MIMFormer: Multiscale Inception Mixer Transformer for Hyperspectral and Multispectral Image Fusion

Rumei Li, Liyan Zhang, Zun Wang, and Xiaojuan Li

***Abstract*—The fusion of low-spatial-resolution hyperspectral image and high-spatial-resolution multispectral image provides an effective method to obtain high-spatial-resolution hyperspectral image. However, existing hybrid fusion architectures combining convolutional neural networks (CNNs) and transformers face significant challenges. Sequential approaches struggle with simultaneous local and global modeling, while parallel approaches often result in information redundancy. In this article, to meet diverse information demands at different layers, we propose a novel multiscale inception mixer transformer network (MIMFormer), a multiscale hybrid network based on the Inception structure integrating CNN and transformer. The core of this network is the multiscale spatial transformer (MST) structure, which enhances the detail richness of fused images by integrating local and global information at various scales. The inception spatial–spectral mixer (ISSM) module within the MST leverages an Inception architecture and employs a spectral splitting mechanism to regulate spectral channel counts across different branches. This design allows the ISSM module to efficiently extract local spatial–spectral features through convolution and max pooling, while global features are captured using a self-attention mechanism, ensuring comprehensive feature fusion across spectral groups. Experimental results on three benchmark datasets and one real remote sensing dataset demonstrate that MIMFormer outperforms ten advanced fusion methods.**

***Index Terms*—Deep learning, hyperspectral image (HSI), image fusion, multispectral image (MSI), transformer.**

I. Introduction

COMPARED to multispectral images, hyperspectral images possess exceptional capabilities for distinguishing key features in illuminated scenes, a fact that has been confirmed by numerous studies [1]. However, due to the physical limitations of imaging sensors, there is an inherent tradeoff between spatial resolution and spectral resolution in imaging, resulting in existing hyperspectral satellites often having lower spatial resolutions, such as EO1-30 m [2] and ZY1E-30 m [3]. Consequently, some hyperspectral applications experience significant performance degradation due to insufficient spatial resolution, including soil composition estimation [4], vegetation classification [5], and urban change detection [6]. Unlike breakthroughs in imaging hardware, fusing low-spatial-resolution hyperspectral image (LR-HSI) with high-spatial-resolution multispectral image (HR-MSI) offers an economically viable method to acquire high-spatial-resolution hyperspectral image (HR-HSI).

Pansharpening methods [7] are used for fusing LR-HSI with the panchromatic band extracted from HR-MSI to generate HR-HSI. Model-based approaches [8] utilize matrix or tensor decomposition of LR-HSI, followed by the recovery of HR-HSI using predefined priors. However, these methods often overly rely on handcrafted prior assumptions about the unknown HR-HSI, leading to spatial distortions and spectral inaccuracies in the fused results [9], [10].

In recent years, deep learning methods, such as convolutional neural networks (CNNs) [11] and transformers [12], [13], have been widely applied to address this challenge. While deep learning methods automatically extract image features, eliminating the limitations of handcrafted features, they each have their own limitations. Transformers emphasize global feature extraction, whereas CNNs excel at local feature extraction. During the fusion process, both global and local features play crucial roles in accurately reconstructing spatial details and preserving spectral fidelity. Global features primarily capture overall semantic and structural context, aiding in maintaining visual consistency and identifying major patterns and trends within the image. Simultaneously, local features focus on small-scale structures, such as edges, lines, and fine textures, which are essential elements in the fusion process.

In order to better combine global and local information, many studies combine CNN and transformer to form a CNN-transformer hybrid architecture [14] to make better use of global and local information, thereby significantly improving the fusion performance. At present, there are two ways to mix this architecture, sequential and parallel, and the combination of sequential results in each layer modeling only one aspect, such as local modeling in the convolutional layer and global modeling in the transformer layer, which is difficult to achieve both in this way. In parallel combination, one branch handles local information

Received 16 May 2024; revised 20 July 2024 and 3 August 2024; accepted 17 August 2024. Date of publication 22 August 2024; date of current version 5 September 2024. This work was supported by the National Natural Science Foundation of China under Grant 42271487. *(Corresponding author: Liyan Zhang.)*

Rumei Li and Zun Wang are with the College of Resource Environment and Tourism, Capital Normal University, Beijing 100048, China (e-mail: 2220902206@cnu.edu.cn; 13824163756@163.com).

Liyan Zhang and Xiaojuan Li are with the Key Laboratory of 3-D Information Acquisition and Application, MOE, and the College of Resource Environment and Tourism, Capital Normal University, Beijing 100048, China (e-mail: zhangliyan@cnu.edu.cn; lixiaojuan@cnu.edu.cn).

Source code is available online at https://github.com/meiruni/MIMFormer.

Digital Object Identifier 10.1109/JSTARS.2024.3447648

and the other branch handles global information. However, this approach can lead to information redundancy if all channels are processed. Study [15] has shown that the lower layers of the transformer require more local information and the higher layers require more global information. Therefore, the simple parallel processing method cannot fully meet the information requirements at all levels, and a more flexible structure must be introduced to optimize the processing and transmission of information. The inception structure [16] is a good solution to the problem of parallel joining, as long as the appropriate channel is divided before entering the branch. At the same time, due to the different requirements for high and low-frequency information in different depths, the number of channels can be divided by control to meet the needs of different depths.

Therefore, we propose a novel network, the multiscale inception mixer transformer (MIMFormer), which integrates CNN and transformer architectures through an inception-based multiscale hybrid approach. Central to MIMFormer is the multiscale spatial transformer (MST) structure, which incorporates an inception spatial–spectral mixer (ISSM). The ISSM regulates the number of spectral channels in various Inception branches via a spectral splitting mechanism, effectively combining CNN and transformer advantages to capture both spectral and spatial information across bands. The main contributions of this article are as follows.

1) We introduce MIMFormer, a multiscale hybrid network based on the inception structure that combines CNN and transformer for fusing LR-HSI and HR-MSI. This architecture capture spectral and spatial information across various bands and scales, enhancing the fused images' quality and accuracy.
2) We develop the ISSM module, constructed upon the inception framework, which ensures image precision through meticulous processing of localized regions and maintains consistency with global sharpening across the entire image. By simultaneously integrating global and local information, it enhances fusion quality while preserving the integrity and authenticity of the image content.
3) We design a spectral splitting mechanism, which regulates the number of spectral channels across different Inception branches. This mechanism reduces feature redundancy and promotes a comprehensive integration of global and local features, thereby further improving the performance of the fusion algorithm.

The rest of this article is organized as follows. Section II reviews related work, including pansharpening, model-based approaches, and deep learning methods. Section III describes the proposed network architectures, MIMFormer, and ISSM. Section IV presents experimental results on benchmark and real datasets. Section V presents the discussion. Finally, Section VI concludes this article.

