Butterfly Residual Network: A Hybrid Approach With Spectral Transformers and Depth-Wise Convolutions for Hyperspectral Image Super-Resolution

Yuchao Yang, *Student Member, IEEE*, Yulei Wang, *Member, IEEE*, Xin Xu, and Enyu Zhao, *Member, IEEE*

***Abstract*—Hyperspectral image (HSI) super-resolution reconstruction is a challenging ill-posed inverse problem, which seeks to enhance the spatial resolution of low-resolution hyperspectral images (LR-HSIs) by integrating complementary information from high-resolution multispectral images (HR-MSIs), ultimately generating high-resolution HSIs (HR-HSIs). Existing methods commonly employ residual connections and deep layer stacking to facilitate information propagation. While residual connections effectively preserve gradient flow, we observe that naively increasing network depth in high-dimensional spectral tasks can lead to feature redundancy and performance saturation. To address these challenges, this article presents a novel Butterfly residual network (BRNet) that incorporates spectral Transformers and depth-wise convolutions to optimize both accuracy and computational efficiency of hyperspectral super-resolution reconstruction from two perspectives: learning strategy and feature extraction. Regarding learning strategy, a recursive structure coupled with a fusion parameter generation technique is proposed to promote efficient feature fusion and enable adaptive network pruning, thereby reducing redundant information and enhancing computational efficiency. For feature extraction, spectral Transformer and depth-wise convolution are employed to capture spectral and spatial features, respectively, effectively leveraging their complementary advantages across different dimensions. A specialized spectral-spatial interaction (SSI) module is then incorporated to effectively fuse the extracted features, thereby enriching the diversity of network features. Additionally, the convolutional gated feed-forward network (FFN) is designed to bolster the network's ability to capture local features while significantly reducing the computational complexity. Experimental evaluations on three hyperspectral datasets demonstrate that the proposed method outperforms existing state-of-the-art super-resolution reconstruction methods across various performance metrics, validating its effectiveness and superiority.**

***Index Terms*—Butterfly residual network (BRnet), depth-wise convolution, gating mechanism, hyperspectral image (HSI), spectral transformer, super-resolution.**

Received 30 April 2025; revised 7 September 2025; accepted 5 November 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 42271355 and Grant 61801075, in part by the Natural Science Foundation of Liaoning Province under Grant 2022-MS-160, and in part by the Fundamental Research Funds for the Central Universities under Grant 3132025251. *(Corresponding author: Yulei Wang.)*

The authors are with the Center of Hyperspectral Imaging in Remote Sensing (CHIRS), Information Science and Technology College, Dalian Maritime University, Dalian, Liaoning 116026, China (e-mail: yyc@dlmu.edu.cn; wangyulei@dlmu.edu.cn; dmu1120250287@dlmu.edu.cn; zhaoenyu@dlmu.edu.cn).

Digital Object Identifier 10.1109/TNNLS.2025.3631243

I. Introduction

HYPERSPECTRAL imaging (HSI) is a critical technology in remote sensing applications, capturing the reflectance or radiance spectrum of target objects across contiguous narrow spectral bands to form a 3-D spectral data cube [1], [2]. These images not only contain the 2-D spatial information of objects but also provide continuous spectral information capable of detecting subtle material differences. As a result, hyperspectral remote sensing images have been widely used in fields such as image classification [3], [4], anomaly detection [5], [6], and target detection [7], [8]. However, despite the significant advantage of HSIs in spectral resolution, current hyperspectral imaging devices suffer from limited spatial resolution due to hardware constraints, including detector size, transmission systems, as well as the trade-offs aimed at improving the signal-to-noise ratio, which restricts the broader application of HSI in practice [9]. To address this issue, hyperspectral image super-resolution reconstruction (HSI-SR) has emerged as an effective postprocessing technique to enhance the spatial resolution at low cost. Among existing HSI-SR strategies, fusion-based methods, which integrate high-resolution multispectral images (HR-MSIs) from the same scene, are the most widely adopted [10].

In general, mainstream fusion-based HSI-SR techniques can be broadly categorized into two types: model-based and learning-based approaches [11]. Model-based approaches first formulate a degradation model, then they commonly require manual introduction of various prior information to formulate regularization terms, which are incorporated into optimization frameworks to constrain the solution space. Finally, the reconstruction image is obtained through inversion using optimization algorithms. While these methods benefit from strong physical interpretability [12], their manually defined priors often fall short in capturing the intricate and nonlinear nature of real-world degradation process. Moreover, the iterative optimization process incurs significant computational overhead, posing challenges for real-time processing scenarios, particularly for large-scale hyperspectral datasets [13].

In recent years, the rapid advancement of deep learning (DL) has garnered considerable attention, owing to its powerful capabilities in automatic feature extraction and excellent nonlinear representation [14], [15], [16], [17], [18]. Compared to traditional model-based approaches, learning-based methods can adaptively learn the complex spectral–spatial correlations from data, enabling superior reconstruction performance. Based on the fusion strategy, learning-based methods are generally divided into late integration and early integration approaches [19]. Late integration methods encode spatial and spectral information independently through separate branches, fusing them at a later stage. This modular design allows each component to specialize in specific feature extraction tasks, thereby enhancing the precision and efficiency of feature learning. For example, Zheng et al. [20] employ a dual-stream convolutional autoencoder based on a linear mixing model, which estimates endmembers and abundances of HSI and MSI data through self-reconstruction. Hu et al. [21] incorporate channel attention and spatial attention modules to improve the spectral fidelity and spatial details of the reconstructed images. Despite the improved specificity of feature learning in late integration methods, such methods may inadequately capture cross-scale spatial–spectral interactions during the feature fusion stage. To address this issue, Sun et al. [22] propose a DSPNet network, extracting multiscale spectral and spatial information via SpePy and SpaPy, and designs a self-attention-based MLSIF module to establish long-range spectral–spatial interactions (SSIs) across different scales. Despite its precision, this method comes at the cost of increased model complexity and computational overhead, posing considerable challenges for real-world applications.

In contrast, early integration methods merge the data at the input stage, allowing a single network to simultaneously learn both spectral and spatial representations, thus avoiding the information loss caused by the independent learning modules in late integration methods. These approaches preserve the details and features of the data to the greatest extent and enhance the network's ability to represent high-dimensional features. Mei et al. [23] introduce 3-D full convolutions to achieve consistent learning of spatial–spectral features, but as the expense of increased redundancy, leading to wasted computational resources. Jiang et al. [24] address this issue by proposing a band grouping strategy that leverages the correlation between HSI bands, progressively up-sampling to generate high-resolution images, thereby effectively reducing computational overhead. Wang et al. [25] propose a self-calibrated attention residual network with efficient self-calibrated convolutions and attention mechanisms to precisely capture long-range spatial features and spectral dependencies. Recognizing the limitations of convolutional neural networks (CNNs) in global feature extraction, Hu et al. [26] introduce an encoder–decoder structure based on Vision Transformers (ViT) [27], using their larger receptive field to model global relationships across the entire feature space. Despite the notable advances achieved by these above methods in HSI-SR tasks, the inherent complexity and variability of hyperspectral data often necessitate deeper network architectures to model complex mapping relationships, which exacerbates the issue of feature degradation. This is particularly pronounced in Transformer-based models, where successive self-attention layers tend to produce homogenized feature representations, diminishing the model's ability to capture fine-grained local details [28].

To address the aforementioned issues, this article proposes a butterfly residual network (BRNet) built upon spectral Transformers and depth-wise convolutions. Specifically, to tackle the problem of degraded feature representation associated with deep network stacking, a butterfly recursive learning structure is designed, incorporating a fusion parameter generation strategy (FPGS) to dynamically generate pruning parameters from the intrinsic characteristics of the input data, thereby enhancing the diversity of the network structure while enabling efficient and adaptive feature fusion. To fully exploit the spectral–spatial information of hyperspectral data, a dual-branch feature extraction module is designed to align with the dual attributes of HSI: spectral Transformers are introduced to capture spectral self-similarity along the sequence for spectral feature extraction, while depthwise convolutions are employed to learn fine-grained local spatial structures for spatial feature extraction. In addition, a SSI module is designed to promote effective fusion of spectral and spatial features at the token level. Finally, a convolutional gated feed-forward network (CGFF) is incorporated to further enhance local details learning ability and facilitate more effective channel-wise feature interaction.

The main contributions of this article are as follows.

1) A BRNet with a recursive architecture is proposed, enhancing the structural diversity and flexibility of the network structure and enabling efficient fusion of multi-depths features, thereby preserving feature integrity and enhancing reconstruction accuracy.
2) An improved token mixer is developed by extending the self-attention mechanism to jointly model spectral and spatial representations, dynamically balancing the contributions and enabling efficient SSI, thereby improving the spectral fidelity and spatial consistency of the reconstruction results.
3) A CGFF is constructed to regulate feature flow via a dynamic gating mechanism, enhancing interchannel feature interaction, improving detail representation ability of the network, and reducing computational overhead.

The remainder of this article is organized as follows. Section II provides a review of related work in HSI-SR. Section III details the architecture of the proposed method. Section IV presents experimental results and performance evaluations. Finally, 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 two types: model-based methods and learning-based methods. Additionally, the limitations of two prevalent network architectures commonly employed in existing learning-based methods are analyzed in detail.

