.

### A. Network Architecture

In HSI-SR tasks, the attainment of a larger receptive field is often crucial for achieving superior reconstruction results. However, conventional CNN architectures, constrained by the inherent limitations of convolutional operations, tend to exhibit deficiencies in modeling long-range dependencies effectively. Recognizing the potential of transformers in addressing this limitation, a SEST network is proposed, specifically designed to simultaneously explore nonlocal spatial similarities and spectral low-rank characteristics inherent in HSIs. The network structure of the proposed SEST method is depicted in Fig. 1.

The process begins with two images, the LR-HSI and the HR-MSI of the same observed scene, represented by  $Y \in \mathbb{R}^{h \times w \times L}$  and  $Z \in \mathbb{R}^{H \times W \times l}$ , respectively, where  $W(w)$  and  $H(h)$  represent the width and height of the spatial dimension, and  $L(l)$  denotes the number of spectral bands in the image ( $w \ll W$ ,  $h \ll H$ , and  $l \ll L$ ). The objective of super-resolution reconstruction is to estimate HR-HSI, represented by  $X \in \mathbb{R}^{H \times W \times L}$ , with both high spatial and high spectral resolution from these two images. Commonly, two strategies are employed for LR-HSI and HR-MSI fusion: image-domain and feature-domain concatenation. To better preserve the spatial and spectral details in the original image, this article adopts the image-domain concatenation framework. Initially, the LR-HSI data undergoes bicubic interpolation during the input data preprocessing stage to obtain

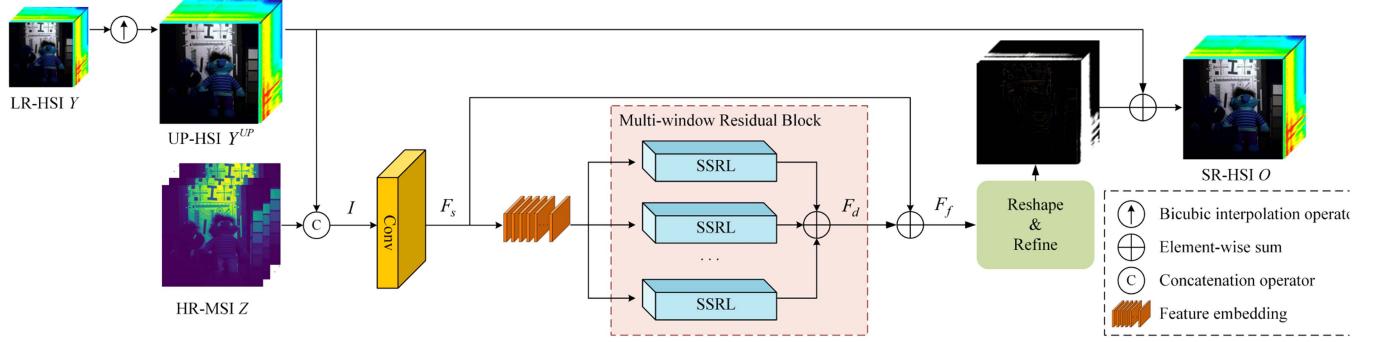


Fig. 1. Overall structure of the proposed SEST network.

the up-sampled image  $Y^{up} \in \mathbb{R}^{H \times W \times L}$ , alleviating the learning burden on the model [49]. Subsequently, the two images ( $Y^{up}$  and HR-MSI) are concatenated along the spectral dimension to form the network input  $I \in \mathbb{R}^{H \times W \times (L+l)}$ . Knowing the fact that deep features contain more semantic information to enhance the realistic texture of the reconstructed image, while shallow features preserve more accurate details and textures, which is crucial for peak signal-to-noise ratio (PSNR)-oriented models, the proposed SEST begins to extract shallow features  $F_s \in \mathbb{R}^{H \times W \times C}$  by employing a  $3 \times 3$  convolution layer, learning the details and textures in the fused image. While for deep feature extraction, to better leverage the transformer network's exceptional sequence data modeling capabilities [40], this article capitalizes on the unique advantages of high spectral resolution of HSI by designing a specialized feature embedding layer. This layer splits the feature map  $F_s$  into individual pixel vectors, enabling the serialization of HSI with an emphasis on spectral information. These vectors are then fed in parallel into multiwindow residual blocks to better capture the multiscale spectral-spatial features of the image. The multiwindow residual blocks consist of SSRLs with different window sizes. Subsequently, these multiscale spectral-spatial features are fused together using a weighted linear combination as the semantic information  $F_d \in \mathbb{R}^{H \times W \times C}$  extracted by the network from the fused image. In the spatial information exploration phase, pixel vectors serve as inputs to effectively preserve the original spectral structure. On this basis, a spectral-enhanced (SE) module is introduced to generate weight coefficients for different channels in the window blocks, facilitating the activation of more pixels in the self-attention matrix calculation. Global skip connections are then employed to combine  $F_s$  and  $F_d$  to obtain  $F_f \in \mathbb{R}^{H \times W \times C}$ , thereby enhancing the robustness of the network, reducing training difficulty, extracting finer high-frequency details, and reducing spectral distortion during the spatial feature extraction process. Finally, the image reconstruction block reduces the number of feature channels in  $F_f$  to the number of spectral bands and adds it to  $Y^{up}$  to yield the final reconstructed result  $O \in \mathbb{R}^{H \times W \times L}$ .

### B. SE Sparse Transformer Residual Layer

Applying transformers to HSI-SR tasks faces three primary challenges. First, unlike RGB images where spatial processing often suffices, HSI-SR tasks require careful handling of

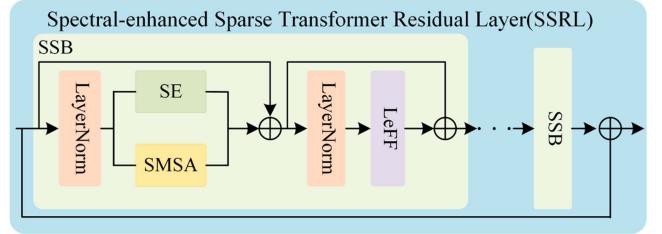


Fig. 2. Structure of SSRL.

rich spectral information crucial for applications such as classification and object detection. Therefore, ensuring that the reconstructed spectrum remains undistorted while promoting effective interaction between spatial and spectral information is quite challenging. This aspect is often overlooked by existing transformer networks, which predominantly focus on spatial attributes. Second, the foundational mechanism of Vanilla Transformer involves computing global self-attention across all tokens, facilitating the modeling of long-distance dependencies, but it also leads to a quadratic growth in complexity relative to the number of tokens, making it impractical for super-resolution reconstruction tasks involving high image resolution. Finally, local contextual information is valuable for capturing image details and textures, providing more semantic information for a better understanding of objects and structures in the image. While essential for image super-resolution tasks, previous work has demonstrated limitations in transformer's ability to capture local dependencies.

To address these challenges, an SSRL is specifically designed, as depicted in Fig. 2, which leverages the advantages of sparse transformer to model long-distance dependencies with a lower computational cost, while at the same time, depth-wise convolutional operators and SE modules are integrated to capture useful local contextual information and spectral features, respectively.

The process can be described as follows:

$$F_l = H_{SSB_l}(F_{l-1}), l = 1, 2, \dots, L \quad (1)$$

$$F_{out} = H_{SSB_L}(F_L) + F_s \quad (2)$$

where  $H_{SSB_l}(\cdot)$  denotes the  $l$ th SE sparse transformer block, and  $F_l$  and  $F_{l-1}$  represent its output and input, respectively.