II. Related Work

In recent years, researchers have developed numerous innovative approaches to the fusion of LR-HSI and HR-MSI from diverse perspectives [17], [18]. These methods can generally be categorized into pansharpening [7], [19], model-based approaches [20], [21], and deep learning methods [22].

A. Pansharpening

Pansharpening is one of the earliest developed methods for fusing LR-HSI and HR-MSI. It involves merging the LR-HSI with the panchromatic band extracted from the HR-MSI, transforming it into HR-HSI [23]. Component substitution (CS) and multiresolution analysis (MRA) are common pansharpening fusion techniques. CS enhances the spatial resolution of hyperspectral images by separating and replacing the spatial components of a multispectral image. Representative methods include principal component analysis [24], [25], intensity hue saturation [26], and Gram–Schmidt spectral sharpening methods [27], [28]. CS-based methods are computationally inexpensive and can recover the main spatial features similar to the original image. However, such approaches often lead to a degradation of spectral information [29]. MRA uses multiresolution decomposition techniques to extract high-frequency spatial details from multispectral images and fuse them into LR-HSI to enhance its spatial resolution. Typical methods based on MRA include high-pass filters [30] and wavelet transforms [31]. While these techniques are computationally straightforward and efficient, significant discrepancies in spatial resolution between multispectral and hyperspectral images may lead to noticeable distortions in the fused results.

B. Model-Based Approaches

Model-based approaches primarily include matrix decomposition, tensor decomposition, and Bayesian-based methods [21], [32]. Yokoya et al. [33] introduced coupled nonnegative matrix factorization (CNMF) for the fusion of LR-HSI and HR-MSI, exploring its impact on HSI classification. Their method demixed the sources of the two images to identify the characteristics and abundances of endmembers. Although this method showed performance improvements, it is computationally intensive and sensitive to parameter selection. Tensor decomposition approaches [34] utilize multidimensional tensors to represent multispectral and hyperspectral images, achieving fusion through tensor decomposition or operations. Tucker decomposition [35], a commonly used method, decomposes high-dimensional tensors into a core tensor and dictionaries for each dimension, extracting and representing information across dimensions. The first known Bayesian fusion approach was designed by Zhang et al. [36]. This method assumes an additive noise imaging model for LR-HSI and uses interpolation as a prior, circumventing the need to estimate the spatial degradation operator and performing super-resolution in a blind manner. Overall, these methods frame the fusion of LR-HSI with HR-MSI as an optimization problem constrained by various handcrafted priors, which may not adequately represent the required HR-HSI, thus limiting their fusion accuracy.

C. Deep Learning Methods

Recent advancements in deep learning have significantly impacted image processing [37], [38], [39], inspiring research in LR-HSI and HR-MSI fusion. Dian et al. [40] introduced a deep hyperspectral image enhancement model that integrates image priors into a CNN fusion framework, outperforming traditional methods. Han et al. [41] developed the MS-SSFNet network, which uses a multilevel loss function to mitigate gradient vanishing in fusing LR-HSI and HR-MSI. Zhang et al. [22] employed CNNs to regularize spatial and spectral degradation and used generative networks to model HR-HSI. Li et al. [42] proposed the cross spectral-scale and shift-window-based cross spatial-scale nonlocal attention network (CSSNet) to explicitly learn spectral and spatial correlations between two input images. To further enhance CNN-based fusion algorithms, researchers introduced a multitask, multiobjective evolutionary network [43], [44] to address spectral distortion caused by LR-HSI upsampling. These CNN-based methods have significantly advanced fusion algorithms, providing robust solutions to the limitations of prior methods and achieving satisfactory results.

Although CNN-based methods significantly improve over traditional approaches, their limited receptive fields and lack of remote modeling ability prevent the full extraction of global image features, reducing fusion quality [45]. To address this, vision transformer (ViT), which excels in modeling global dependencies, has recently been applied to LR-HSI and HR-MSI fusion with notable success [46]. For instance, Hu et al. [47] introduced Fusformer, the first ViT-based solution for fusion, while Jia et al. [48] developed the multiscale spatial–spectral transformer network (MSST-Net) to enhance network performance and generalization. Fang et al. [49] integrated spatiotemporal frequency information of LR-HSI and HR-MSI. However, focusing too much on global information can neglect local feature extraction.

Combining CNN's local feature extraction with transformer's global feature modeling [50], [51] aims to address these issues and has shown promising results. This hybrid approach leverages the strengths of both architectures, allowing for a more comprehensive extraction of image features. However, challenges remain. Ma et al. [52] used Swin transformer's window attention with 3D-CNN to learn LR-HSI's implicit priors. While this outperforms pure transformer networks, the sequential mode struggles to balance global and local information extraction. Interactformer [53] combines Swin transformer and 3D-CNN in parallel to improve spatial resolution and preserve spectral information, but this can lead to feature redundancy and weakened performance. Lower transformer layers need more local information, while higher layers require global information [15]. Therefore, a simple parallel method cannot fully satisfy the information needs at all levels.

Inspired by the Inception structure [16], we propose a novel multiscale hybrid network, MIMFormer, which optimizes information processing through a spectral splitting mechanism. By adjusting the number of channels, MIMFormer meets different depth requirements, providing a flexible structure to enhance fusion. This innovative architecture effectively balances the extraction of global and local features, addressing the limitations of previous methods and achieving superior fusion quality.

III. Methodology

A. MIMFormer Architecture

The proposed MIMFormer fusion network architecture, depicted in Fig. 1, comprises three primary modules: a shallow feature extraction module, a deep feature extraction module, and an image reconstruction module. The shallow feature extraction module utilizes two residual network modules to extract preliminary features. The deep feature extraction module consists of three MST branches, each functioning at a distinct scale. Lastly, the image reconstruction module includes two convolutional layers paired with a LeakyReLU activation function to reconstruct the final image.

Let $Z \in \mathbb{R}^{h\times w\times S}$ represent the observed LR-HSI, where h, w, and S denote the number of rows, columns, and spectral bands of the LR-HSI, respectively. Let $Y \in \mathbb{R}^{H\times W\times s}$ represent the observed HR-MSI, where H, W, and s denote the number of rows, columns, and spectral bands of the HR-MSI, respectively. Our objective is to fuse Y and Z to obtain an image $\hat{X} \in \mathbb{R}^{H\times W\times S}$ that possesses both high spatial and spectral resolutions. Initially, this article employs a bilinear interpolation method to upsample LR-HSI to obtain $Z_{\text{up}} \in \mathbb{R}^{H\times W\times S}$, facilitating channel-wise concatenation with HR-MSI. This process can be represented by the following equation:

$$Z_{\text{up}} = \text{Up}(Z) \tag{1}$$

where $\text{Up}(\cdot)$ denotes the bilinear interpolation upsampling function. Subsequently, Y and Z_{up} are aggregated along the channel dimension to form $D_{\text{cat}} \in \mathbb{R}^{H\times W\times(S+s)}$, which can be expressed as

$$D_{\text{cat}} = \text{Concat}\,(Y, Z_{\text{up}}) \tag{2}$$

where $\text{Concat}(\cdot)$ represents concatenation along the channel dimension. Given that convolution is a simple and effective method for mapping images to a higher dimensional feature space, this article uses 2-D convolution with a kernel size of 3, 180 channels, and a stride of 1, constructing a residual network with two blocks to extract shallow features $F_s \in \mathbb{R}^{H\times W\times 180}$, represented as

$$F_s = \text{SF}\,(D_{\text{cat}}) \tag{3}$$

where $\text{SF}(\cdot)$ denotes the shallow feature extraction module.