A. Model-Based Methods

Early research on HSI-SR often addresses the reconstruction task by extending traditional pansharpening strategies, mainly divided into two categories: Component substitution (CS) and multiresolution analysis (MRAs) approaches. For example, Aiazzi et al. [29] propose the adaptive Gram–Schmidt algorithm, which improves traditional CS-based image fusion methods by introducing a multivariate regression model. Vivone et al. [30] employ an MRA-based approach to extract high-frequency spatial details and estimate regression injection coefficients at full resolution using an iterative algorithm. Although these approaches effectively enhance spatial details, they tend to ignore the high interband similarity in HSI data, leading to spectral distortion. To address this issue, Bayesian methods have been introduced, leveraging predefined prior knowledge to constrain the fusion process. For instance, Wei et al. [31] propose a fast multiband image fusion algorithm based on solving the Sylvester equation (FUSE), which avoids iterative steps and achieves an efficient closed-form solution by utilizing the properties of convolution and downsampling matrices. Simões et al. [32] combine subspace dimensionality reduction with vector total variation regularization to improve the quality of reconstructed images and efficiently solved the problem using the ADMM framework. While Bayesian methods can naturally model uncertainty in the data, they are typically associated with high computational complexity. To further improve computational efficiency, decomposition-based algorithms have gained popularity. These methods reconstruct images by decomposing high-dimensional data into low-dimensional components. For example, Yokoya et al. [33] propose a coupled non-negative matrix factorization (CNMF) method that effectively addresses the spectral unmixing and fusion problem. Lanaras et al. [34] apply prior information to the linear mixing model and estimated spectral bases and coefficients using the proximal alternating linearized minimization (PALM) algorithm, further optimizing the image reconstruction process. Dian et al. [35] propose a nonlocal sparse tensor decomposition method that achieves more accurate super-resolution reconstruction by estimating sparse core tensors and dictionaries. Kanatsoulis et al. [36] further introduce a coupled tensor decomposition approach capable of recovering high-resolution images even in the presence of unknown or inaccurate spatial degradation operators.

B. Learning-Based Methods

Traditional model-based approaches rely heavily on manually defined priors, making them sensitive to parameter tuning and less adaptable to diverse data conditions. With the rapid development of DL technology, particularly CNNs, learning-based approaches have shown strong performance in HSI-SR due to their powerful feature extraction capabilities and simple end-to-end training processes. For instance, Xu et al. [37] propose a parallel dual-branch structure that combines multiscale feature extraction within the network with an adaptive angle and Laplacian (RAP) loss function, enabling joint optimization of low-resolution hyperspectral image (LR-HSI) and HR-MSI. Wu et al. [38] introduce a cross-modality nonlocal (CMNL) module to inject spatial information from HR-MSI into LR-HSI, employing a cross-scale self-calibrating convolution structure to enhance the fusion of multiscale spatial-spectral features. While effective in feature representation, these models often suffer from inefficient fusion due to the independent processing of LR-HSI and HR-MSI, resulting in suboptimal consistency between spatial and spectral components. To address this challenge, Zhu et al. [39] introduce a zero-centric residual learning module, which progressively enhances spatial detail representation in images through staged high-frequency detail learning in the spectral dimension during the fusion process. This incremental learning mechanism effectively alleviates the inconsistency between features. Furthermore, Li et al. [40] propose a three-stage fusion framework based on a degradation model, which enhances the model's interpretability and robustness of the fusion process by progressively injecting different levels of prior information during degradation information learning, initialized image establishment, and deep image generation.

In contrast to stepwise fusion strategies, an alternative approach treats LR-HSI and HR-MSI as a unified whole input, thereby enabling more comprehensive exploitation of the representational capacity of deep CNNs. Zhang et al. [41] propose a physically intuitive spatial–spectral reconstruction network (SSRNet) based on residual networks, which progressively restores lost information in HSIs through residual connections, achieving dual correction of spatial and spectral dimensions. However, the receptive field of CNNs is limited by the convolution kernel, making it difficult to capture long-range spatial dependencies over larger areas. To address this issue, Ran et al. [42] introduce nonlocal dilated convolutions and adaptive pointwise convolution modules to break through the limitations of traditional convolutions and capture broader spatial contextual information. Although these improvements have shown some effectiveness, the inherent limitations of convolutional structures still persist in practical applications, especially in complex spatial–spectral interactions, where expanding the receptive field remains a challenge. Consequently, compared to CNNs, Transformer architectures have rapidly gained widespread application in recent years due to their outstanding performance in modeling long-range dependencies. [43] Hu et al. [26] pioneer the application of Transformer to HSI-SR task through Fusformer, enabling effective modeling of long-range dependencies. PSRT [44] further improves upon this by employing a shuffle-and-reshuffle (SaR) strategy to facilitate feature interactions between different windows, demonstrating more efficient fusion effects in global information extraction. To further address the inconsistency between LR-HSI and HR-MSI, DCTransformer [45] propose a dual-cross attention mechanism, injecting advantageous features from both images into each other, combined with the self-attention enhancement mechanism of the Swin Transformer [46], achieving refined modeling of complex spectral–spatial structures.

While these methods have demonstrated notable progress in HSI-SR, they remain constrained in two key aspects. First, from the perspective of data representation, they fail to capture the joint spectral–spatial characteristics inherent in hyperspectral data. Second, at the architectural level, they do not effectively integrate the complementary advantages of CNN-based and Transformer-based models. Moreover, most existing network architectures still rely on simple skip

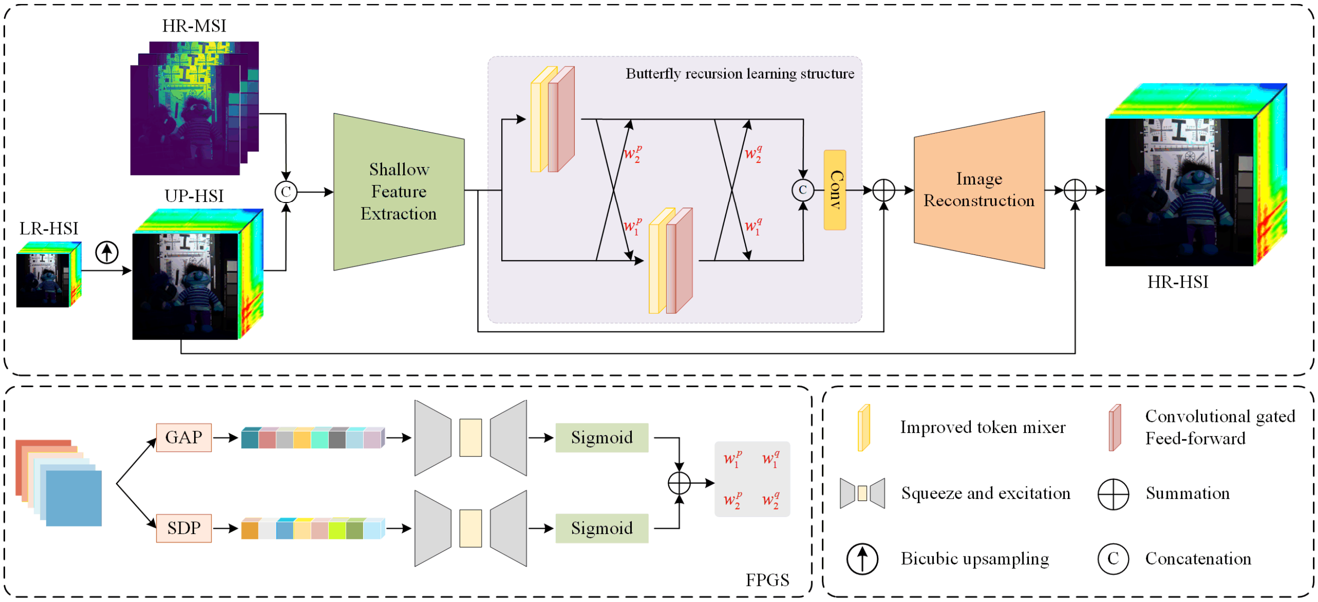

Fig. 1. Overall structure of the proposed method and the FPGS.

connections for interlayer feature propagation, which—while beneficial for information flow—may result in feature redundancy or insufficient feature refinement, especially in the presence of complex spatial–spectral couplings, thereby limiting the overall capacity of the network to fully exploit HSI information.

III. Proposed Method

In this section, the overall network architecture of the proposed method and its HSI-SR process are first described. The proposed framework is then elaborated from two key perspectives: the butterfly recursive learning structure, which addresses learning strategy and network efficiency; and the token-level SSI network, which focuses on feature extraction and spectral–spatial information fusion.

A. Network Architecture

Fusion-based HSI-SR aims to combine the rich spectral information of HSI with the high spatial resolution of MSI to generate high-resolution HSI (HR-HSI). Due to the limitations of imaging devices, HSI often has low spatial resolution, while MSI is limited in the spectral dimension. By fusing these two types of images, the spatial resolution of HSI can be significantly improved while retaining its spectral information. The relationship among these three types of images can be described as

$$\mathbf{Y} = \mathbf{R}\mathbf{X} \tag{1}$$

$$\mathbf{Z} = \mathbf{X}\mathbf{B} \tag{2}$$

where $\mathbf{Y} \in \mathbb{R}^{h\times w\times L}$, $\mathbf{Z} \in \mathbb{R}^{H\times W\times l}$ and $\mathbf{X} \in \mathbb{R}^{H\times W\times L}$ denote the LR-HSI, the HR-MSI and the reconstructed HR-HSI, respectively. $\mathbf{R} \in \mathbb{R}^{wh\times WH}$ and $\mathbf{B} \in \mathbb{R}^{L\times l}$ are the spatial and spectral degradation matrices, respectively.