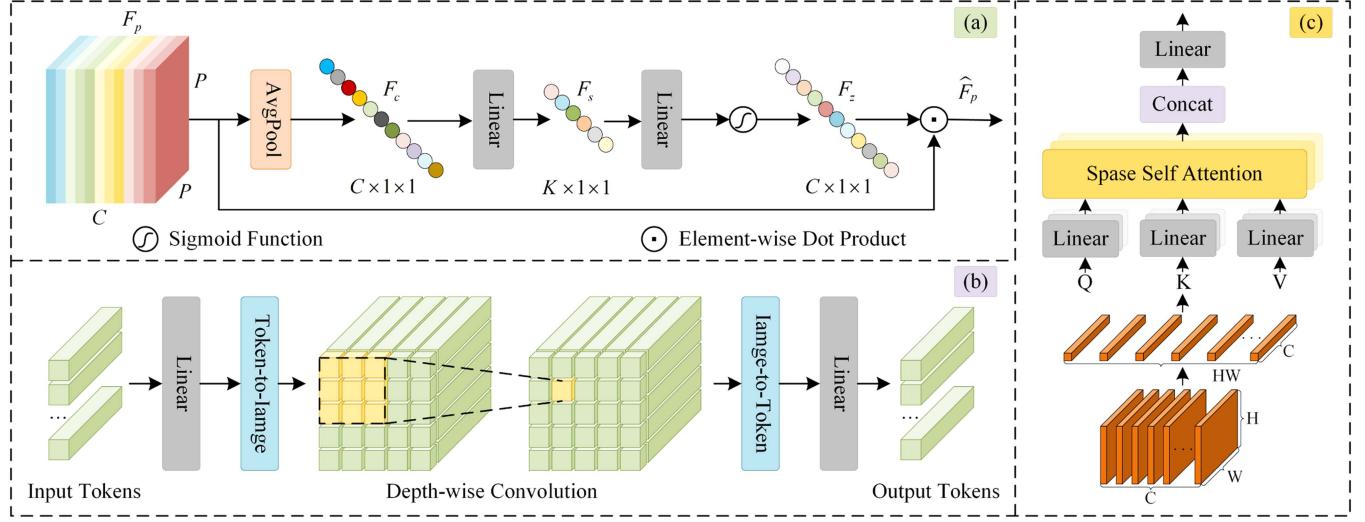


Fig. 3. Modules of the proposed SSB, with (a) SE module (SE), (b) SMSA module, and (c) LeFF.

For a single spectral-enhanced sparse transformer block (SSB), as illustrated in Fig. 3, three core designs are adopted as follows: SE module, sparse multihead self-attention (SMSA), and LeFF.

The computation of a single SSB can be described as follows:

$$F_N = LN(F) \quad (3)$$

$$F_M = SMSA(F_N) + \alpha SE(F_N) + F \quad (4)$$

$$F' = LeFF(LN(F_M)) + F_M \quad (5)$$

where  $F_M$  and  $F'$  denote the outputs of the hybrid attention module and the LeFF module, respectively,  $LN(\cdot)$  denotes layer normalization, and  $SE(\cdot)$  represents the SE module. The parameter  $\alpha$  is an adaptive weight coefficient used to balance the two attention modules.

1) *Spectral Enhancement*: Given that the spectra in HSI usually exhibit low-rank properties due to high correlation among different spectral bands, which have been proven to be useful for guidance in HSI tasks such as denoising, compressive sensing, and unmixing [50], this module aims to leverage these properties for efficient spectral handling. Knowing that the biggest bottleneck in the HSI super-resolution task lies in designing an appropriate regularization method to map the low-resolution HSI into the proper subspace, inspired by the channel attention mechanism, a SE module is introduced into the self-attention calculation module of the Vanilla Transformer to facilitate the automatic learning of appropriate representations in the subspace.

Specifically, as depicted in Fig. 3(a), the input feature  $F \in \mathbb{R}^{C \times H \times W}$  is initially divided into nonoverlapping data cubes, denoted as  $F_p \in \mathbb{R}^{C \times P \times P}$ . Subsequently, average pooling is applied to  $F_p$  to aggregate its channel features and obtain the mapped spectral vector  $F_c \in \mathbb{R}^{C \times 1 \times 1}$ , improving the model efficiency and feature stability. Finally, a linear layer is used to compress its channel dimension to obtain  $F_s \in \mathbb{R}^{K \times 1 \times 1}$ , aiming to map the spectral vector into a suitable low-rank subspace.

This process can be described as follows:

$$F_c = Avgpool(F_p) \quad (6)$$

$$F_s = W_c F_c \quad (7)$$

where  $W_c$  is the weight of the linear layer and  $Avgpool(\cdot)$  denotes the average pooling layer.

It is noteworthy that these operations all act on the internal information of each data cube, thereby primarily focusing on learning spectral statistical information between adjacent pixels. Given that  $F_s$  contains rich spectral statistical information, a linear layer is utilized to scale the obtained low-rank vector to match the dimensions of the input spectral vector. Subsequently, these vectors are fed into the sigmoid function to convert into weight coefficients  $F_z$ , which serve as guidance to recalibrate the input data cube  $F_p$ , thus enhancing spatial-spectral correlation and promoting the super-resolution process. This process can be described as follows:

$$\hat{F}_p = F_p \cdot F_z = F_p \cdot Sigmoid(W_z F_s) \quad (8)$$

where  $W_z$  is the weight of the linear layer and “ $\cdot$ ” denotes the element-wise dot product.

2) *Sparse Multihead Self-Attention*: The inherent quadratic computational complexity of the Vanilla Transformer poses substantial challenges for practical applications. To address this issue, innovations in the field of NLP have led to the development of two improved self-attention structures: linear self-attention and sparse self-attention. Leveraging the established principle that the attention matrix naturally exhibits mathematical sparsity, sparse self-attention strategies can effectively reduce computational cost and enhance model efficiency by pruning or distilling the attention matrix. Inspired by this, as shown in Fig. 3(b), the input HSI data is initially divided into individual pixel vectors by drawing upon the characteristics of high spectral resolution of HSI. Subsequently, due to the local spatial similarity of HSI, several nonoverlapping windows of the same size are employed to partition these pixel vectors,

obtaining  $F_p^i \in \mathbb{R}^{P^2 \times C}$  by flattening and transposing the data within each window. Finally, self-attention is computed on the flattened features, and the outputs of all attention heads are concatenated and linearly projected to obtain the final result. The  $k$ th self-attention head can be described as follows:

$$F = \{F_p^1, F_p^2, \dots, F_p^N\}, N = HW/P^2 \quad (9)$$

$$SA_k^i = \text{Attention}\left(F_p^i W_k^Q, F_p^i W_k^K, F_p^i W_k^V\right), i = 1, 2, \dots, N \quad (10)$$

$$\hat{F}_k = \{SA_k^1, SA_k^2, \dots, SA_k^N\} \quad (11)$$

where  $W_k^Q$ ,  $W_k^K$ , and  $W_k^V$  in (10) denote the projection matrix of the *query*, *key*, and *value* (Q, K, V), respectively.  $\hat{F}_k$  in (11) represents the output of the  $k$ th self-attention head. It is worth noting that the self-attention computation here is restricted within each block.

Although this approach shares similarities with the nonoverlapping window-based multihead self-attention mechanism employed in ViT, the purposes are significantly different. In this work, the approach is primarily employed to encourage the model to learn a sparser attention matrix. In contrast to global self-attention, this strategy reduces the computational complexity for a given input feature map  $F \in \mathbb{R}^{C \times H \times W}$  from  $O(H^2 W^2 C)$  to  $O(P^2 HWC)$ .

3) *Local Enhanced Feed-Forward Network*: The Vanilla Transformer architecture comprises a multihead self-attention (MSA) module and a feed-forward network (FFN). While the MSA module calculates correlations between tokens and performs linear fusion to achieve global modeling, the FFN, consisting of a simple multilayer perceptron, performs nonlinear transformations on features to enhance their representation capability. However, the conventional FFN designed in most transformer models often neglects crucial neighboring spatial information for images [51]. To overcome this challenge, the designed LeFF modifies the traditional FFN by incorporating a depth-wise convolution block. Specifically, as shown in Fig. 3(c), a linear projection layer is first applied to each token to increase its feature dimension. The projected tokens are then spatially reshaped according to their original positions and a  $3 \times 3$  depth-wise convolution is performed on each channel of the reshaped features to better capture local spatial contextual information. Finally, the features are restored to tokens, and the channel dimension is matched with the input through another linear projection, which serves as the final output.