In the deep feature extraction module, we designed three MST to extract features at different scales. For the simulated dataset, the patch size values in MST were set to 8, 12, and 16, whereas for the real dataset, the patch sizes were set to 3, 15, and 6. Each MST module includes a patch embed layer, three ISSMs, and a patch unembed layer. Initially, the extracted shallow features F_s are fed into the patch embed layer, whose goal is to divide the shallow feature map into a series of equally sized image blocks, each mapped to a higher dimensional feature representation. This process is described by the following equation:

$$I^i = \text{GELU}\,(\text{Norm}\,(\text{Conv}\,(F_s)))\,. \tag{4}$$

The feature maps, represented as $I^i \in \mathbb{R}^{H/P\times W/P\times C}$ and $i = 1, 2, 3$, are processed through a convolutional layer after passing through the patch embed layer. The convolutional layer uses a

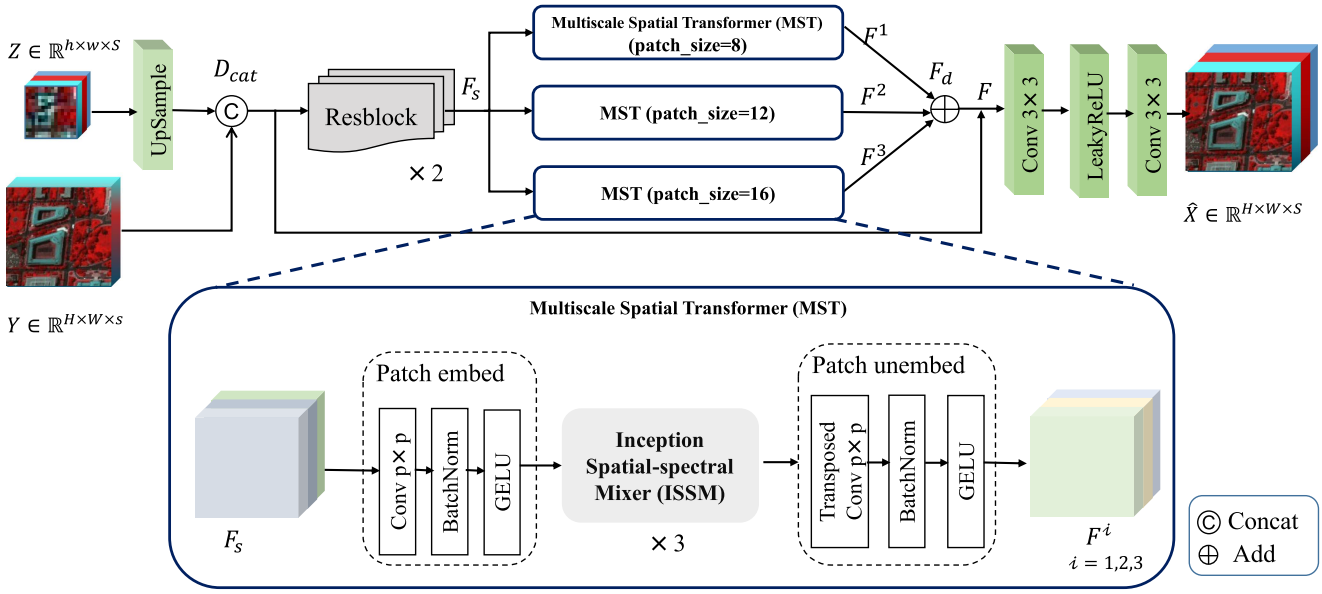


Fig. 1. Overall architecture diagram of the proposed MIMFormer network.

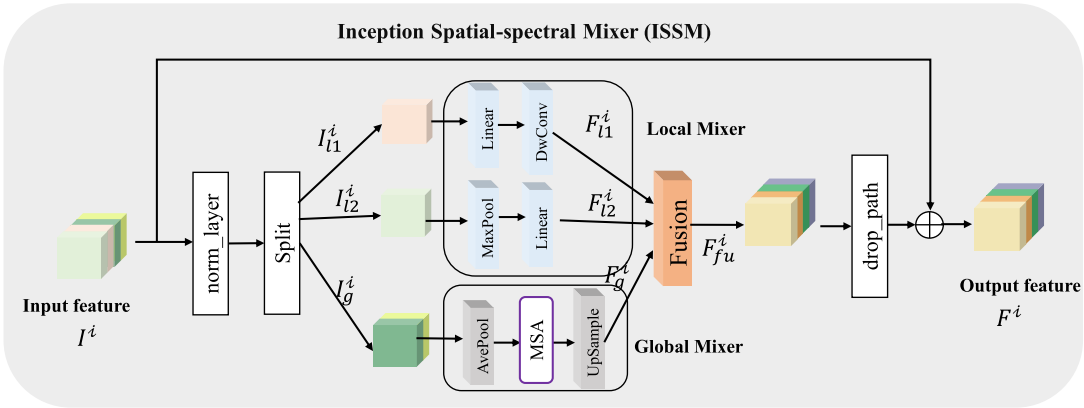


Fig. 2. Architecture diagram of the proposed ISSM.

kernel size and stride of p (with p values being 8, 12, 16, or 3, 15, 6), a channel count of $C = 120$, and a padding of 1.

B. ISSM and Spectral Splitting Mechanism

The architecture of the ISSM is shown in Fig. 2. After partitioning the input features across the spectral dimensions, local and global mixers are employed to learn features across different frequency ranges. The local mixer includes a MaxPool path, consisting of a maximum pooling operation and a linear layer, and a parallel convolutional path, consisting of a linear layer and a DwConv layer. The global mixer includes an attention path, which consists of average pooling, a multihead self-attention mechanism (MSA), and an upsample layer. The rationale is as follows.