Based on this, this article proposes a BRNet that integrates spectral Transformers and depth-wise convolutions, specifically tailored to accommodate the dual spectral–spatial characteristics of hyperspectral data, thereby achieving lightweight and efficient HSI-SR.

As illustrated in Fig. 1, the overall architecture consists of three main components: data preprocessing, feature extraction, and image reconstruction. In the data preprocessing stage, bicubic interpolation is first used to upsample LR-HSI to match the spatial dimensions of HR-MSI, compensating for the low spatial resolution of LR-HSI and ensuring spatial alignment between the two. The two images are then concatenated along the spectral dimension to achieve preliminary fusion in the image domain. This simple yet effective strategy provides the subsequent network with rich input information. In the feature extraction stage, to enhance the effectiveness and stability of the network, feature extraction is divided into shallow and deep branches. A simple convolution operation is initially used to quickly capture shallow spatial and spectral features from the input data, serving as a foundation for subsequent processing. Deep feature extraction is then performed through a designed butterfly recursive learning structure, which alternately applies spectral Transformers and depth-wise convolutions to extract complex features from different scales and dimensions. This design allows the network to effectively capture both fine spectral corrections and detailed spatial structures. To integrate shallow and deep features, residual connections are employed, where the shallow branch preserves coarse global contextual information and the deep branch captures localized textures. This fusion strategy not only ensures spatial consistency but also enhances the network's ability to reconstruct fine-grained details with high fidelity. In reconstruction stage, the designed reconstruction module converts the fused features into a high-resolution image consistent with HR-HSI. This module consists of two convolution layers and an activation layer, progressively recovering spatial and spectral details of the image. Finally, the reconstruction result is fused with the upsampled LR-HSI through residual learning to reduce error accumulation during the reconstruction process, yielding final HR-HSI result.

B. Butterfly Recursion Learning Structure

ResNet resolves vanishing gradients in deep networks via skip connections, enabling direct learning of input–output differences without complex feature mappings, thus enhancing network depth stability [47]. However, affected by the high-dimensional nature of HSI data, standard residual blocks tend to introduce redundant features during the feature connection process, ultimately limiting the model's performance. Existing methods often introduce adaptive learnable parameters in the residual branches to adjust the weights of different feature channels, while this approach helps suppress redundancy between spectral bands, there are still limitations in adaptive feature fusion and network pruning: 1) adaptive parameters are typically initialized randomly and are decoupled from the input features, leading to suboptimal learning efficiency; and 2) adaptive parameters are generally applied to only one branch, limiting the full potential of multibranch feature fusion. To overcome these challenges, this article proposes a butterfly recursive learning structure that enables more efficient feature learning and more compact network design.

To address the first issue, the channel attention mechanism, which generates adaptive weights for each channel based on the characteristics of the input data, inspired the FPGS in this article. Specifically, global average pooling (GAP) and standard deviation pooling (SDP) are first applied to the input features separately. GAP captures the overall information of each channel, ensuring spectral consistency, while SDP captures the internal feature variations within each channel, enhancing the detail differences between bands. Subsequently, operations similar to those in SENet are used to squeeze and excitation (SE) these features, and they are transformed into weight coefficients using a sigmoid function [48]. Finally, the two sets of weight coefficients are combined through a simple weighted fusion to obtain the final fusion parameters, which can better couple the input features. This process can be described as

$$W = \frac{1}{2}\left(\mathrm{SE}\left(\mathrm{GAP}\left(x\right)\right) + \mathrm{SE}\left(\mathrm{SDP}\left(x\right)\right)\right) \tag{3}$$

where W denotes the fusion parameter, GAP$(\cdot)$ and SDP$(\cdot)$ denote the GAP and the SDP, respectively. SE$(\cdot)$ is the transformation process from feature to weight.

Regarding the second issue, simply introducing adaptive parameters in another branch does not effectively address the problem of insufficient feature fusion. Mathematically, this approach is essentially equivalent to the behavior of the original residual network and lacks the capability to deeply explore the complex relationships between multibranch features, thus failing to achieve true multibranch feature fusion. To overcome this limitation, this article draws inspiration from the recursive decomposition structure in fast Fourier transform (FFT). FFT recursively decomposes the input signal into multiple subproblems, significantly improving the efficiency of high-dimensional data processing. Inspired by this, multiple residual blocks are integrated into a recursive structure to capture input features layer by layer. At the same time, the FPGS is employed to generate multiple independent adaptive pruning coefficients, enabling the network to dynamically filter out unimportant redundant features based on changes in the input features, thereby achieving efficient adaptive pruning.

Taking the integration of two residual blocks, $p(\cdot)$ and $q(\cdot)$, as an example, the recursive process of the butterfly block can be described as follows.

The first residual block $p(\cdot)$

$$m_1 = w_1^p p\left(x_1\right) + x_2 \tag{4}$$

$$m_2 = p\left(x_1\right) + w_2^p x_2. \tag{5}$$

The second residual block $q(\cdot)$

$$n_1 = q\left(m_1\right) + w_2^q m_2 \tag{6}$$

$$n_2 = w_1^q q\left(m_1\right) + m_2 \tag{7}$$

where x_i, m_i, n_i denote the input feature, the recursive intermediate variable, and the final output, respectively. w_i^p and w_i^q $i = \{1, 2\}$ are the adaptive pruning coefficients, the superscript p and q represent the corresponding residual block, and the subscript i represents the ith branch.

Finally, the features output by different branches are concatenated along the spectral dimension, and a convolutional layer is applied to further promote deep fusion and interaction between multiple branches. This process can be described as

$$y = \mathrm{Conv}\left(\mathrm{Cat}\left(n_1, n_2\right)\right) \tag{8}$$

where Conv$(\cdot)$ and Cat$(\cdot)$ represent the convolution and concatenation operations, respectively. Compared to the simple cascading of multiple residual blocks, the proposed recursive structure enhances the depth of information processing by capturing hierarchical features, while the FPGS dynamically adjusts the network structure. This enables efficient feature fusion and adaptive pruning, significantly improving computational efficiency and reducing the model size.

C. Token-Level SSI Network

In Section III-B, by introducing a recursive structure and a FPGS, this article designed a butterfly recursive learning structure that effectively alleviates the shortcomings of existing methods in feature fusion and network pruning. Based on this architecture, this section further explores the issue of feature learning within the residual blocks. Specifically, this article innovatively decomposes and optimizes the Transformer structure, proposing two improved feature mixing mechanisms: a token mixer based on the interaction between spectral Transformers and depth-wise convolutions, and a channel mixer based on a CGFF.

Improved Token Mixer: In the Vanilla transformer, the token mixer is the self-attention mechanism, which models long-range dependencies by calculating the similarity between each feature token and all other tokens [49]. Additionally, since the self-attention mechanism does not rely on a fixed receptive field, it performs particularly well when processing data with

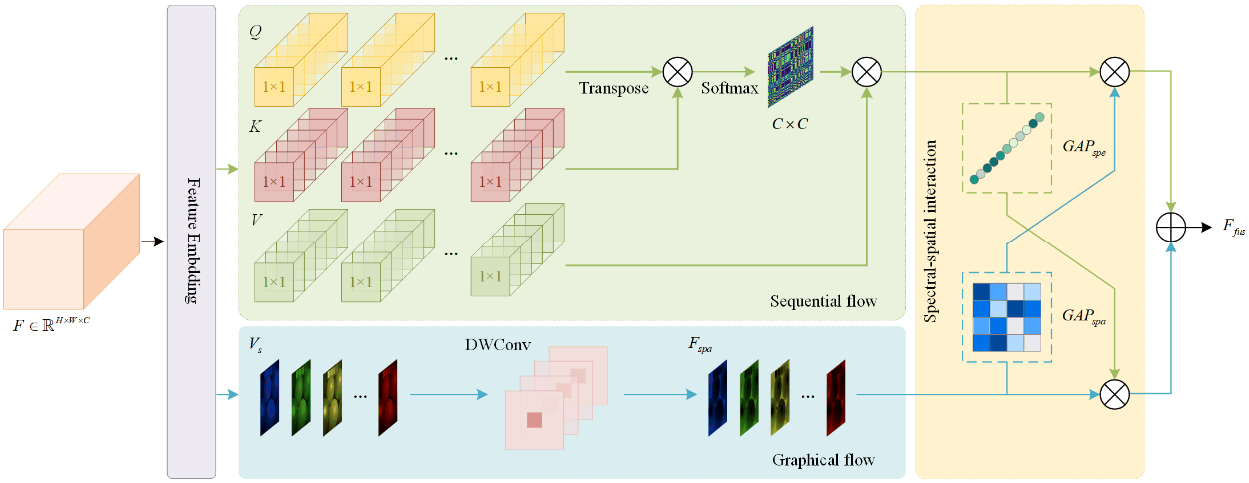

Fig. 2. Network structure of improved token mixer, which consists of a spectral feature branch (sequential flow), a spatial feature branch (graphical flow), and the SSI module.

sequential characteristics, making it especially suitable for modeling spectral continuity in hyperspectral data. These properties effectively preserve spectral fidelity in HSI-SR tasks, which have led to widespread application and attention for the self-attention mechanism in this field. However, despite the clear advantages of the self-attention mechanism in the spectral domain, it faces two major limitations in spatial reconstruction: 1) its ability to capture local detail features is limited, often resulting in detail loss; and 2) it cannot fully exploit spatial structural information, leading to insufficient integration of spatial features. To address these shortcomings, as shown in Fig. 2, this article introduces depth-wise convolution into the original token mixer as a parallel branch to the self-attention mechanism. Depth-wise convolution, with its strong capability for local feature extraction, effectively compensates for the self-attention mechanism's weaknesses in spatial feature processing, allowing for better capture of image detail features.