### C. Loss Function

In the reconstruction process, the most crucial aspect is restoring the high-frequency details, which encapsulate critical spatial information that is normally lost during lower-resolution image acquisition processes. To effectively restore these details, the mean absolute error (MAE) is employed in this article as a primary loss function. The choice of MAE is driven by its sensitivity to minor discrepancies with a better convergence of the network. By minimizing the MAE between reconstructed images and ground truth, the network learns to accurately reconstruct the

spatial information, thereby enhancing the overall fidelity of the reconstructed HR-HSI. The MAE is depicted as follows:

$$L_{MAE}(\theta) = \frac{1}{M} \sum_{m=1}^M \|O^m - X^m\|_1 \quad (12)$$

where  $O^m$  and  $X^m$  are the  $m$ th reconstructed HR-HSI and the ground truth, respectively.  $M$  is the number of images in a training batch, and  $\theta$  denotes the parameter set of the network.

To address the critical challenge of spectral distortion in HSI super-resolution, a spatial-spectral total variation (SSTV) loss, as initially proposed in [52], is introduced as another loss. The SSTV loss is particularly designed to minimize artifacts and ensure fidelity in both spatial and spectral domains, which is crucial for maintaining the essential characteristics of the original scene. The mathematical representation of SSTV loss is expressed as follows:

$$L_{SSTV}(\theta) = \frac{1}{M} \sum_{m=1}^M (\|\nabla_h O^m\|_1 + \|\nabla_w O^m\|_1 + \|\nabla_l O^m\|_1) \quad (13)$$

where  $\nabla_h$ ,  $\nabla_w$ , and  $\nabla_l$  denote the gradient functions of the computed horizontal, vertical, and spectral dimensions, respectively.

The final loss function is a composite loss function, taking into account both MAE and SSTV, expressed as follows:

$$L(\theta) = L_{MAE}(\theta) + \beta L_{SSTV}(\theta) \quad (14)$$

where  $\beta$  is the tradeoff parameter, which is used to adjust the weight between space and spectral reconstruction errors.

## IV. EXPERIMENTAL RESULTS

This section presents the experimental results to demonstrate the effectiveness of the proposed method. Initially, the experimental configurations are introduced, encompassing descriptions of the utilized datasets, data simulation procedures, and implementation details. Following this, the reconstruction performance on three public datasets is illustrated and compared against the state-of-the-art algorithms, supplemented by a concise analysis. Finally, an ablation study is provided to validate the effectiveness of the proposed method.

### A. Experimental Configurations

1) *Datasets*: Experiments are conducted on three publicly available hyperspectral datasets, the CAVE dataset [53], the Harvard dataset [54], and the Washington DC Mall (WDC) dataset [32]. The CAVE dataset was captured by a cooled charge-coupled device camera and consists of 32 different indoor scenes, with each HSI presenting a spatial resolution of  $512 \times 512$  pixels and encompassing 31 spectral bands at 10 nm intervals in the range of 400–700 nm. The Harvard dataset, captured by Nuance FX and CRI INC cameras, contains 77 real indoor and outdoor scenes, with each HSI presenting a spatial resolution of  $1024 \times 1392$  pixels and encompassing 31 spectral bands at 10 nm intervals in the range of 420–720 nm. The WDC dataset was captured by the HYDICE sensor and contains one

TABLE I  
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON CAVE, HARVARD, AND WDC DATASETS

Dataset	Metrics	CNMF	Hysure	FUSE	SSRNet	Fusformer	PSRT	SEST
CAVE	PSNR	41.79	43.09	39.68	47.36	<u>48.56</u>	47.24	<b>51.11</b>
	SAM	6.99	6.03	4.97	3.13	<u>2.52</u>	2.70	<b>1.81</b>
	SSIM	0.9778	0.9786	0.9791	0.9931	<u>0.9953</u>	0.9947	<b>0.9971</b>
	ERGAS	3.37	2.58	3.88	1.97	<u>1.30</u>	1.59	<b>1.04</b>
Harvard	PSNR	45.09	44.76	42.70	41.55	44.42	<u>45.36</u>	<b>45.80</b>
	SAM	2.61	2.51	2.69	5.10	2.66	<u>2.26</u>	<b>2.25</b>
	SSIM	0.9785	0.9773	0.9705	0.9539	<u>0.9841</u>	0.9783	<b>0.9843</b>
	ERGAS	<b>2.06</b>	2.36	2.53	13.69	2.48	<u>2.23</u>	2.35
WDC	PSNR	46.00	45.33	43.90	46.32	37.44	<u>47.09</u>	<b>48.26</b>
	SAM	4.98	4.62	6.22	<u>2.87</u>	9.56	3.40	<b>2.59</b>
	SSIM	0.9555	0.9557	0.9454	0.9341	0.7511	<u>0.9760</u>	<b>0.9640</b>
	ERGAS	<b>4.44</b>	4.94	5.41	17.83	21.12	<u>4.83</u>	15.75

The best values are highlighted in bold, and the second-best values are underlined.

HSI of a large urban area, which presents a spatial resolution of  $1280 \times 307$  pixels and encompasses 191 spectral bands with a range of 400–2400 nm.

2) *Data Simulation:* In this article, considering that both the CAVE and Harvard datasets have 31 bands and similar coverage ranges, 20 images from the CAVE dataset are selected for the training set. The remaining 11 images, along with 9 randomly selected images from the Harvard dataset, are set aside as the test set to evaluate the network's generalization ability. For the WDC dataset, four  $128 \times 128$  images are cropped from the original image for testing network performance, while the rest are used for training. Due to the limited number of training samples, the training images from the CAVE dataset are segmented into 4275 overlapping image patches of size  $64 \times 64 \times 31$ . These overlapping patches are then downsampled to a spatial resolution of  $16 \times 16 \times 31$  using a Gaussian filter with a kernel size of  $3 \times 3$  and a standard deviation of 0.5 to generate LR-HSI. Additionally, HR-MSI patches are generated using the spectral response function of the Nikon D700 camera. For the training images in the WDC dataset, 921 overlapping image patches of size  $64 \times 64 \times 191$  are obtained and downsampled to generate LR-HSI using the same method. The spectral response function from blue to SWIR2 of the Landsat 8 is selected to generate HR-MSI.

3) *Implementation Details:* The proposed SEST model is implemented using Pytorch 1.13.1 and Python 3.7.16 on the Windows operating system with an NVIDIA GPU GeForce RTX4080. Regarding the hyperparameters in the network, the channel feature mapping C is set to 48 in the shallow feature extraction process. The number of transformer blocks in the SEST residual block is set to 6, and the window sizes in the multiwindow residual block are set to 8, 16, and 32, respectively. To train the proposed network, the Adam optimizer is employed with  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . The initial learning rate is set to  $1e-4$  and halved every 75 epochs. The network is trained for a total of 350 epochs.

### B. Performance Evaluation

To verify the performance of the proposed method, a comparative analysis is conducted against six state-of-the-art HSI-SR methods, including three traditional methods, namely CNMF [32], FUSE [30], and HySure [31], along with three DL-based methods, namely SSRNet [55], Fusformer [46], and PSRT [48]. To ensure a fair and consistent comparison, all DL models are trained with the same input, and the hyperparameter settings are aligned with the specifications outlined in their respective original papers. Moreover, to provide a more intuitive and quantitative comparison of the performance of the aforementioned methods, four widely used HSI-SR quality indices (QIs) are employed, including PSNR, spectral angle mapping (SAM), structural similarity (SSIM), and erreur relative globale adimensionnelle de synthèse (ERGAS). Among these QIs, superior reconstruction performance is indicated by higher PSNR and SSIM values, alongside lower SAM and ERGAS values.

1) *Experimental Results on the CAVE Dataset:* Table I presents a comprehensive quantitative evaluation of the reconstruction performance achieved by the proposed method along with a comparison to state-of-the-art methods on the CAVE dataset, with the best results highlighted in bold and the second-best results underlined. Notably, the proposed SEST method consistently outperforms comparable methods across all four QIs on the CAVE dataset, establishing its consistently superior efficacy.