Local mixer: Given the sharpness sensitivity of maximum pooling and the detail perception capability of convolution operations, we propose two local paths to leverage the sharpness sensitivity of maximum pooling and the detail perception capability of convolution layers to learn local features. Initially, the input I^i_l (where $i = 1, 2, 3$) is divided into $I^i_{l1} \in \mathbb{R}^{H/P\times W/P\times C_l/2}$ and $I^i_{l2} \in \mathbb{R}^{H/P\times W/P\times C_l/2}$. In our experiments, to reduce feature redundancy, enhance feature extraction efficiency, and ensure that both pathways receive spectral information representing global characteristics while retaining sufficient detail, we designed a spectral splitting mechanism. This mechanism sets C_l as $C - (i \times \text{dim})$, dim as 40. I^i_{l1} is routed to the MaxPool path, and I^i_{l2} is routed to the parallel convolution path. The outputs of the local mixer, F^i_{l1} and F^i_{l2}, can be represented by the following equation:

$$F^i_{l1} = \text{Linear}\left(\text{MaxPool}\left(I^i_{l1}\right)\right) \tag{5}$$

$$F^i_{l2} = \text{DwConv}\left(\text{Linear}\left(I^i_{l2}\right)\right). \tag{6}$$

Global mixer: Considering the powerful capability of attention mechanisms in learning global representations, we use an MSA to establish long-distance dependencies to learn global information. Before applying the attention operation, we use an average pooling operation to reduce the scale of $I^i_g \in$

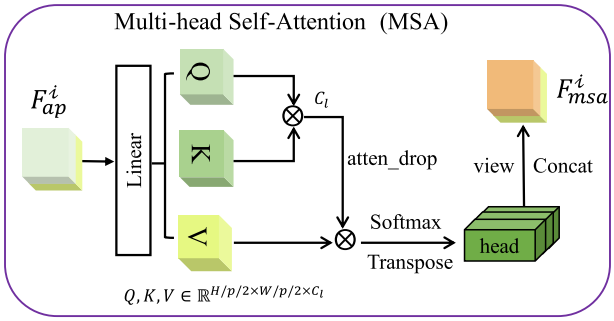


Fig. 3. Structure MSA.

$\mathbb{R}^{H/P\times W/P\times C_g}$ to decrease computational complexity. The kernel size and stride for the average pooling in this experiment are both 2. Then, a multihead self-attention mechanism is applied to calculate the attention for redundant bands

$$F_{\text{ap}}^i = \text{AvgPool}\left(I_g^i\right) \tag{7}$$

where $F_{\text{ap}}^i \in \mathbb{R}^{H/P/2\times W/P/2\times C_g}$ is the output of average pooling, C_g is set as $i \times \text{dim}$.

Multihead self-attention, as shown in Fig. 3, captures the correlation among spectral bands by computing self-attention. Initially, we project the input F_{ap}^i through trainable linear projections to obtain the query matrix $Q \in \mathbb{R}^{H/P/2\times W/P/2\times C_g}$, key matrix $K \in \mathbb{R}^{H/P/2\times W/P/2\times C_g}$, and value matrix $V \in \mathbb{R}^{H/P/2\times W/P/2\times C_g}$, represented as

$$Q^i = F_{\text{ap}}^i W_Q^i, \quad K^i = F_{\text{ap}}^i W_K^i, \quad V^i = F_{\text{ap}}^i W_V^i \tag{8}$$

where $W_Q^i, W_K^i, W_V^i \in \mathbb{R}^{C_g\times C_g}$ are learnable projection matrices. The scaled dot-product attention function, using the queries, keys, and values, is defined as

$$\text{Atten}(Q^i, K^i, V^i) = V^i\left(\text{softmax}\left(\frac{K^{i^T}Q^i}{\sqrt{C_g}}\right)\right). \tag{9}$$

Atten$(\cdot)$ is the scaled dot-product attention function. The MSA, similar to that used in ViT, enhances the network's feature extraction capability. The function form of the multihead self-attention is as follows:

$$\text{head}_h^i = \text{Atten}\left(Q_h^i, K_h^i, V_h^i\right), \quad h = 1, 2, 3 \tag{10}$$

$$F_{\text{msa}}^i = \text{view}\left(\text{Concat}\left(\text{head}_h^i\right)\right) \tag{11}$$

where head_h is the output of the hth head, computed through the Atten function. Concat denotes concatenation of all head outputs. Finally, the spectral multihead self-attention output is reshaped.

C. Future Fusion

Ultimately, we employ an upsampling layer to restore the original scale, consistent with a ratio of 2 used in max pooling

$$F_g^i = \text{UpSample}\left(F_{\text{msa}}^i\right). \tag{12}$$

The outputs F_{l1}^i, F_{l2}^i, and F_g^i are concatenated along the channel dimension and fused to form F_{fu}^i

$$F_{fu}^i = \text{Fusion}\left(\text{Concat}\left(F_{l1}^i, F_{l2}^i, F_g^i\right)\right). \tag{13}$$

To ensure the feature dimensions are consistent before reconstructing the image, the fused feature map is passed through a patch unembed layer and a transposed convolution layer to restore the features to the original number of hyperspectral bands S, resulting in a single-scale feature $F^i \in \mathbb{R}^{H\times W\times S}$. Subsequently, the extracted multiscale features are aggregated using learnable weights to form the deep features $F_d \in \mathbb{R}^{H\times W\times S}$. Finally, F_d and the shallow features F_s are input into an image reconstruction module composed of two convolutional layers and a LeakyReLU activation function, yielding the estimated high-resolution hyperspectral image $\hat{X} \in \mathbb{R}^{H\times W\times S}$

$$\hat{X} = \text{Conv}\left(\text{LReLU}\left(\text{Conv}(F)\right)\right). \tag{14}$$

Ultimately, the network's parameters are optimized by minimizing the $L1$ pixel loss

$$l_1 = \|\hat{X} - X\| \tag{15}$$

where $X \in \mathbb{R}^{H\times W\times S}$ is the true HR-HSI.

IV. Experiments and Analyzes

To effectively evaluate the performance of the proposed methods, this study selected ten state-of-the-art fusion technologies for comparative analysis. These techniques include two traditional methods: HySure [54] and CNMF [33]; four CNN-based approaches: MSDCNN [55], TFNet [56], MHF-Net [57], and CSSNet [42]; along with four novel transformer-based technologies: Fusformer [47] PSRT [58], MSST-Net [59], and 3DT-Net [52]. The parameters for each method were set according to the original authors' code or literature recommendations. Traditional methods were tested on a Windows 10 system equipped with an Intel Core i9 processor and 32GB RAM, using MATLAB R2014a. Deep learning methods were primarily implemented using Python 3.8 and PyTorch 1.7, with GPU acceleration provided by NVIDIA RTX 4060TI. Data preprocessing and analysis were conducted using MATLAB R2014a and Python's NumPy and Pandas libraries.