Specifically, for a given input feature $F \in \mathbb{R}^{H\times W\times C}$, it is first linearly mapped through the feature embedding module into the sequential data $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V} \in \mathbb{R}^{HW\times C}$ required for self-attention calculations, while generating image-formatted features $\mathbf{V}_s \in \mathbb{R}^{H\times W\times C}$ for the depth-wise convolution branch. In the spectral feature branch (sequential flow), a band-by-band self-attention calculation is employed to effectively explore the self-similarity between spectral bands and reduce the computational cost of self-attention. This process can be described as

$$F_{\text{spe}} = \mathbf{V}\cdot \text{Softmax}\left(\frac{\mathbf{Q}^T\mathbf{K}}{\beta}\right) \tag{9}$$

where F_{spe} is the spectral features, β is a learnable temperature parameter to adjust the smoothness of the attention distribution.

In the spatial feature branch (graphical flow), depth-wise convolution is used to learn the spatial features of each band in the input data. This process can be described as follows:

$$F_{\text{spa}} = \text{DWConv}\left(\mathbf{V}_s\right) \tag{10}$$

where F_{spa} represents the spatial feature representation, and DWConv$(\cdot)$ refers to the depth-wise convolution operation, which is capable of better capturing spatial details and local structural information.

To further enhance the interaction and fusion between spatial and spectral features, this article introduces a SSI module. This module is based on an attention mechanism, incorporating both spatial and spectral attention derived from GAP. It dynamically adjusts and fuses the aforementioned spatial and spectral features. The calculation process can be expressed as

$$F_{\text{fus}} = F_{\text{spe}} \odot \sigma\left(\text{GAP}_{\text{spa}}\left(F_{\text{spa}}\right)\right) + F_{\text{spa}} \odot \sigma\left(\text{GAP}_{\text{spe}}\left(F_{\text{spe}}\right)\right) \tag{11}$$

where F_{fus} represents the fused features, $\odot$ denotes element-wise multiplication, and $\sigma(\cdot)$ represents the sigmoid function. This fusion strategy dynamically adjusts the weights of the spatial and spectral features, ensuring effective interaction and flow between them, thereby enhancing the network's ability to jointly learn spatial and spectral information.

Improved Channel Mixer: In the Vanilla transformer, the channel mixer is the feed-forward network (FFN), which primarily operates through a series of linear transformations and nonlinear activation functions to process input features at each position, enhancing the network's expressive power [49]. However, FFN has significant limitations in HSI-SR tasks. First, since FFN performs position-wise transformations on each token, it cannot capture contextual information between neighboring pixels, leading to insufficient learning of local features. Second, FFN processes each pixel's features independently and cannot adaptively adjust according to the characteristics of input data, resulting in wasted computational resources. To address these issues, recent studies have shown that incorporating a 3 × 3 depth-wise convolution into FFN acts as a form of conditional encoding, effectively improving the extraction of local features [50]. Additionally, the gated linear unit (GLU) has been demonstrated to outperform the multilayer perceptron (MLP) in several natural language processing tasks. Specifically, GLU relies on the token itself to

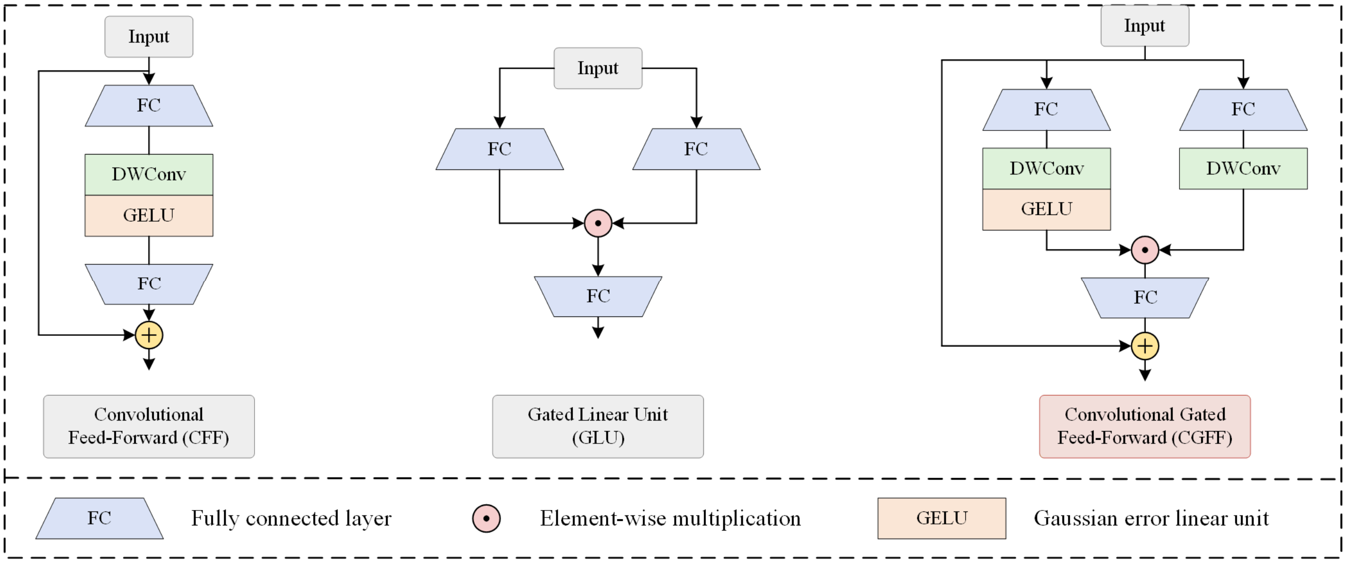
Input
FC
DWConv
GELU
FC
Convolutional Feed-Forward (CFF)
Input
FC
FC
FC
Gated Linear Unit (GLU)
Input
FC
FC
DWConv
GELU
DWConv
FC
Convolutional Gated Feed-Forward (CGFF)
FC
Fully connected layer
Element-wise multiplication
GELU
Gaussian error linear unit

Fig. 3. Overall structure of the CGFF, and comparison with other channel mixer designs: convolutional feed-forward and GLU.

generate both a main signal and a gating signal through two parallel linear transformations, dynamically adjusting the flow of the main signal using the gating signal, which significantly reduces the computational cost and learning complexity. Based on these insights, this article incorporates a 3 × 3 depth-wise convolution layer after each linear transformation layer in the original GLU structure, constructing a CGFF. This network elegantly integrates the advantages of both depth-wise convolution and the gating mechanism, as shown in Fig. 3. For a given feature input $F_{\text{fus}} \in \mathbb{R}^{H \times W \times C}$, the calculation process is described as

$$\text{Gated}\left(F_{\text{fus}}\right) = \varphi\left(\text{DWConv}\left(\text{FC}\left(F_{\text{fus}}\right)\right)\right) \tag{12}$$

$$F_o = \text{FC}\left(\text{Gated}\left(F_{\text{fus}}\right) \odot \text{DWConv}\left(\text{FC}\left(F_{\text{fus}}\right)\right)\right) \tag{13}$$

where F_o represents the output of the CGFF, Gated$(\cdot)$ refers to the gated branch, FC$(\cdot)$ represents the fully connected layer, and $\varphi(\cdot)$ denotes the GELU nonlinear activation function.

IV. Experimental Results

This section presents a detailed explanation of the experimental results to verify the effectiveness of the proposed method. Initially, the experimental setup is introduced, including a description of the datasets used, the data simulation process, and implementation details. Next, the selection of comparative methods and the standards for quantitative evaluation metrics are explained; then, the reconstruction performance on three public datasets is presented, with comparisons made to the current state-of-the-art algorithms, followed by a brief analysis. Finally, ablation experiments are conducted to further validate the effectiveness of each component of the proposed method, thoroughly demonstrating its advantages in the super-resolution reconstruction task.