To substantiate the quantitative findings through visual aspect, two representative images, specifically *balloon* and *feather*, are selected from the CAVE dataset for in-depth visualization. Fig. 4 shows the reconstructed images and their corresponding residual images produced by different methods, with highlighted regions (within the red boxes) for detailed comparison. The visual inspection reveals the superior performance of the SEST method, particularly evident in the residual images (generated by a randomly selected band), underscoring the ability of SEST

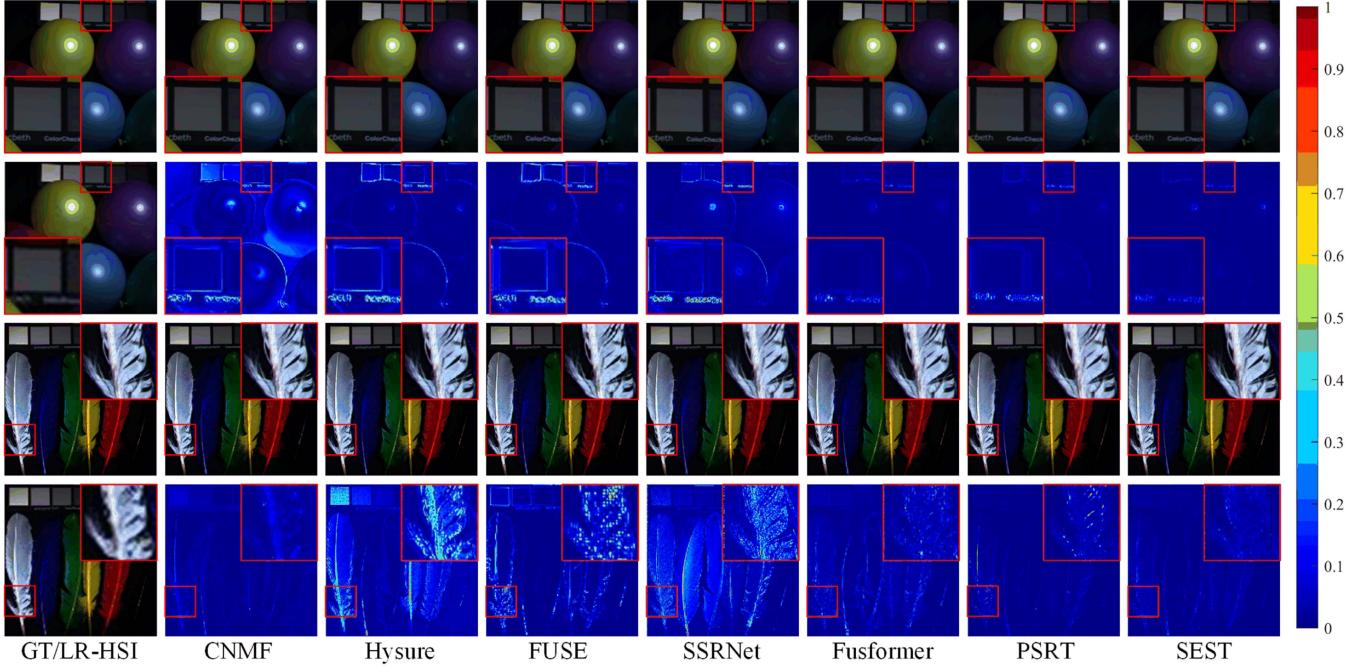


Fig. 4. First column: the GTs and the corresponding LR-HSI images (in pseudo-colors) for the *balloon* (first and second rows) and the *feather* (third and fourth rows) test cases from the CAVE dataset. The second to eighth columns: the visual results and the residuals (generated by a randomly selected band) between the GT and the fused products for all the compared approaches. A zoomed area has been added to aid the visual inspection.

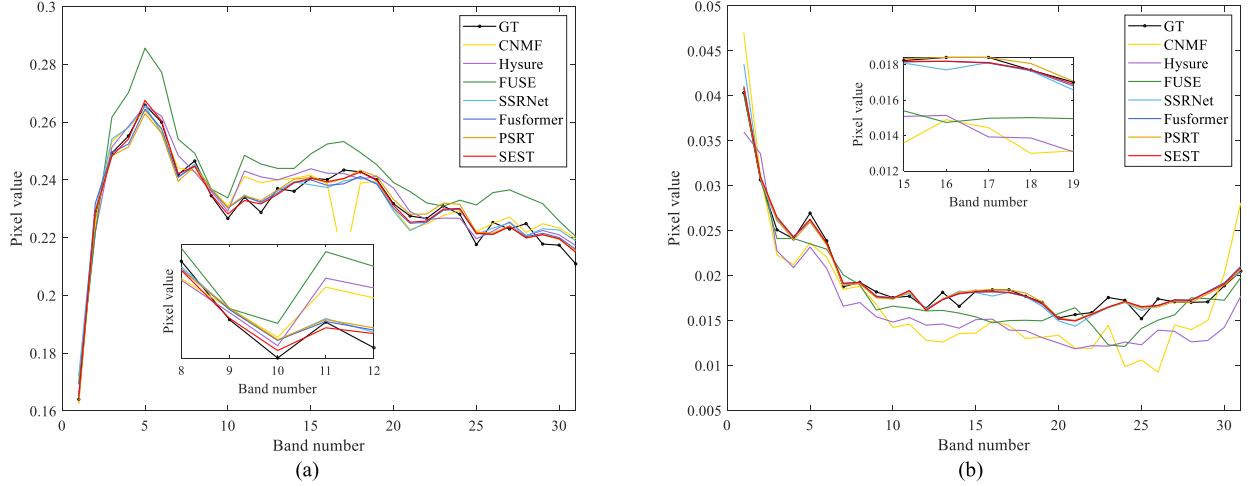


Fig. 5. Spectral vectors analysis of the GT and results of the compared approaches for the (a) *balloon* located at (60, 270) and (b) *feather* located at (50, 420).

to achieve reconstructions that closely align with the ground truth (GT), thereby demonstrating better recovery of spatial details.

To better compare the spectral fidelity of different methods, two pixels from *balloon* and *feather* are randomly selected to compare spectral differences against the GT, as shown in Fig. 5. The spectral vectors generated by the proposed SEST closely resemble the GT, indicating a significant reduction in spectral distortion and an enhanced preservation of spectral fidelity of the proposed SEST method.

2) *Experimental Results on Harvard Dataset*: To evaluate the generalization capability of the proposed SEST method, the

CAVE dataset is exclusively used to train the proposed network, which is then tested on the Harvard dataset. The quantitative results, as presented in Table I, showcase that the proposed SEST outperforms all compared methods in terms of PSNR, SAM, and SSIM, while securing not that good in terms of ERGAS due to the more complex noise types present in real-world scenarios. For visual assessment, similar experiments are conducted with the results shown in Figs. 6 and 7.

As shown in Fig. 6, two test images from the Harvard dataset are selected for analysis, displaying pseudo-color images and residual images of the reconstructed results. Similarly, Fig. 7 illustrates the spectral vector diagrams of the reconstructed