To comprehensively evaluate the performance of image fusion algorithms, a variety of quantitative metrics are commonly employed for comparative analysis [10], [60]. These metrics include the spectral angle mapper (SAM), peak signal-to-noise ratio (PSNR), erreur relative globale adimensionnelle de synthèse (ERGAS), structural similarity index metric (SSIM), root mean squared error (RMSE), and the quality with no reference (QNR) index. SAM assesses spectral quality, with lower values indicating minimal loss of spectral information. PSNR evaluates spatial effects, where higher values denote lesser loss of spatial details. SSIM is used to appraise structural correlation, with higher values suggesting superior fusion outcomes. RMSE measures the similarity between images, where lower values denote a more effective fusion algorithm. ERGAS serves as a comprehensive metric, with lower values indicating higher fusion quality. Particularly, QNR is apt for evaluating the fusion

quality of no-reference imagery, such as the ZY1E real remote sensing dataset, encompassing all aspects of distortion including both spectral and spatial distortions. Higher QNR values signify optimal fusion quality with more complete information preservation. Collectively, these metrics reflect the efficacy of fusion algorithms in retaining both spatial and spectral information.

A. Datasets

The experiments in this article utilized three mainstream hyperspectral image benchmark datasets: CAVE [61], Washington DC Mall (WDCM) [62], Pavia University (PU) [63], and a real remote sensing dataset ZY1E.

The CAVE dataset contains 32 indoor hyperspectral images, each with dimensions of 512×512 pixels, covering the wavelength range of 400–700 nm with 31 bands. Experiments followed the Wald protocol [64], using the spectral response function of a Nikon D700 camera to generate HR-MSI. The original hyperspectral images of CAVE served as reference HR-HSI. They were filtered with a Gaussian kernel of size 8×8 and standard deviation of 2, then subsampled by a factor of eight in both horizontal and vertical directions to generate LR-HSI. A total of 20 image pairs were randomly selected from the dataset for training, and the remaining 12 pairs for testing. During training, patches of 64×64 were randomly extracted from each 512×512 image, making the dimensions of HR-HSI, HR-MSI, and LR-HSI during training $64 \times 64 \times 31$, $64 \times 64 \times 3$, and $8 \times 8 \times 31$, respectively, and during testing, $512 \times 512 \times 31$, $512 \times 512 \times 3$, and $64 \times 64 \times 31$.

The WDCM dataset, captured by the Hydice sensor in 1995, consists of 191 bands covering the wavelength range of 400–2400 nm. Each band has a resolution of 1280×307 pixels, with a spatial resolution of 2.5 m. Two subimages of 128×128 pixels were cropped for testing, with the remainder used for training. The setup was the same as that used for MSST-Net [59], with HR-MSI generated using the Sentinel-2A spectral response matrix and LR-HSI produced in the same manner as the CAVE dataset. The dimensions of HR-HSI, HR-MSI, and LR-HSI during training were $64 \times 64 \times 191$, $64 \times 64 \times 10$, and $8 \times 8 \times 191$, respectively, and for testing, $128 \times 128 \times 191$, $128 \times 128 \times 10$, and $16 \times 16 \times 191$.

The PU dataset was collected by the ROSIS sensor in 2003, originally measuring 610×340 pixels in dimensions, covering a wavelength range of 430–860 nm. After removing 22 water vapor absorption bands, 93 bands remained. Consistent with the WDCM dataset, after Gaussian filtering, the images were subsampled by a factor of eight to generate LR-HSI. Two subimages of 128×128 pixels were cropped for testing, with the remainder used for training. HR-MSI was generated using a spectral response function similar to IKONOS. The dimensions of HR-HSI, HR-MSI, and LR-HSI during training were $64 \times 64 \times 93$, $64 \times 64 \times 4$, and $8 \times 8 \times 93$, respectively, and for testing, $128 \times 128 \times 93$, $128 \times 128 \times 4$, and $16 \times 16 \times 93$.

The ZY1E dataset consists of hyperspectral and multispectral data acquired from the ZY1E satellite, specifically from the ZY1E satellite, equipped with both visible-near infrared and hyperspectral cameras. This study utilized an image captured on 19 April 2023, over the Pinggu district of Beijing, consisting of AHSI hyperspectral and VNIC multispectral imagery. In the ENVI software, a series of preprocessing steps were applied to the acquired hyperspectral and multispectral data, including radiometric calibration, atmospheric correction, orthorectification, and cropping, followed by registration of the hyperspectral and multispectral images. The processed data comprised 166 bands of LR-HSI and 8 bands of HR-MSI, with spatial resolutions of 30 m and 10 m, and spatial dimensions of 4986×4581 and 1662×1527 pixels, respectively. Notably, lacking HR-HSI as a reference image for training, we followed the Wald protocol [64], using a 9×9 Gaussian kernel to perform a threefold spatial downsampling of the original LR-HSI and HR-MSI to generate the training dataset. Subsequently, the original LR-HSI was considered as HR-HSI. Given the low signal-to-noise ratio in some bands of the ZY-1E data, we selected the first 76 bands of the LR-HSI for the fusion experiments. During the training phase, image pairs were randomly cropped from the training dataset (HR-HSI: $60 \times 60 \times 76$, HR-MSI: $60 \times 60 \times 8$, LR-HSI: $20 \times 20 \times 76$) for training purposes. In the testing phase, image pairs (HR-MSI: $540 \times 540 \times 8$, LR-HSI: $180 \times 180 \times 76$) were cropped for testing.

TABLE I
ABLATION STUDY OF THE ISSM AND MULTISCALE STRUCTURE ON THE WDCM DATASET

Model	M-scale	NAttention	NMaxPool	NDwConv	BASE
M-scale	×	✓	✓	✓	✓
NAttention	✓	×	✓	✓	✓
NMaxPool	✓	✓	×	✓	✓
NDwConv	✓	✓	✓	×	✓
BASE	✓	✓	✓	✓	✓
PSNR↑	46.83	45.42	46.41	48.63	**52.20**
SAM↓	1.45	1.67	1.49	1.19	**0.78**
ERGAS↓	1.10	1.28	1.14	0.80	**0.53**
SSIM↑	0.9947	0.9935	0.9945	0.9934	**0.9983**
RMSE↓	0.0046	0.0054	0.0047	0.0021	**0.0014**

The best indicators are displayed in bold.

B. Ablation Experiments

To better understand MIMFormer, a series of ablation experiments were conducted. All models were trained on the WDCM dataset for 100 epochs, with training configurations consistent with those previously described in the document.