A. Experimental Configurations

1) Datasets: To demonstrate the effectiveness of the proposed method, comparative experiments are conducted on three publicly available hyperspectral datasets: the CAVE dataset [51], the Harvard dataset [52], and the Chikusei dataset [53]. The first two datasets consist of HSIs captured in controlled experimental environments at close range, while the last dataset includes real satellite remote sensing data covering large surface areas, allowing for a comprehensive evaluation of the model's adaptability in remote sensing scenarios. The CAVE dataset was captured using a cooled charge-coupled device (CCD) camera with a spectral interval of 10 nm, consisting of 32 HSIs of various indoor scenes. Each image has a spatial resolution of 512 × 512 and includes 31 spectral bands, covering the visible spectrum from 400 to 700 nm. The Harvard dataset was collected using a Nuance FX camera with a 10 nm spectral interval and consists of 50 HSIs captured under natural light and 27 images taken under artificial or mixed light sources. Each image has a spatial resolution of 1392 × 1040, with 31 spectral bands covering the visible spectrum from 420 to 720 nm. The Chikusei dataset was acquired using a Headwall Hyperspec-VNIR-C hyperspectral sensor and includes a single hyperspectral remote sensing image covering agricultural and urban areas surrounding Chikusei city in Japan. The image has a spatial resolution of 2.5 m, a size of 2517 × 2335 pixels, and includes 128 spectral bands, covering the visible and near-infrared spectrum from 343 to 1018 nm.

2) Data Simulation: The proposed method is based on supervised learning, which requires real HR-HSI as guidance for network training. However, in practical applications, obtaining real HR-HSI is challenging. Therefore, in the experiments, the original HSI is treated as ground truth, and LR-HSI and HR-MSI are generated through data simulation. Specifically, the performance evaluation (Section IV-C) is conducted under two scenarios: experimental settings and real remote sensing scenarios. Details of data processing are as follows: under experimental settings, verification is conducted based on the CAVE and Harvard datasets. Since these two datasets share the same number of spectral bands and similar spectral coverage, 20 images from the CAVE dataset are randomly selected as the training set, while the remaining 11 images, together with 20 randomly selected images from the Harvard dataset, are used to form the test set to evaluate the effectiveness and generalization ability of the proposed method. To increase the number of training samples, the training images from the CAVE dataset are cropped into 4275 overlapping patches of size 64 × 64 × 31. In the data simulation process, spatial degradation is simulated using a 3 × 3 Gaussian blur kernel with a standard deviation of 0.5, followed by downsampling with a scale factor of 4 to obtain LR-HSI of size 16 × 16 × 31. Spectral degradation is simulated using the spectral response function of a Nikon D700 camera, resulting in HR-MSI of size 64 × 64 × 3. In real remote sensing scenarios, the Chikusei dataset is used. The zero-value regions at the edges of the image are first cropped, retaining the central region of 2048 × 2048 pixels. The upper half of the image (1536 × 2048 pixels) is cropped to generate 2961 overlapping patches of size 64 × 64 × 128, which are used as the training set; the lower half (512 × 2048 pixels) is cropped to obtain 4 nonoverlapping patches of size 512 × 512 × 128, which are used as the test set. In the data simulation process, the spatial degradation is the same as for the CAVE dataset, while the spectral degradation is simulated using an IKONOS-like reflectance spectral response filter, resulting in HR-MSI of size 64 × 64 × 4. In addition, to simplify the experimental process, the ablation experiments (Section IV-D) are conducted solely on the CAVE dataset.

3) Implementation Details: The proposed network is implemented using Pytorch 1.13.1 and Python 3.7.16, and the experiments are conducted on a Windows operating system with an NVIDIA GeForce RTX4080 GPU. The network parameters are initialized using the method described in [54], with the number of channel mappings in the shallow feature extraction set to $C = 48$. To train the proposed network, the commonly used content loss function [55] for HSI-SR tasks is applied for supervised learning, and the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) is used for parameter optimization. The initial learning rate is set to 4×10^{-4}, and it is halved after every 50 epochs, with the entire training process lasting 500 epochs.

B. Benchmark and Metrics

To comprehensively validate the effectiveness of the proposed method, six state-of-the-art HSI-SR methods are selected for comparative analysis with a unified scale factor of 4. Specifically, the experiment includes three traditional methods with well-established theoretical foundations and widespread application: CNMF [33], Hysure [31], and FUSE [32]. These traditional methods provide a reliable benchmark for comparing the performance of DL-based methods. Additionally, five cutting-edge DL-based methods are selected: SSRNet [41], S3-Net [56], Fusformer [26], PSRT [44], and LGCT [57]. SSRNet and S3-Net represents a typical application of convolutional networks, while Fusformer, PSRT, and LGCT facilitate a more direct performance comparison with the proposed Transformer-based method.

For evaluation metrics, six quantitative measures are adopted: peak signal-to-noise ratio (PSNR), spectral angle mapper (SAM), erreur relative globale adimensionnelle de synthèse (ERGAS), structural similarity index measure (SSIM), correlation coefficient (CC), and root mean square error (RMSE), ensuring a comprehensive assessment of the reconstructed image quality from multiple dimensions. PSNR and SSIM are primarily used to evaluate the spatial reconstruction quality and structural similarity, reflecting the fidelity of the images. SAM focuses on the consistency of spectral information, making it an important metric for HSI quality assessment. ERGAS normalizes the relative error for each band, providing a comprehensive error analysis in both spectral and spatial dimensions. CC evaluates the correlation between the reconstructed image and the reference image, while RMSE measures the absolute error between the images. These multidimensional evaluation metrics enable a systematic quantitative analysis of each method's performance in terms of spatial, spectral, and structural aspects, ensuring the comprehensiveness and scientific rigor of the experimental results.

C. Performance Evaluation

1) Experimental Results on the CAVE Dataset: From a quantitative perspective, Table I shows the quantitative comparison results of the seven methods on the CAVE dataset, where the best results are highlighted in bold and the second-best results are underlined. Compared to traditional methods, DL-based methods achieve more reliable results due to their strong inductive bias. Transformer networks, with their superior ability to model long-range dependencies, capture more global features and more effectively predict missing information during the reconstruction process. Specifically, the proposed method achieves the best results across all six quantitative evaluation metrics, with significantly fewer network parameters than other methods. Notably, compared to the second-best method, the proposed method improves PSNR by 0.27 dB and reduces SAM by 0.07, clearly demonstrating its significant advantages in both spatial and spectral domains.

From the qualitative results shown in Fig. 4, the pseudocolor images of the reconstruction results by different methods are presented in the spatial dimension, along with the error maps compared to the ground truth (generated from a randomly selected band). The highlighted regions (yellow boxes) are used to compare the details. As observed, the pseudocolor map generated by the proposed method is clearer, with richer detail restoration. Moreover, in the error map, the proposed method exhibits the smallest error, further confirming its superiority in spatial reconstruction accuracy. In the spectral dimension, the spectral angle maps of each method are shown, where the proposed method demonstrates a smaller overall spectral error. Additionally, to compare spectral differences in more detail, Fig. 5 presents the spectral curve of a randomly selected pixel, which nearly coincides with the ground truth, further validating the excellent spectral fidelity of the proposed method.

To evaluate the computational burden of different fusion methods, Table II reports the training time, testing time, floating point operations (FLOPs), and number of parameters

TABLE I
AVERAGE QUANTITATIVE RESULT AND CORRESPONDING PARAMETERS ON THE CAVE, HARVARD AND CHIKUSEI DATASET WITH THE SCALE FACTOR OF 4

Dataset	Metrics	CNMF [33]	Hysure [31]	FUSE [32]	SSRNet [41]	Fusformer [26]	PSRT [44]	LGCT [57]	S3-Net [56]	BRNet (Proposed)
CAVE	PSNR↑	42.46	43.26	41.43	47.36	50.77	47.24	51.01	50.82	**51.28**
	SAM↓	8.17	6.63	4.30	3.14	2.26	2.71	2.18	2.71	**2.11**
	ERGAS↓	2.8420	2.5709	3.1071	1.9763	1.0540	1.5990	1.0188	1.0375	**0.9859**
	SSIM↑	0.9772	0.9777	0.9838	0.9931	0.9967	0.9948	0.9968	0.9965	**0.9970**
	CC↑	0.9961	0.9976	0.9967	0.9985	0.9994	0.9990	0.9994	0.9994	**0.9995**
	RMSE↓	0.0090	0.0080	0.0100	0.0051	0.0035	0.0048	0.0034	0.0035	**0.0033**
	# params	/	/	/	**0.03M**	0.50M	0.25M	5.16M	0.07M	0.19M
Harvard	PSNR↑	47.11	48.25	46.93	48.36	50.59	50.15	49.38	49.56	**50.72**
	SAM↓	2.99	3.17	3.08	4.08	2.73	2.79	2.82	2.89	**2.72**
	ERGAS↓	2.7173	2.5636	2.6469	8.4567	2.1832	2.1381	2.4044	2.3488	**2.1271**
	SSIM↑	0.9875	0.9878	0.9862	0.9807	0.9922	0.9915	0.9904	0.9906	**0.9924**
	CC↑	0.9783	0.9849	0.9796	0.9750	0.9838	0.9826	**0.9850**	0.9849	0.9843
	RMSE↓	0.0057	0.0046	0.0057	0.0049	**0.0034**	0.0036	0.0040	0.0040	**0.0034**
	# params	/	/	/	**0.03M**	0.50M	0.25M	5.16M	0.07M	0.19M
Chikusei	PSNR↑	48.07	47.52	46.26	57.96	52.88	52.40	56.78	56.45	**59.30**
	SAM↓	1.80	1.93	1.99	1.04	1.21	1.21	1.07	1.08	**0.96**
	ERGAS↓	2.8880	2.9200	3.1647	1.6981	1.8127	2.1973	1.8269	1.8639	**1.6530**
	SSIM↑	0.9927	0.9916	0.9899	0.9979	0.9965	0.9965	0.9976	0.9975	**0.9980**
	CC↑	0.9865	0.9895	0.9855	**0.9945**	0.9937	0.9914	0.9939	0.9937	**0.9945**
	RMSE↓	0.0048	0.0051	0.0056	0.0020	0.0026	0.0027	0.0020	0.0021	**0.0018**
	# params	/	/	/	0.44M	0.55M	0.30M	5.37M	1.07M	**0.28M**