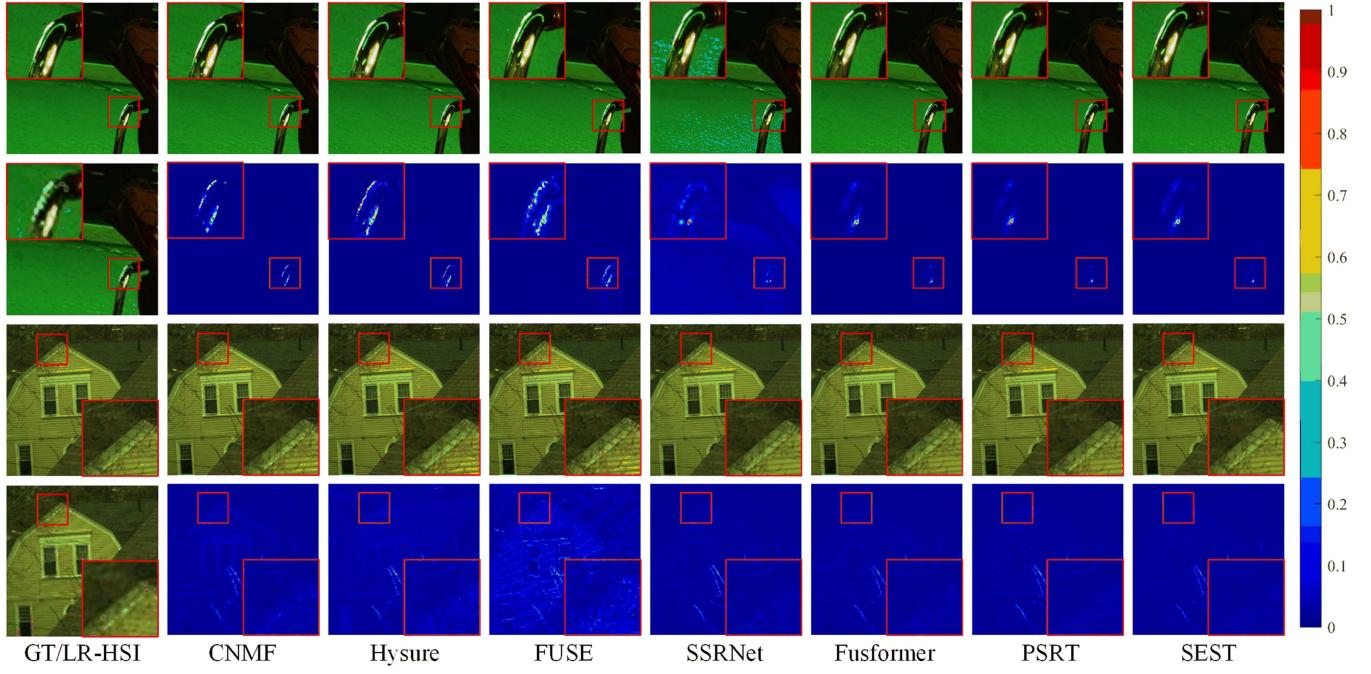


Fig. 6. First column: the GTs and the corresponding LR-HSI images (in pseudo-colors) for the *imga3* (first and second rows) and the *imgc9* (third and fourth rows) test cases from the Harvard dataset. The second to eighth columns: the visual results and the residuals (generated by a randomly selected band) between the GT and the fused products for all the compared approaches. A zoomed area has been added to aid the visual inspection.

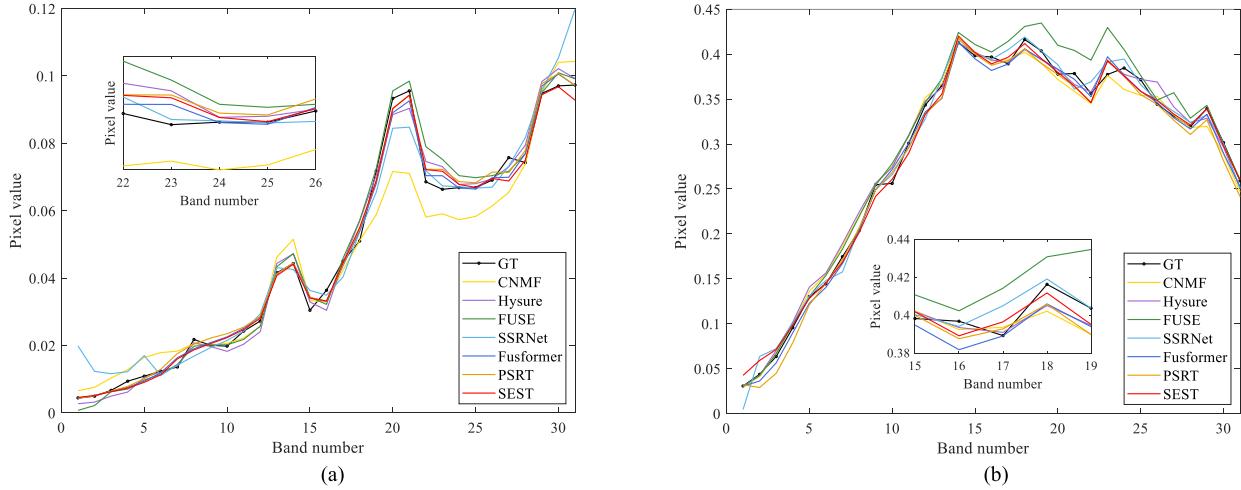


Fig. 7. Spectral vectors analysis of the GT and results of the compared approaches for the (a) *imga3* located at (200, 400) and (b) *imgc9* located at (200, 250).

results by different methods, facilitating a comparison of spectral fidelity. The SEST consistently delivers superior results in both spatial and spectral domains, aligning with the quantitative analysis.

3) *Experimental Results on the WDC Dataset:* To evaluate the robustness of the proposed SEST method on real-world data, the majority of the WDC dataset is used for training the network, while the remaining portion is used for testing its performance. Quantitative results in Table I demonstrate that the proposed SEST method outperforms all comparison methods in terms of PSNR and SAM, and ranks second in SSIM. Similar performance as the Harvard dataset of the ERGAS metric could be

drawn by the results. For visual assessment, similar experiments are conducted with the results shown in Figs. 8 and 9.

As shown in Fig. 8, two test images from the WDC dataset are selected for analysis, displaying pseudo-color images and residual images of the reconstructed results. Similarly, Fig. 9 illustrates the spectral vector diagrams of the reconstructed results by different methods, facilitating a comparison of spectral fidelity. The SEST consistently delivers superior results in both spatial and spectral domains, aligning with the quantitative analysis.

4) *Discussion and Analysis:* The experimental results on the CAVE, Harvard, and WDC datasets demonstrate that the

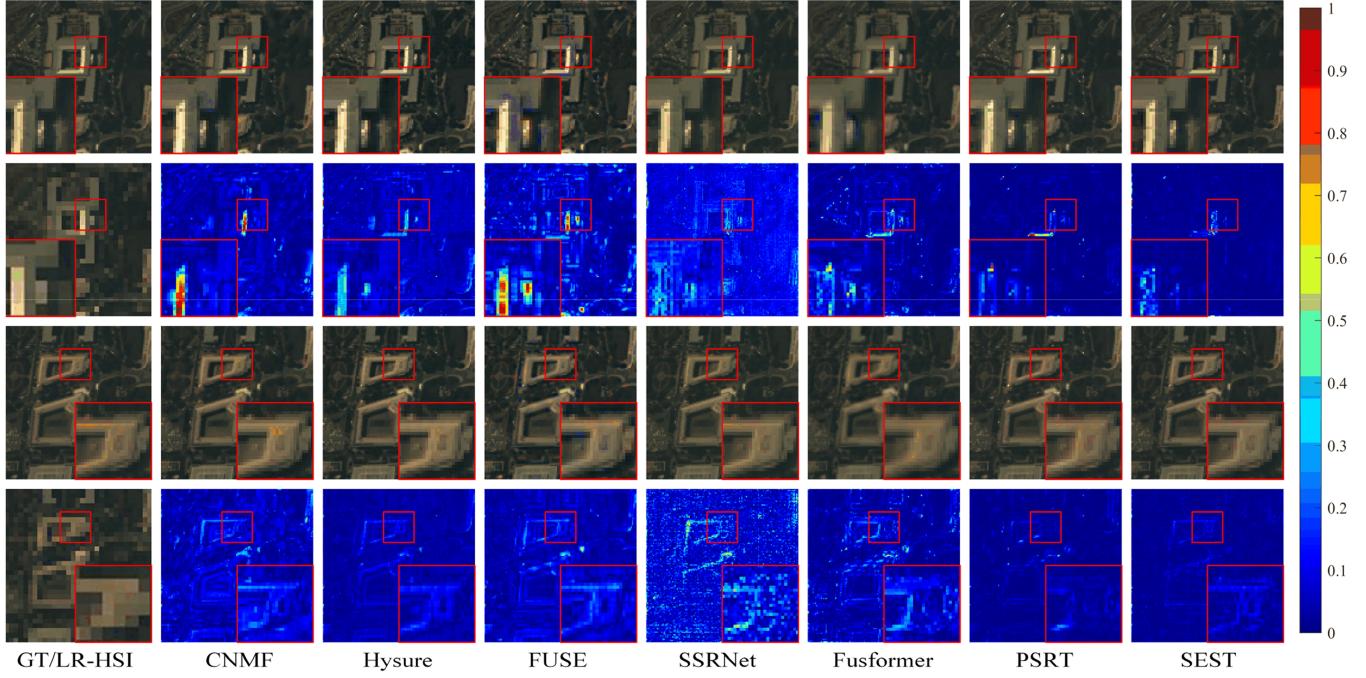


Fig. 8. First column: the GTs and the corresponding LR-HSI images (in pseudo-colors) for the *img1* (first and second rows) and the *img2* (third and fourth rows) test cases from the WDC dataset. The second to eighth columns: the visual results and the residuals (generated by a randomly selected band) between the GT and the fused products for all the compared approaches. A zoomed area has been added to aid the visual inspection.