In terms of multiscale feature extraction: To adequately extract features from LR-HSI and HR-MSI, this study employed a multiscale approach for feature extraction. To evaluate the effectiveness of multiscale features, we conducted single-scale feature extraction experiments by modifying the original three-branch feature structure into a single-branch structure. "M-scale" denotes the removal of the multiscale architecture, utilizing only a single MST structure. The quantitative evaluation metrics are presented in the first column of Table I. In addition, we plotted the decline in the loss function and the increase in PSNR values on the test set with the single MST structure as training epochs progressed, as shown by the orange lines in Fig. 4. The results demonstrate that multiscale feature extraction effectively captures features at various scales, enabling the model to identify patterns and details that are challenging

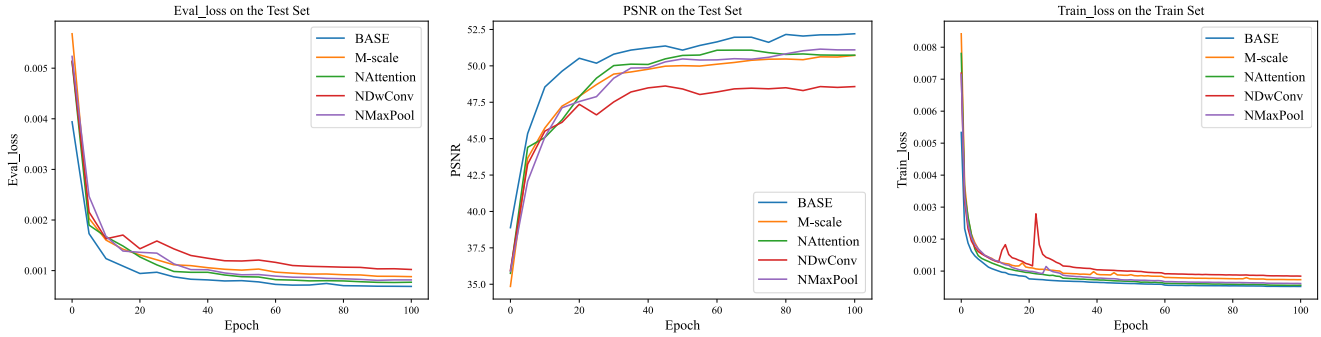


Fig. 4. Ablation results on the WDCM dataset, showing evaluation loss on the test set, PSNR on the test set, and training loss on the train set, for different models (BASE, M-scale, NAttention, NDwConv, NMaxPool).

TABLE II
ABLATION STUDY OF SPECTRAL SPLITTING MECHANISM ON THE WDCM DATASET

Model	PSNR↑	SAM↓	ERGAS↓	SSIM↑	RMSE↓
NSSM	51.40	0.86	0.58	**0.9983**	0.0015
AC	50.07	1.00	0.67	0.9967	0.0017
BASE	**52.20**	**0.78**	**0.53**	**0.9983**	**0.0014**

The best indicators are displayed in bold.

TABLE III
COMPARISON OF DIFFERENT FUSION METHODS ON THE CAVE DATASET

Method	PSNR↑	SAM↓	ERGAS↓	SSIM↑	RMSE↓
CNMF [33]	37.95	3.35	3.16	0.9762	0.0157
HySure [54]	40.39	2.61	2.62	0.9832	0.0096
MSDCNN [55]	39.84	3.04	2.90	0.9775	0.0100
TFNet [56]	44.29	1.90	1.45	0.9912	0.0062
MHF-Net [57]	45.93	1.58	1.21	0.9934	0.0050
CSSNet [42]	46.20	1.53	1.16	0.9943	0.0049
Fusformer [58]	46.74	1.45	1.11	0.9945	0.0046
PSRT [59]	45.51	1.65	1.26	0.9936	0.0053
MSST-Net [60]	47.16	1.39	1.06	0.9950	0.0044
3DT-Net [61]	46.84	1.44	1.10	0.9950	0.0045
MIMFormer	**48.36**	**1.25**	**0.95**	**0.9960**	**0.0039**

The best indicators are displayed in bold.

to detect at a single scale, thereby achieving superior fusion performance.

Regarding the ISSM: To integrate the local feature extraction capabilities of CNN with the strengths of transformer, we introduces the ISSM, aimed at enhancing the transformer's perceptual ability in the spectral dimension. To evaluate the effectiveness of each component within the inception mixer, we progressively removed each branch from the complete model and recorded the results. As shown in Table I, combining the attention mechanism with convolution and max-pooling yields higher accuracy compared to using only the mixer. This validates the efficacy of the ISSM. As illustrated in Fig. 4, the blue line labeled "BASE" represents our proposed MIMFormer architecture; the green, red, and purple lines correspond to the removal of the attention branch, the convolution branch, and the max-pooling branch from the ISM's three-branch structure, respectively. The results indicate that MIMFormer exhibits the fastest and most stable loss function decline, both in the test and validation sets, and achieves the most significant improvement in PSNR values on the test set.

Regarding the spectral splitting mechanism: This mechanism can reduce feature redundancy and improve feature extraction efficiency, but improper allocation may lead to insufficient extraction of both global and local information. Therefore, the key is to balance the splitting mechanism to ensure comprehensive information extraction through the full integration of global and local features. In designing the spectral splitting mechanism, we ensure that both paths receive spectral information that represents global features while containing sufficient detail. By setting $C_l = C - (i \cdot \text{dim})$ and gradually increasing the value of i, we can dynamically adjust the information received by each path. As shown in Table II, ac represents an average allocation of channels (both global and local paths receiving $C/2$ channels), BASE represents our MIMFormer method using the spectral splitting mechanism, and NSSM represents the absence of the spectral splitting mechanism (both global and local paths receiving C channels). BASE outperforms in all metrics. This design ensures that the spectral splitting mechanism effectively enhances network performance by fully leveraging and extracting both global and local information.

C. Experiments With Benchmark Data

Results on CAVE dataset: We evaluated 12 test images on the CAVE dataset and presented the average evaluation metrics for different methods in Table III, with the best results highlighted in bold. It is observable that the proposed MIMFormer method significantly outperforms the comparison methods in terms of performance. Notably, the PSNR values for MIMFormer are substantially higher than those of other methods. This aligns with previous analyses of the network architecture, suggesting that the proposed ISSM effectively preserves the spectral characteristics of the scene.

To ease the burden on readers, we only display the fusion results for the "watercolors" test image. Fig. 5 showcases the enlarged local images using bicubic interpolation, and the synthesized false-color images of bands 29, 19, and 9. The second and fourth rows display the error images, where the error values are the average of all band errors. Compared to the methods assessed, the proposed MIMFormer method reconstructs high-resolution details more effectively, significantly reducing errors in the error images, especially in regions with prominent edge information.