TABLE II
INFERENCE EFFICIENCY OF DIFFERENT FUSION METHODS ON THE CAVE DATASET WITH A SCALE FACTOR OF 4

Method	CAVE			
	Training time	Testing time	FLOPs	#params
CNMF[33]	/	161.26 s	/	/
Hysure[31]	/	2418.31 s	/	/
FUSE[32]	/	11.72 s	/	/
SSRNet[41]	1.82×10^4 s	1.97 s	0.32G	0.03M
Fusformer[26]	4.37×10^4 s	4.51 s	1.37G	0.50M
PSRT[44]	3.62×10^4 s	4.09 s	3.15G	0.25M
LGCT[57]	3.96×10^4 s	4.47 s	4.33G	5.16M
S3-Net[56]	5.02×10^4 s	3.15 s	0.31G	0.07M
BRNet	1.19×10^4 s	3.86 s	2.28G	0.19M

on the CAVE dataset. The results clearly demonstrate that traditional methods incur significantly higher testing times compared to DL-based approaches. Although Transformer architectures are generally more computationally intensive, and thus our proposed method does not achieve the lowest training or testing time when compared to CNN-based models such as SSRNet and S3-Net, it exhibits substantially lower computational cost relative to other Transformer-based methods, including Fusformer, PSRT, and LGCT. Specifically, the proposed model achieves shorter inference times while maintaining competitive FLOPs and parameter counts. These findings substantiate the effectiveness of the proposed BRNet and its architectural components, which strike a favorable balance between reconstruction accuracy and computational efficiency.

2) Experimental Results on the Harvard Dataset: While DL-based methods typically achieve better results, their generalization issues cannot be overlooked. Especially in the context of high-cost hyperspectral data acquisition and limited training datasets, deep models for HSI-SR tasks are prone to overfitting. To explore this issue, the deep models are trained on the CAVE dataset and tested for generalization on the Harvard dataset. Table I presents the quantitative results of the comparison methods on the Harvard dataset, showing that the performance gap between traditional methods and DL methods has narrowed. However, Transformer-based models still exhibit a clear advantage, with the proposed method only differs from the optimal method by 0.0007 in the quantitative metric of CC, while achieving the best reconstruction results across all quantitative evaluation metrics, demonstrating its superior robustness and generalization ability across different datasets.

In terms of visual evaluation, a test image is selected from the Harvard dataset for similar experiments, with the results shown in Fig. 4. Fig. 4 displays the pseudo-color images, the error maps, and the spectral angle maps of the reconstruction results from different methods, comparing the overall performance in spatial reconstruction and spectral preservation. Fig. 5 further illustrates the spectral curves of the reconstruction results to provide a more detailed comparison of the spectral fidelity between the methods and the ground truth. The experimental results show that the proposed method achieves significant advantages in both spatial reconstruction

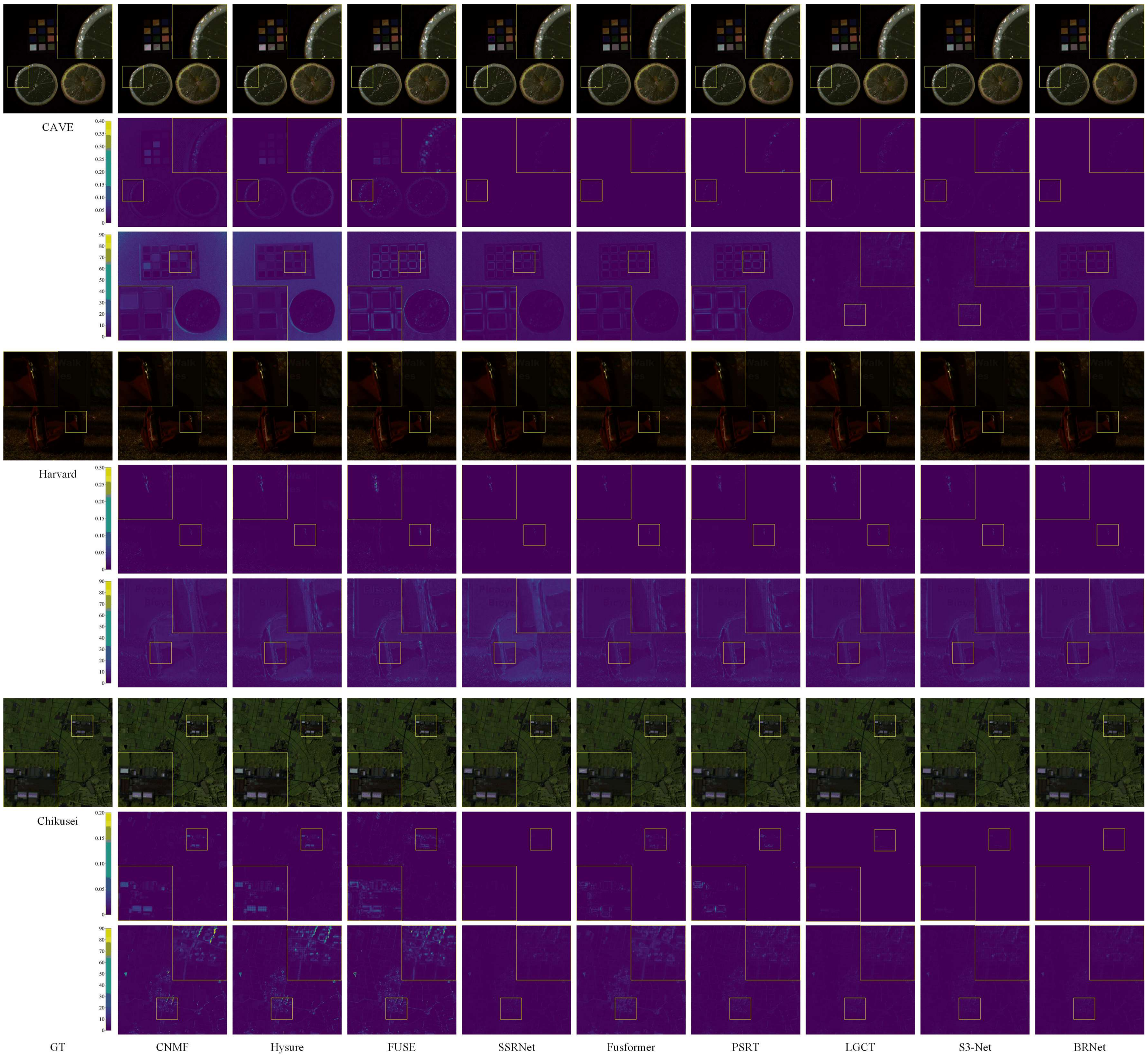

Fig. 4. Visual evaluation results on CAVE (rows 1–3), Harvard (rows 4–6), and Chikusei (rows 7–9) datasets. The first row of each dataset shows the pseudo-color images of reconstructed outputs, the second row presents the heatmap relative absolute error, and the third row displays the heatmap of SAM error.

and spectral information retention, consistent with the quantitative analysis results.

3) Experimental Results on the Chikusei Dataset: Both CAVE and Harvard datasets consist of simulated data collected from a close distance. While these datasets provide a controlled testing environment, they are limited in replicating the complexity of real-world application scenarios. Therefore, this article further conducts experiments on a real-world satellite hyperspectral dataset, the Chikusei dataset. Compared to simulated data, satellite imagery is more significantly affected by factors such as atmospheric interference, leading to higher noise levels. Additionally, the ground objects in the scene are more complex, and spectral mixing becomes more pronounced, which increases the challenges of HSI-SR. Table I presents the quantitative results for the various methods on the Chikusei dataset. As shown in the table, Transformer-based methods exhibit varying degrees of performance degradation, due to the fact that deeper models facing overfitting or insufficient feature learning when dealing with more complex and sparse remote sensing data. However, despite these challenges, the proposed method achieves the best results across all quantitative evaluation metrics, further demonstrating the robustness and superiority of the proposed network structure when applied to real, complex data. Additionally, Figs. 4 and 5 present the pseudo-color images, error maps, spectral angle maps, and spectral curves for all methods on Chikusei dataset. The results show that the proposed method not only outperforms others in terms of visual quality but also exhibits