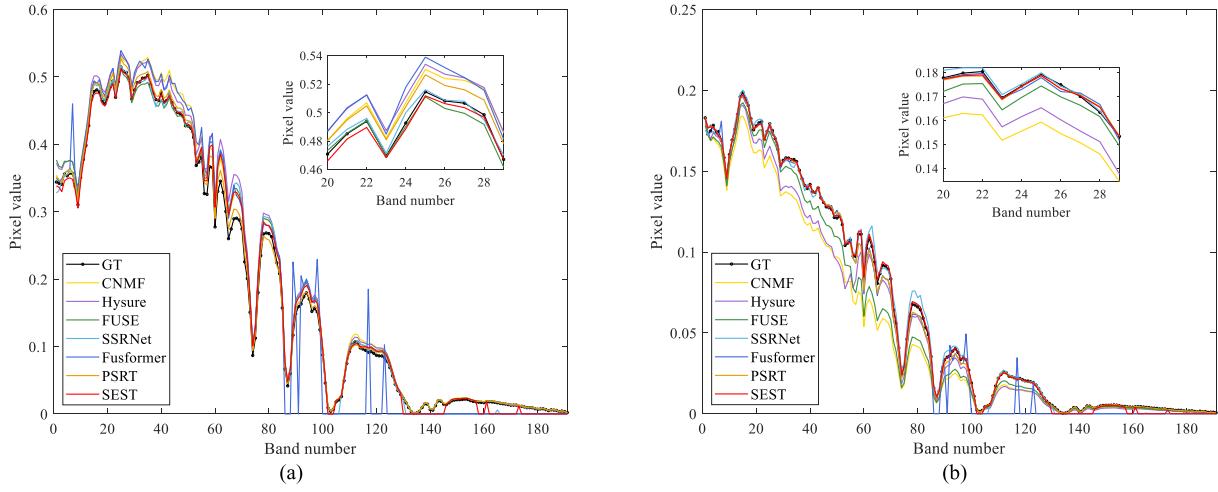


Fig. 9. Spectral vectors analysis of the GT and results of the compared approaches for the (a) *img1* located at (40, 70) and (b) *img2* located at (45, 80).

proposed method exhibits superior reconstruction performance across the quantitative metrics of PSNR, SAM, and SSIM. However, despite these positive outcomes indicating the method's overall effectiveness in reconstructing spectral information, it performs suboptimally in the ERGAS metric. Further analysis reveals that although the reconstructed spectral curves closely match the ground truth in general, the presence of extreme outliers in certain bands results in significant deviations from the ground truth, which substantially increases the ERGAS value.

A deeper investigation suggests that these outliers may be attributed to two inherent limitations of the proposed method: 1) To

reduce the computational cost of the self-attention mechanism and enhance the network's ability to handle high-dimensional data, the proposed method restricts sparse attention to fixed windows. While this design effectively reduces computational overhead, it also compromises the transformer model's ability to capture long-range dependencies, thereby affecting the precise reconstruction of spectral information and leading to large errors in specific bands or pixels. 2) Although the multiwindow residual block design reduces reliance on manual parameter tuning, its inherent structural limitations may lead to inconsistent performance when dealing with hyperspectral data of varying

TABLE II  
RECONSTRUCTION RESULTS ON DIFFERENT WAYS OF COMBINING KEY MODULES

Components	Different combinations of components			
LeFF	x	x	✓	✓
SE	x	✓	x	✓
PSNR	50.31	50.48	50.47	<b>51.11</b>
SAM	1.93	1.91	1.89	<b>1.81</b>
ERGAS	1.27	1.15	1.07	<b>1.04</b>
SSIM	0.9964	0.9966	0.9969	<b>0.9971</b>

The best values are highlighted in bold.

characteristics. In certain cases, these limitations can result in increased errors in specific bands, thus causing larger deviations.

### C. Ablation Study

This section evaluates the contributions of individual components within the proposed SEST method through a series of ablation experiments, mainly focusing on four critical aspects: the integration strategy of key modules, the sizes of the residual windows, the location of spectral enhancement, and the influence of loss functions. To be concise and not affect the generality, the ablation experiments are exclusively conducted on the CAVE dataset with a detailed analysis of different factors of the proposed SEST model on the reconstruction performance, allowing for a focused analysis of how each factor influences reconstruction performance.

1) *Integration Strategy of Key Modules*: As has been demonstrated before, the SEST framework incorporates three pivotal modules: the multiwindow residual block, spectral enhancement, and LeFF. While the multiwindow residual block is comprehensively analyzed in terms of window sizes separately (discussed in the subsequent subsection), this subsection assesses the significance of the remaining SE and LeFF modules. Results of the quantitative reconstruction metrics in Table II indicate that integrating either the LeFF or the SE module individually yields improvements in reconstruction quality compared to a baseline model devoid of these two modules, confirming the necessity of each module. Notably, when both modules are combined, as shown in the final column of Table II, there is a markable improvement in performance, surpassing the individual contributions of each module. This synergistic effect underscores the complementary nature of these two modules, enhancing the overall efficacy of the HSI-SR process.

2) *Size of Residual Windows*: Experimental results concerning various window sizes are detailed in Table III, illustrating the effect of different window sizes on reconstruction quality, where rows 2–9 delineate the impact of the single-window method with window sizes varying from 4 to 32. Interestingly, the reconstruction quality does not exhibit a straightforward improvement with an increase of the window size. This observation underscores the significance of selecting an appropriate window size, emphasizing the critical importance of the multiscale window design.

Moreover, Table III also provides a comparative analysis of the training time for each method, with the average time for

TABLE III  
RECONSTRUCTION RESULTS ON DIFFERENT SIZES OF RESIDUAL WINDOWS

Method	PSNR	SAM	ERGAS	SSIM	Time/s
Fusformer	48.56	2.52	1.30	0.9953	163
Win=4	50.78	1.87	1.09	0.9969	<b>15</b>
Win=8	51.04	1.86	1.05	0.9970	16
Win=12	50.84	1.82	1.08	0.9967	22
Win=16	50.96	1.84	1.12	0.9969	23
Win=20	50.76	1.87	1.07	0.9970	39
Win=24	51.06	1.87	1.13	0.9969	41
Win=28	50.99	1.83	1.05	0.9970	67
Win=32	50.95	1.85	1.07	0.9969	51
Win=8 16 32	<b>51.11</b>	<b>1.81</b>	<b>1.04</b>	<b>0.9971</b>	88

The best values are highlighted in bold.

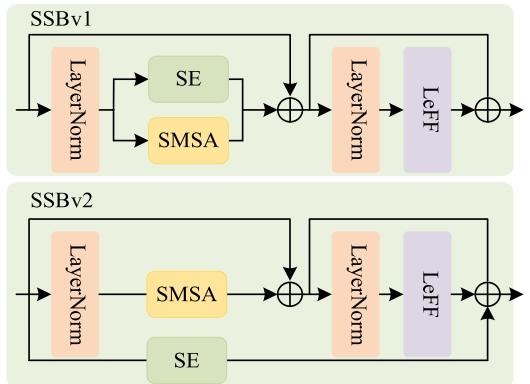


Fig. 10. Structure of the SSBv1 and SSBv2.