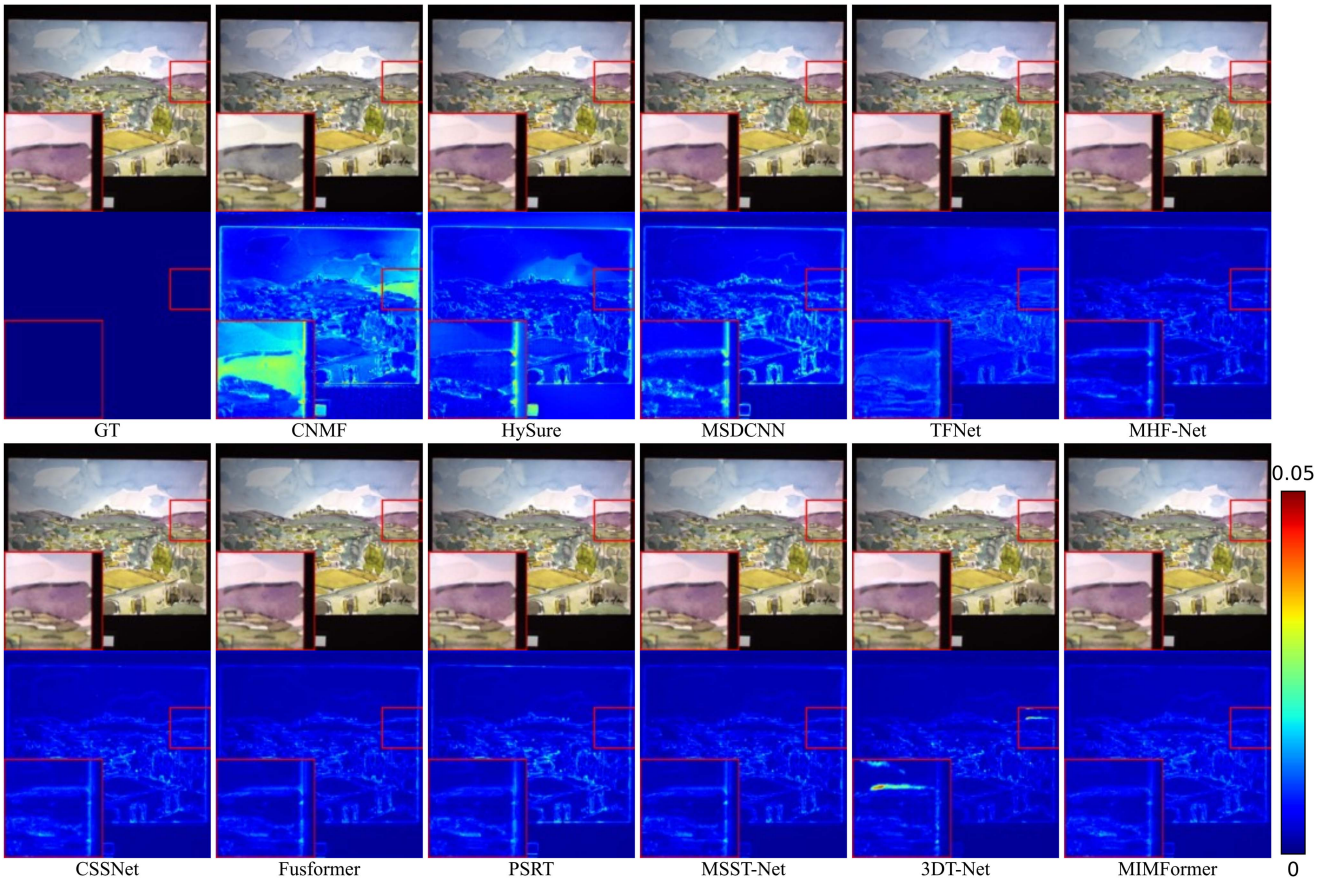


Fig. 5. Fusion results of the "watercolors" image from the CAVE dataset. The first row presents a false-color image synthesized from the 29th, 19th, and 9th spectral bands. The second row depicts the error images between the fused and the ground truth.

TABLE IV
COMPARISON OF DIFFERENT FUSION METHODS ON THE WDCM AND PU DATASETS

Method	WDCM Dataset					PU Dataset						
	PSNR↑	SAM↓	ERGAS↓	SSIM↑	RMSE↓	PSNR↑	SAM↓	ERGAS↓	SSIM↑	RMSE↓	FLOPs (G)	Para (M)
CNMF [33]	37.28	4.10	2.93	0.9938	0.0077	34.73	5.07	2.91	0.9716	0.0309	/	/
HySure [54]	35.65	4.74	3.52	0.9915	0.0094	36.05	4.16	2.30	0.9747	0.0256	/	/
MSDCNN [55]	45.90	1.28	0.86	0.9986	0.0022	42.07	2.24	1.24	0.9885	0.0125	2.209	0.527
TFNet [56]	47.67	1.32	0.88	0.9983	0.0023	42.56	2.12	1.08	0.9904	0.0119	2.020	2.387
MHF-Net [57]	48.22	1.24	0.83	0.9985	0.0021	41.19	2.48	1.23	0.9841	0.0139	22.460	3.630
CSSNet [42]	48.82	1.16	0.78	0.9941	0.0020	42.11	2.21	1.14	0.9892	0.0125	2.507	1.226
Fusformer [58]	49.24	1.10	0.74	0.9989	0.0019	42.22	2.20	1.12	0.9897	0.0124	456.327	0.109
PSRT [59]	51.31	0.87	0.58	0.9992	0.0015	41.62	2.37	1.21	0.9866	0.0133	1.291	0.302
MSST-Net [60]	49.98	1.01	0.67	0.9990	0.0018	42.61	2.11	1.08	0.9910	0.0118	188.720	34.400
3DT-Net [61]	52.10	0.79	**0.53**	**0.9993**	**0.0014**	42.83	2.05	1.05	0.9916	0.0116	66.122	3.455
MIMFormer	**52.20**	**0.78**	**0.53**	0.9983	**0.0014**	**43.09**	**2.00**	**1.02**	**0.9919**	**0.0112**	35.537	16.206

The best indicators are displayed in bold.

Results on the WDCM dataset: Table IV presents the objective results of various comparative algorithms on the WDCM dataset, with the best metrics highlighted in bold. Deep learning methods based on CNNs, such as MSDCNN and TFNet, exhibit superior performance compared to traditional approaches, with PSNR values reaching 45.90 and 47.67, respectively. MHF-Net and CSSNet excel further in detail preservation, achieving PSNR values of 48.22 and 48.82. Transformer-based methods also demonstrate commendable efficacy; for instance, MSST-Net and 3DT-Net showcase exceptional integration capabilities. Notably, 3DT-Net attains an SSIM of 0.9993 on the WDCM dataset, with its metrics being on par with our proposed MIMFormer, attributable to its utilization of 3-D CNNs to accommodate the characteristics of hyperspectral data cubes. However, it is worth noting that our proposed MIMFormer shows slightly lower SSIM performance on the WDCM dataset compared to other deep learning networks. This may be due to the inherent challenges MIMFormer faces when handling extremely fine details. Nevertheless, this 0.001 discrepancy is negligible in terms of overall performance impact. Overall, the proposed MIMFormer outperforms other methods across most critical performance metrics.