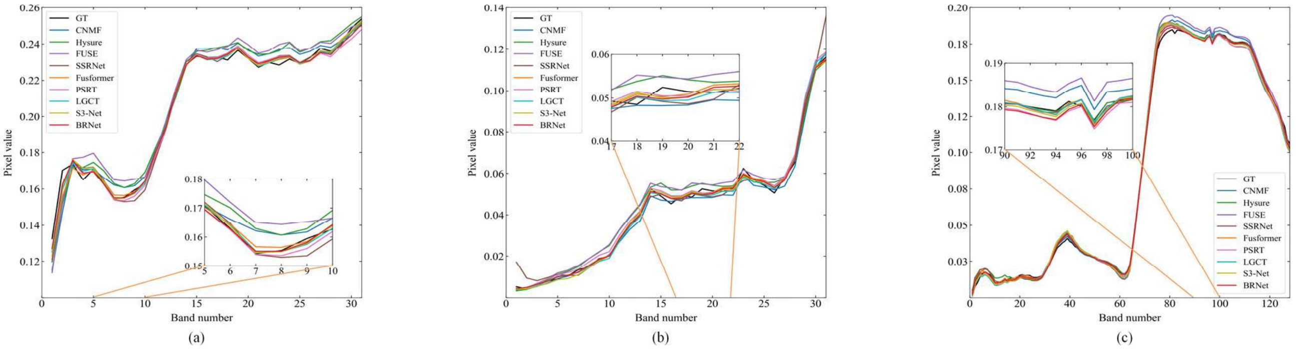

Fig. 5. Illustration for the spectral curves of the reconstruction results. (a) CAVE dataset: *lemon slices* located at position (307, 205). (b) Harvard dataset: *imgf6* located at position (273, 474). (c) Chikusei dataset: *area0* located at position (475, 394).

superior spectral fidelity compared to the ground truth, in alignment with the quantitative analysis.

D. Ablation Study

This section evaluates the contribution of each component of the proposed method through a series of ablation experiments, focusing on three key aspects: the effectiveness of the butterfly recursive learning structure, the effectiveness of the FPGS, and the effectiveness of key modules in the token-level SSI network. To simplify the experimental process and ensure the generality of the results, all ablation experiments are trained and tested on the CAVE dataset. This provides a detailed analysis of the impact of different components on reconstruction performance, allowing us to explore how each factor influences the reconstruction results and enabling a more precise evaluation of the overall performance of the proposed method.

1) Effectiveness of the Butterfly Recursion Learning Structure: To validate the advantages of the proposed recursive structure in addressing the challenge of depth degradation, a comparative analysis is conducted with the simple cascade connection approach. Specifically, to ensure the fairness of the comparison, the experiment adopts the improved Transformer network proposed in Section III-C as the feature extraction module while keeping the overall network framework shown in Fig. 1 unchanged, aiming to explore the performance differences under the two different connection methods. For the cascade connection method, four comparative methods are designed, where 2, 4, 6, and 8 feature extraction modules are cascaded via skip connections (denoted as cascade 2, cascade 4, cascade 6, and cascade 8 in sequence). For the recursive connection method, this article constructs the butterfly recursive learning structure (Butterfly) in Fig. 1 with 2 feature extraction modules.

The results shown in Table III indicate that the proposed recursive structure achieves the best performance in terms of quantitative metrics such as PSNR, SAM, and ERGAS. For the cascade connection approach, as illustrated in Fig. 6, these metrics first improve then deteriorate as module count increases. Specifically, cascade 6 performs best, while cascade 8's metrics start to decline, further confirming that simple stacking of standard residual blocks is insufficient to alleviate deep degradation's negative impact on HR-HSI tasks. Although Butterfly and cascade 6 exhibit similar performance across metrics, the former significantly reduces the model's parameter count, demonstrating higher efficiency and practicality.

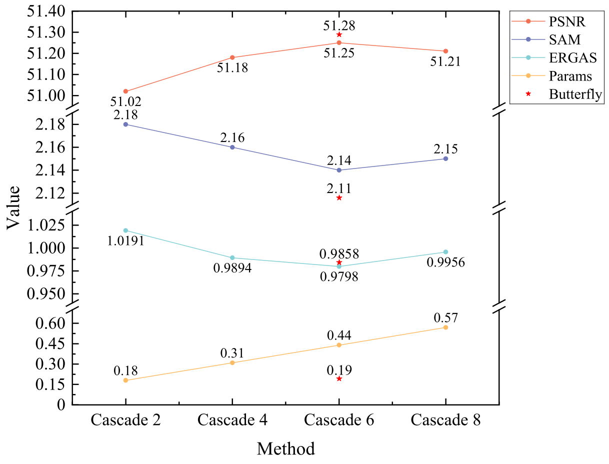

Fig. 6. Illustration for comparison of different connection methods.

TABLE III
QUANTITATIVE RESULTS ON THE EFFECTIVENESS OF THE BUTTERFLY RECURSION LEARNING STRUCTURE ON THE CAVE DATASET

Method	PSNR↑	SAM↓	ERGAS↓	# params
Cascade 2	51.02	2.18	1.0191	**0.18M**
Cascade 4	51.18	2.16	0.9894	0.31M
Cascade 6	<u>51.25</u>	<u>2.14</u>	**0.9798**	0.44M
Cascade 8	51.21	2.15	0.9956	0.57M
Butterfly	**51.28**	**2.11**	<u>0.9858</u>	<u>0.19M</u>

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

2) Effectiveness of the FPGS: To validate the effectiveness of GAP and SDP in the FPGS, four different combinations of the model are tested: "w/o GAP and w/o SDP," "w GAP and w/o SDP," "w/o GAP and w SDP," and "w GAP and w SDP." In the "w/o GAP and w/o SDP" case, the parameters are initialized randomly.

TABLE IV

QUANTITATIVE RESULTS ON EFFECTIVENESS OF THE FPGS ON THE CAVE DATASET

Components	Different Combinations of Components			
GAP	×	√	×	√
SDP	×	×	√	√
PSNR↑	51.11	51.18	51.12	**51.28**
SAM↓	2.13	2.12	2.15	**2.11**
ERGAS↓	1.0043	1.0029	0.9961	**0.9858**
SSIM↑	0.9968	0.9969	0.9969	**0.9970**
CC↑	0.9994	0.9994	0.9995	**0.9995**
RMSE↓	0.0034	0.0034	0.0034	**0.0033**

The best values are highlighted in bold.

TABLE V

QUANTITATIVE RESULTS ON EFFECTIVENESS OF KEY MODULES IN THE TOKEN-LEVEL SSI NETWORK ON THE CAVE DATASET

Components	Different Combinations of Components			
SSI	×	√	×	√
CGFF	×	×	√	√
PSNR↑	50.98	51.03	51.18	**51.28**
SAM↓	2.20	2.19	2.15	**2.11**
ERGAS↓	1.0303	1.0376	0.9949	**0.9858**
SSIM↑	0.9968	0.9968	0.9968	**0.9970**
CC↑	0.9994	0.9994	0.9995	**0.9995**
RMSE↓	0.0035	0.0035	0.0034	**0.0033**

The best values are highlighted in bold.

As shown in Table IV, the method with both GAP and SDP achieves the best performance across all quantitative metrics. In contrast, using only GAP or SDP improves the performance to some extent but falls short compared to the strategy of using both, especially in metrics like SSIM and SAM. This indicates that relying solely on one pooling method is insufficient to capture the complex information in images comprehensively. The strategy with random initialization performs the worst, further emphasizing the importance of establishing a connection between parameter learning and the input data.

3) Effectiveness of Key Modules in the Token-level SSI Network: To verify the effectiveness of key modules in the token-level SSI network, this study explores the impact of different combinations of the SSI and CGFF modules on model performance. Specifically, four module combinations are tested: "w/o SSI and w/o CGFF," "w SSI and w/o CGFF," "w/o SSI and w CGFF," and "w SSI and w CGFF."

As shown in Table V, the strategy using both SSI and CGFF achieves the best performance across all quantitative metrics, particularly with significant improvements in key indicators like PSNR and SAM. In contrast, using only one of these modules results in some performance gains but does not match the effectiveness of using both. This is especially evident in metrics such as SSIM and ERGAS, where models using only one module show clear limitations, indicating that SSI and CGFF play complementary roles in the interaction and fusion of spatial and spectral information. The strategy without either module performs the worst, further demonstrating the necessity of these key modules for enhancing model performance.

V. CONCLUSION

This article proposes a BRNet based on spectral Transformers and depth-wise convolutions to address the critical challenges of depth degradation in HSI-SR tasks. Specifically, this method integrates a recursive structure with FPGS as the backbone for adaptive information propagation, while employing a spectral Transformer and depth-wise convolutions as complementary feature extractors—the former captures cross-band spectral dependencies, and the latter focuses on spatial detail representation. The SSI module further connects these dimensional features to generate token-level representations that effectively preserve spectral fidelity and spatial details. Finally, a CGFF is utilized to enhance local interactions between tokens, enriching the detail information in the reconstruction results. Experimental results demonstrate that the proposed method achieves superior reconstruction performance over existing state-of-the-art methods on multiple benchmark datasets, validating its effectiveness of the designed modules.