TABLE IV  
RECONSTRUCTION RESULTS ON DIFFERENT LOCATIONS OF SPECTRAL ENHANCEMENT

Method	PSNR	SAM	ERGAS	SSIM
SSBv1	<b>51.11</b>	<b>1.813</b>	<b>1.04</b>	<b>0.9970</b>
SSBv2	50.45	1.837	1.15	0.9965

The best values are highlighted in bold.

training one epoch. Notably, window-based methods demonstrate significantly reduced training time compared to Fusformer. While the multiscale window method requires more training time compared to the single window, it should be noted that each window operates independently within the network, allowing for parallel training on multiple GPUs, thereby facilitating a reduction in overall training time.

3) *Location of Spectral Enhancement*: In an effort to ascertain the optimal configuration of the spectral enhancement module within the SEST architecture, two variants of the SSB are tested, denoted as SSBv1 and SSBv2. SSBv1 places the SE module externally to the transformer block, whereas SSBv2 integrates the SE module directly within the self-attention calculation to fine-tune the self-attention map, as depicted in Fig. 10.

The comparative results, presented in Table IV, demonstrate that the SSBv2 configuration proves to be more conducive to image reconstruction in terms of reconstruction quality. This

TABLE V  
RECONSTRUCTION RESULTS ON INFLUENCE OF LOSS FUNCTIONS

Functions	Different combinations of functions		
$L_{MAE}$	×	✓	✓
$L_{SSTV}$	✓	✗	✓
PSNR	49.77	49.82	<b>51.11</b>
SAM	2.31	2.24	<b>1.81</b>
ERGAS	1.29	1.26	<b>1.04</b>
SSIM	0.9960	0.9960	<b>0.9971</b>

The best values are highlighted in bold.

insightful analysis sheds light on the impact of the location of the spectral enhancement module within the sparse transformer block, emphasizing its significance in improving HSI super-resolution reconstruction efficacy.

4) *Influence of Loss Functions*: To balance the recovery of high-frequency spatial details and spectral fidelity, this article introduces the use of MAE and SSTV as loss functions, namely  $L_{MAE}$  and  $L_{SSTV}$ .  $L_{MAE}$  aims to reduce pixel-level errors between the reconstructed and true images, thereby enhancing overall image quality and detail.  $L_{SSTV}$ , on the other hand, constrains spatial and spectral variations to reduce noise and maintain spectral consistency and fidelity.

To determine the impact of these loss functions on super-resolution reconstruction results, this article tests three different combinations: “w  $L_{MAE}$  and w/o  $L_{SSTV}$ ,” “w/o  $L_{MAE}$  and w  $L_{SSTV}$ ,” and “w  $L_{MAE}$  and w  $L_{SSTV}$ .” Table V presents the experimental results for these combinations. The comparison shows that using either  $L_{MAE}$  or  $L_{SSTV}$  alone results in comparable reconstruction quality. However, using both together significantly enhances the quality of the reconstructed images. This finding indicates that a reasonable combination of different loss functions can fully exploit their respective advantages, leading to higher-quality image reconstruction.

## V. CONCLUSION

This article presents a novel SEST network tailored for HSI-SR tasks. This innovative model, engineered to operate within the spatial domain, incorporates sparse self-attention and a local enhancement feedforward network to capture global features while preserving local details. Addressing the spectral distortions in HSI-SR tasks, an SE module is specially designed within the self-attention calculation process, forming a powerful hybrid attention mechanism. Furthermore, to exploit multiscale information inherent in the image, a well-crafted multiwindow residual block is devised, contributing significantly to the overall improvement in reconstruction quality. The comprehensive experiments conducted on the CAVE, Harvard, and WDC datasets convincingly validate the superior performance of the proposed approach, underscoring its significance in the field of HSI-SR. In light of the limitations identified in the experimental analysis, future work could explore the integration of dynamic convolutions to enable more flexible window size adjustments and diversified attention patterns based on the characteristics of the input data, thereby mitigating the occurrence of outliers.

## REFERENCES

- [1] Y. Wang, Q. Zhu, H. Ma, and H. Yu, “A hybrid gray wolf optimizer for hyperspectral image band selection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5527713.
- [2] Y. Wang, X. Chen, E. Zhao, C. Zhao, M. Song, and C. Yu, “An unsupervised momentum contrastive learning based transformer network for hyperspectral target detection,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 9053–9068, Apr. 2024.
- [3] C. Yu, B. Gong, M. Song, E. Zhao, and C. I. Chang, “Multiview calibrated prototype learning for few-shot hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Dec. 2022, Art. no. 5544713.
- [4] C. Cheng, J. Peng, and W. Cui, “A two-stage convolutional sparse coding network for hyperspectral image classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Feb. 2023, Art. no. 5501905.
- [5] Y. Wang, X. Chen, F. Wang, M. Song, and C. Yu, “Meta-learning based hyperspectral target detection using Siamese network,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5527913.
- [6] Y. Wang, X. Chen, E. Zhao, and M. Song, “Self-supervised spectral-level contrastive learning for hyperspectral target detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Apr. 2023, Art. no. 5510515.
- [7] F. Luo, T. Zhou, J. Liu, T. Guo, X. Gong, and J. Ren, “Multiscale diff-changed feature fusion network for hyperspectral image change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jan. 2023, Art. no. 5502713.
- [8] R. Dian, S. Li, B. Sun, and A. Guo, “Recent advances and new guidelines on hyperspectral and multispectral image fusion,” *Inf. Fusion*, vol. 69, pp. 40–51, May 2021.
- [9] J. Jiang, H. Sun, X. Liu, and J. Ma, “Learning spatial-spectral prior for super-resolution of hyperspectral imagery,” *IEEE Trans. Comput. Imag.*, vol. 6, pp. 1082–1096, May 2020.
- [10] T. Xu, T. Z. Huang, Y. Chen, J. Huang, and L. J. Deng, “A variational approach with nonlocal self-similarity and joint-sparsity for hyperspectral image super-resolution,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2021, pp. 2444–2447.
- [11] X. Fu, S. Jia, M. Xu, J. Zhou, and Q. Li, “Sparsity constrained fusion of hyperspectral and multispectral images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jan. 2022, Art. no. 6006705.
- [12] W. Wan, W. Guo, H. Huang, and J. Liu, “Nonnegative and nonlocal sparse tensor factorization-based hyperspectral image super-resolution,” *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8384–8394, Dec. 2020.
- [13] J. Xue, Y. Q. Zhao, Y. Bu, W. Liao, J. C. W. Chan, and W. Philips, “Spatial-spectral structured sparse low-rank representation for hyperspectral image super-resolution,” *IEEE Trans. Image Process.*, vol. 30, pp. 3084–3097, Feb. 2021.
- [14] C. Liu, Z. Fan, and G. Zhang, “GJTD-LR: A trainable grouped joint tensor dictionary with low-rank prior for single hyperspectral image super-resolution,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5537617.
- [15] J. Zhang, J. Liu, J. Yang, and Z. Wu, “Crossed dual-branch U-Net for hyperspectral image super-resolution,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 2296–2307, Dec. 2024.
- [16] Q. Xie, M. Zhou, Q. Zhao, Z. Xu, and D. Meng, “MHF-Net: An interpretable deep network for multispectral and hyperspectral image fusion,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1457–1473, Mar. 2022.
- [17] A. Khader, J. Yang, and L. Xiao, “Model-guided deep unfolded fusion network with nonlocal spatial-spectral priors for hyperspectral image super-resolution,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4607–4625, May 2023.
- [18] Y. Wang, H. Wang, E. Zhao, M. Song, and C. Zhao, “Tucker decomposition-based network compression for anomaly detection with large-scale hyperspectral images,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 10674–10689, May 2024.
- [19] 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.
- [20] Q. Ma, J. Jiang, X. Liu, and J. Ma, “Learning a 3D-CNN and transformer prior for hyperspectral image-resolution,” *Inf. Fusion*, vol. 100, Dec. 2023, Art. no. 101907.
- [21] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
- [22] Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Proc. Int. Conf. Learn. Representations*, Jan. 2021.

- [23] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 9992–10002.
- [24] K. Choromanski et al., "Rethinking attention with performers," in *Proc. Int. Conf. Learn. Representations*, Jan. 2021.
- [25] R. Child, S. Gray., A. Radford., and L. Sutskever., "Generating long sequences with sparse transformers," 2019, *arXiv:1904.10509*.
- [26] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
- [27] B. Aiazzi, S. Baronti, and M. Selva, "Improving component substitution pansharpening through multivariate regression of MS + pan data," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3230–3239, Oct. 2007.
- [28] M. Selva, B. Aiazzi, F. Butera, L. Chiarantini, and S. Baronti, "Hyper-sharpening: A first approach on SIM-GA data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 3008–3024, Jun. 2015.
- [29] N. Akhtar, F. Shafait, and A. Mian, "Bayesian sparse representation for hyperspectral image super resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3631–3640.
- [30] Q. Wei, N. Dobigeon, and J. Y. Tourneret, "Fast fusion of multi-band images based on solving a Sylvester equation," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4109–4121, Nov. 2015.
- [31] M. Simões, J. Bioucas-Dias, L. B. Almeida, and J. Chanussot, "A convex formulation for hyperspectral image superresolution via subspace-based regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3373–3388, Jun. 2015.
- [32] N. Yokoya, C. Grohnfeldt, and J. Chanussot, "Hyperspectral and multispectral data fusion: A comparative review of the recent literature," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 2, pp. 29–56, Jun. 2017.
- [33] R. Dian, L. Fang, and S. Li, "Hyperspectral image super-resolution via non-local sparse tensor factorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3862–3871.
- [34] J. Li, K. Zheng, Z. Li, L. Gao, and X. Jia, "X-shaped interactive autoencoders with cross-modality mutual learning for unsupervised hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jul. 2023, Art. no. 5518317.
- [35] Q. Li, M. Gong, Y. Yuan, and Q. Wang, "RGB-induced feature modulation network for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, May 2023, Art. no. 5512611.
- [36] S. Mei, X. Yuan, J. Ji, Y. Zhang, S. Wan, and Q. Du, "Hyperspectral image spatial super-resolution via 3D full convolutional neural network," *Remote Sens.*, vol. 9, no. 11, Nov. 2017, Art. no. 1139.
- [37] Q. Li, Y. Yuan, X. Jia, and Q. Wang, "Dual-stage approach toward hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 31, pp. 7252–7263, Nov. 2022.
- [38] Q. Li, Q. Wang, and X. Li, "Mixed 2D/3D convolutional network for hyperspectral image super-resolution," *Remote Sens.*, vol. 12, no. 10, May 2020, Art. no. 1660.
- [39] Q. Li, M. Gong, Y. Yuan, and Q. Wang, "Symmetrical feature propagation network for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5536912.
- [40] Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process Syst.*, Dec. 2017, pp. 5998–6008.
- [41] Y. Wang et al., "Constrained-target band selection for multiple-target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 6079–6103, Aug. 2019.
- [42] W. Lu et al., "A CNN-transformer hybrid model based on CSWin transformer for UAV image object detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1211–1231, Jan. 2023.
- [43] X. Chen, M. Zhang, and Y. Liu, "Target detection with spectral graph contrast clustering assignment and spectral graph transformer in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Apr. 2024, Art. no. 5516916.
- [44] X. Zhou, L. Zhou, S. Gong, S. Zhong, W. Yan, and Y. Huang, "Swin Transformer embedding dual-stream for semantic segmentation of remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 175–189, Oct. 2024.
- [45] J. Liang, J. Cao, G. Sun, K. Zhang, L. V. Gool, and R. Timofte, "SwinIR: Image restoration using Swin Transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop.*, Oct. 2021, pp. 1833–1844.
- [46] J. F. Hu, T. Z. Huang, L. J. Deng, H. X. Dou, D. Hong, and G. Vivone, "Fusformer: A transformer-based fusion network for hyperspectral image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jul. 2022, Art. no. 6012305.
- [47] Y. Long, X. Wang, M. Xu, S. Zhang, S. Jiang, and S. Jia, "Dual self-attention Swin Transformer for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, May 2023, Art. no. 5512012.
- [48] S.-Q. Deng, L.-J. Deng, X. Wu, R. Ran, D. Hong, and G. Vivone, "PSRT: Pyramid Shuffle-and-Rewhole Transformer for multispectral and hyperspectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 5503715.
- [49] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.
- [50] Y. Chang, L. Yan, and S. Zhong, "Hyper-Laplacian regularized unidirectional low-rank tensor recovery for multispectral image denoising," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5901–5909.
- [51] H. Wu et al., "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 22–31.
- [52] H. K. Aggarwal and A. Majumdar, "Hyperspectral image denoising using spatio-spectral total variation," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 3, pp. 442–446, Mar. 2016.
- [53] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar, "Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2241–2253, Sep. 2010.
- [54] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 193–200.
- [55] X. Zhang, W. Huang, Q. Wang, and X. Li, "SSR-NET: Spatial-spectral reconstruction network for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5953–5965, Jul. 2021.



**Yuchao Yang** (Student Member, IEEE) was born in Xianning, China, in 1998. He received the B.E. degree in electronic information engineering in 2020 from Dalian Maritime University, Dalian, China, where he is currently working toward the Ph.D. degree in information and communication engineering with the Information Science and Technology College.

His research interests include hyperspectral image super-resolution reconstruction and deep learning.



**Yulei Wang** (Member, IEEE) was born in Yantai, China, in 1986. She received the B.S. and Ph.D. degrees in signal and information processing from Harbin Engineering University, Harbin, China, in 2009 and 2015, respectively.

From 2011 to 2013, she was a joint Ph.D. student with the Remote Sensing Signal and Image Processing Laboratory, University of Maryland, Baltimore County. From 2011 to 2013, she was a Research Assistant with the Shock, Trauma and Anesthesiology Research organized research center (STAR-ORC), University of Maryland, School of Medicine. She is currently an Associate Professor and doctoral supervisor with Hyperspectral Imaging in Remote Sensing (CHIRS), Information Science and Technology College, Dalian Maritime University, Dalian, China. Her research interests include hyperspectral image processing, multisource remote sensing fusion, and vital signs signal processing. For more information, please visit the website <https://github.com/YuleiWang1/>.



**Hongzhou Wang** was born in Shenyang, China, in 2000. He received the B.S. degree in communication engineering in 2022 from the Information Science and Technology College, Dalian Maritime University, Dalian, China, where he is currently working toward the M.S. degree in information and communication engineering.

His research interests include hyperspectral target detection and deep learning.



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

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

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



**Enyu Zhao** (Member, IEEE) was born in Dalian, China, in 1987. He received the Ph.D. degree in cartography and geographic information system from the College of Resources and Environment, University of Chinese Academy of Sciences, Beijing, China, in 2017.

From 2014 to 2016, he was a joint Ph.D. student with Engineering Science, Computer Science and Imaging Laboratory, University of Strasbourg, Strasbourg, France. He is currently an Associate Professor with the College of Information Science and Technology, Dalian Maritime University, Dalian, China. His research interests include quantitative remote sensing and hyperspectral image processing.



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

From 2013 to 2014, she was a visiting associate research scholar with the Remote Sensing Signal and Image Processing Laboratory, University of Maryland, Baltimore County. She is currently a Professor and doctoral supervisor with the Hyperspectral Imaging in Remote Sensing (CHIRS), Information Science and Technology College, Dalian Maritime University, Dalian, China. Her research interests include remote sensing and hyperspectral image processing.



**Chunyan Yu** (Senior Member, IEEE) received the B.S. and Ph.D. degrees in environmental engineering from Dalian Maritime University, Dalian, China, in 2004 and 2012, respectively.

In 2004, she joined the College of Computer Science and Technology, Dalian Maritime University. From 2013 to 2016, she was a Postdoctoral Fellow with the Information Science and Technology College, Dalian Maritime University. From 2014 to 2015, she was a Visiting Scholar with the College of Physicians and Surgeons, Columbia University, New York City, NY, USA. She is currently an Associate Professor with the Information Science and Technology College, Dalian Maritime University. Her research interests include image segmentation, hyperspectral image classification, and pattern recognition.