Although the method proposed in this article effectively reduces the model complexity through recursive structures and gating mechanisms, limited by the high-dimensional characteristics of HSIs and the inherent computational overhead of the multihead self-attention mechanism, the network's demand for computing resources remains at a high level, making it difficult for the model to achieve optimal performance in scenarios such as real-time processing and high-magnification super-resolution. In addition, because the recursive structure relies on the output of the previous module as the input of the subsequent module, when there are complex spatial-spectral mutation phenomena in the super-resolution scenario, the wrong feature estimation of the mutation region by the previous module will be amplified during the recursive process, leading to the accumulation of errors in the subsequent fusion, and future research can focus on exploring the adaptive redundancy pruning strategy of self-attention and the dynamic feedback adjustment mechanism of front and rear features to promote the development of HSI-SR technology toward a more efficient and universal direction.

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, 2022, Art. no. 5527713.

[2] 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, 2024.

[3] J. Shi, T. Wu, A. K. Qin, T. Shao, Y. Lei, and G. Jeon, "Deep-growing neural network with manifold constraints for hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 1, pp. 210–221, Jan. 2025.

[4] P. Ghamisi et al., "New frontiers in spectral–spatial hyperspectral image classification: The latest advances based on mathematical morphology, Markov random fields, segmentation, sparse representation, and deep learning," *IEEE Geosci. Remote Sens. Mag.*, vol. 6, no. 3, pp. 10–43, Sep. 2018.

[5] X. Yang, B. Tu, Q. Li, J. Li, and A. Plaza, "Graph evolution-based vertex extraction for hyperspectral anomaly detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 12, pp. 17372–17386, Dec. 2024.

[6] L. Chen, J. He, H. Shi, J. Yang, and W. Li, "SWDiff: Stage-wise hyperspectral diffusion model for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5536217.

[7] 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.

[8] 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, 2023, Art. no. 5510515.

[9] 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, 2023, Art. no. 5512012.

[10] J. Liu, Z. Wu, and L. Xiao, "A spectral diffusion prior for unsupervised hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5528613.

[11] R. Dian, Y. Liu, and S. Li, "Hyperspectral image fusion via a novel generalized tensor nuclear norm regularization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 4, pp. 7437–7448, Apr. 2025.

[12] Q. Hu, X. Wang, J. Jiang, X.-P. Zhang, and J. Ma, "Exploring the spectral prior for hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 33, pp. 5260–5272, 2024.

[13] J. Jia, H. Yu, C. Wang, K. Zheng, J. Li, and J. Hu, "Spectral–spatial collaborative pretraining framework with multi-constraint cooperation for hyperspectral-multispectral image fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, early access, Apr. 17, 2025, doi: 10.1109/JSTARS.2025.3562278.

[14] 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, 2022, Art. no. 5527913.

[15] K. Li, L. Van Gool, and D. Dai, "Test-time training for hyperspectral image super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Sep. 23, 2024, doi: 10.1109/TPAMI.2024.3461807.

[16] Y. Yang et al., "Spectral-enhanced sparse transformer network for hyperspectral super-resolution reconstruction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 17278–17291, 2024.

[17] Q. Ma, J. Jiang, X. Liu, and J. Ma, "Learning a 3D-CNN and transformer prior for hyperspectral image super-resolution," *Inf. Fusion*, vol. 100, pp. 1–12, Dec. 2023.

[18] Z. Lai, Y. Fu, and J. Zhang, "Hyperspectral image super resolution with real unaligned RGB guidance," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 2, pp. 2999–3011, Feb. 2025.

[19] S. Li, Y. Tian, C. Wang, H. Wu, and S. Zheng, "Hyperspectral image super-resolution network based on cross-scale nonlocal attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5509615.

[20] K. Zheng et al., "Coupled convolutional neural network with adaptive response function learning for unsupervised hyperspectral super resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2487–2502, Mar. 2021.

[21] J.-F. Hu, T.-Z. Huang, L.-J. Deng, T.-X. Jiang, G. Vivone, and J. Chanussot, "Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 7251–7265, Dec. 2022.

[22] Y. Sun et al., "Dual spatial–spectral pyramid network with transformer for hyperspectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5526016.

[23] 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, pp. 1–22, Nov. 2017.

[24] 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, 2020.

[25] B. Wang, S. Mei, Y. Feng, and Q. Du, "Hyperspectral imagery super-resolution based on self-calibrated attention residual network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. IGARSS*, Jul. 2021, pp. 3896–3899.

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

[27] A. Dosovitskiy, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2021, pp. 1–22.

[28] S. Zhang, H. Liu, S. Lin, and K. He, "You only need less attention at each stage in vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 6057–6066.

[29] 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.

[30] G. Vivone, R. Restaino, and J. Chanussot, "Full scale regression-based injection coefficients for panchromatic sharpening," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3418–3431, Jul. 2018.

[31] 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.

[32] 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.

[33] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 2, pp. 528–537, Feb. 2012.

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

[35] R. Dian, L. Fang, and S. Li, "Hyperspectral image super-resolution via non-local sparse tensor factorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3862–3871.

[36] C. I. Kanatsoulis, X. Fu, N. D. Sidiropoulos, and W.-K. Ma, "Hyperspectral super-resolution: A coupled tensor factorization approach," *IEEE Trans. Signal Process.*, vol. 66, no. 24, pp. 6503–6517, Dec. 2018.

[37] S. Xu, O. Amira, J. Liu, C.-X. Zhang, J. Zhang, and G. Li, "HAM-MFN: Hyperspectral and multispectral image multiscale fusion network with RAP loss," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4618–4628, Jul. 2020.

[38] H. Wu, J. Gui, Y. Xu, Z. Wu, Y. Y. Tang, and Z. Wei, "An efficient cross-modality self-calibrated network for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5545312.

[39] Z. Zhu, J. Hou, J. Chen, H. Zeng, and J. Zhou, "Hyperspectral image super-resolution via deep progressive zero-centric residual learning," *IEEE Trans. Image Process.*, vol. 30, pp. 1423–1438, 2021.

[40] J. Li, K. Zheng, L. Gao, L. Ni, M. Huang, and J. Chanussot, "Model-informed multistage unsupervised network for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5516117.

[41] 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.

[42] R. Ran, L.-J. Deng, T.-J. Zhang, J. Chang, X. Wu, and Q. Tian, "KNLConv: Kernel-space non-local convolution for hyperspectral image super-resolution," *IEEE Trans. Multimedia*, vol. 26, pp. 8836–8848, 2024.

[43] E. Zhao, N. Qu, Y. Wang, and C. Gao, "Spectral reconstruction from thermal infrared multispectral image using convolutional neural network and transformer joint network," *Remote Sens.*, vol. 16, no. 7, pp. 1–17, Apr. 2024.

[44] S.-Q. Deng, L.-J. Deng, X. Wu, R. Ran, D. Hong, and G. Vivone, "PSRT: Pyramid shuffle-and-reshuffle transformer for multispectral and hyperspectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5503715.

[45] Q. Ma, J. Jiang, X. Liu, and J. Ma, "Reciprocal transformer for hyperspectral and multispectral image fusion," *Inf. Fusion*, vol. 104, Apr. 2024, Art. no. 102148.

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

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

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

[49] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2017, pp. 1–11.

[50] 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, 2024.

[51] 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.

[52] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," in *Proc. CVPR*, Jun. 2011, pp. 193–200.

[53] N. Yokoya and A. Iwasaki, *Airborne hyperspectral data over Chikusei*, Dept. Space Appl. Lab., Univ. Tokyo, Tokyo, Japan, Tech. SAL-2016-05-27, May 2016.

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[55] J. Hu, Y. Tang, Y. Liu, and S. Fan, "Hyperspectral image super-resolution based on multiscale mixed attention network fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[56] X. Wang, Z. Huang, J. Zhu, X. Wang, and L. Feng, "S3-Net: Learning spectral-spatio self-similarity for hyperspectral image super-resolution," *Neural Netw.*, vol. 188, pp. 1–11, Aug. 2025.

[57] W. He, X. Fu, N. Li, Q. Ren, and S. Jia, "LGCT: Local–global collaborative transformer for fusion of hyperspectral and multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5537114.

Yuchao Yang (Student Member, IEEE) was born in Xianning, Hubei, China, in 1998. He received the B.E. degree in electronic information engineering from Dalian Maritime University, Dalian, China, in 2020, where he is currently pursuing the Ph.D. degree in information and communication Engineering, 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, Shandong, 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.

She was a joint Ph.D. student at the Remote Sensing Signal and Image Processing Laboratory, University of Maryland, Baltimore County, Baltimore, MD, USA, from 2011 to 2013. From 2011 to 2013, she was a Research Assistant with the Shock, Trauma and Anesthesiology Research organized research center (STAR-ORC), School of Medicine, University of Maryland, College Park, MD, USA. She is currently an Associate Professor and a Doctoral Supervisor in Hyperspectral Imaging in Remote Sensing (CHIRS), Information Science and Technology College, Dalian Maritime University, Dalian, China. Her current research interests include hyperspectral image processing, multisource remote sensing fusion, and vital signs signal processing.

Xin Xu was born in Harbin, Heilongjiang, China, in 2002. She received the Bachelor of Engineering degree in electronic information science and technology from Dalian Maritime University, Dalian, China, in 2025, where she is currently pursuing the Master of Engineering degree in information and communication engineering, information science and technology college.

Her research interests include hyperspectral image, spectral super resolution, and reconstruction.

Enyu Zhao (Member, IEEE) was born in Dalian, Liaoning, 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.

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