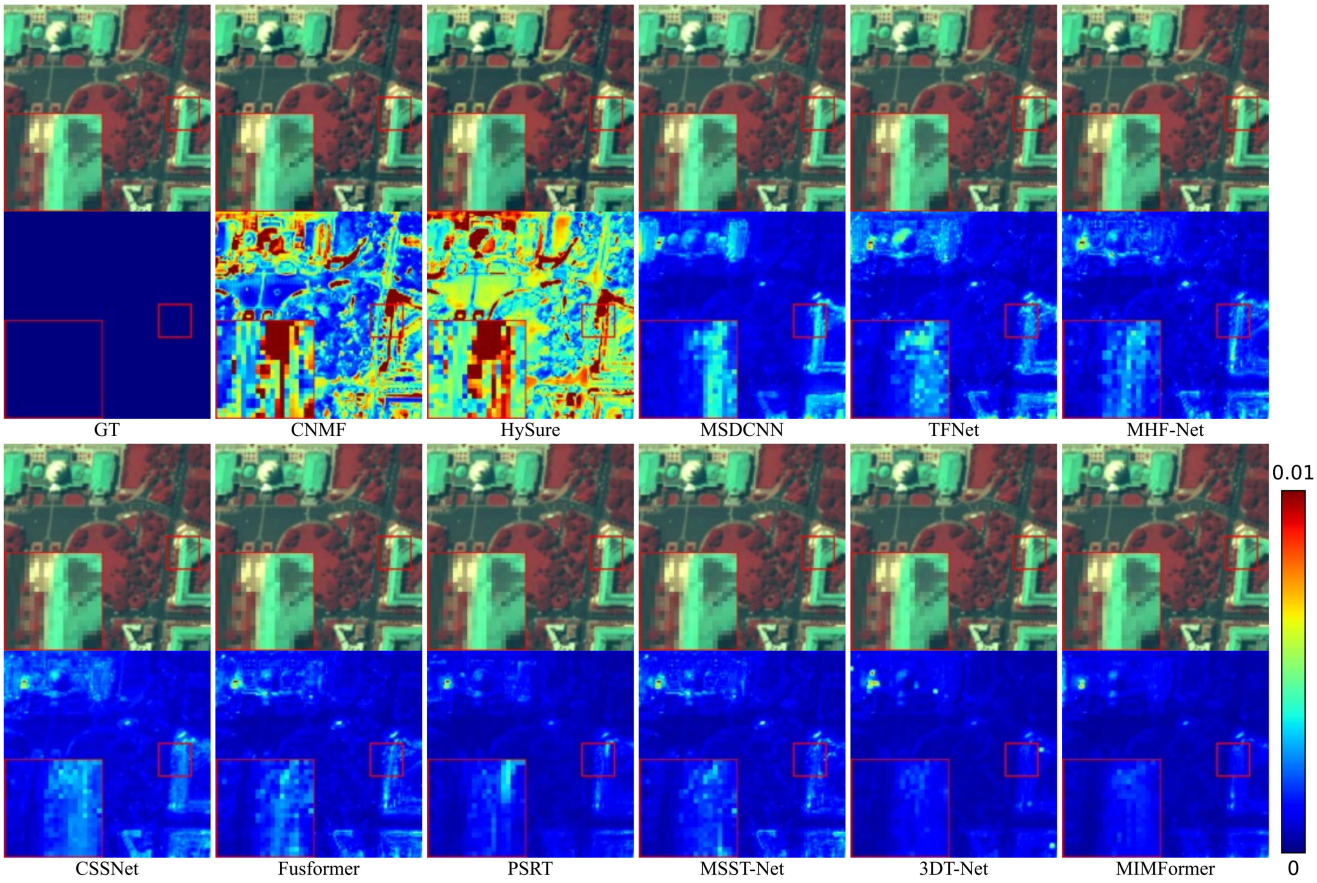


Fig. 6. Fusion results of WDCM test set. The first row presents a false-color image synthesized from the 56th, 21th, and 5th spectral bands. The second row depicts the error comparison between the fused and the ground truth.

This confirms that employing multiscale feature extraction combined with the ISSM structure can significantly improve the performance of hyperspectral and multispectral image fusion networks. Compared to methods that use CNNs alone for fusion, increasing network depth limits the receptive field, leading to the loss of many detail features. In our method, by incorporating the Inception structure and using a three-branch structure to extract high-frequency and low-frequency information, along with the long-distance dependencies of the self-attention mechanism and local feature extraction of deep convolution and max pooling layers, our network can learn more valuable information, effectively addressing this issue. Thus, MIMFormer surpasses other comparative methods in performance. Fig. 6 provides a visual appreciation of the superiority of our network's results over other comparative methods in terms of color and edge granularity. Compared to CNN-based methods, our network achieves more ideal results in color and brightness performance, and shows more notable improvements in edge detail and clarity compared to 3DT-Net and MSST-Net.

Results on the PU dataset: The ground sampling distance of the PU dataset is 1.3 m, with each pixel containing only one or a few types of land cover, making the spectral characteristics relatively simple. The quantitative results of MIMFormer and other comparative methods on the PU dataset are shown in Table IV. It is evident from the table that traditional methods, such as CNMF, still perform poorly on this dataset. In contrast, MIMFormer and MSST-Net outperform Fusformer in key performance metrics such as PSNR, SAM, ERGAS, SSIM, and RMSE, thanks to their multiscale architectures that enable the extraction of deep spectral features at different scales.

Regarding the fusion results of the PU dataset test set, the false-color images and error maps are shown in Fig. 7. MIMFormer, combining CNN and transformer technologies, achieved satisfactory visual results on this dataset. As shown in Fig. 7, the fused image mainly includes land cover types, such as asphalt roads, grasslands, trees, self-adhesive bricks, and buildings. In the error map, the areas marked by red rectangles are magnified to show artificial buildings and grasslands. The density of grasslands varies in different locations, and the materials used in different buildings also vary, leading to spectral signal differences even among the same type of land cover, greatly increasing the difficulty of fusing hyperspectral and multispectral images. Despite these challenges, MIMFormer still demonstrates superior image reconstruction quality in the magnified areas compared to MSST-Net and 3DT-Net, showcasing its exceptional fusion capabilities. To analyze the computational burden, the last two columns of Table IV list the floating-point operations (FLOPs) and the parameters of different fusion methods on the PU dataset. As shown, MSST-Net and Fusformer have higher FLOPs, at 188.72G and 456.327G, respectively. Fusformer causes GPU memory overload due to its use of original transformer layers. MSST-Net likely uses large-scale

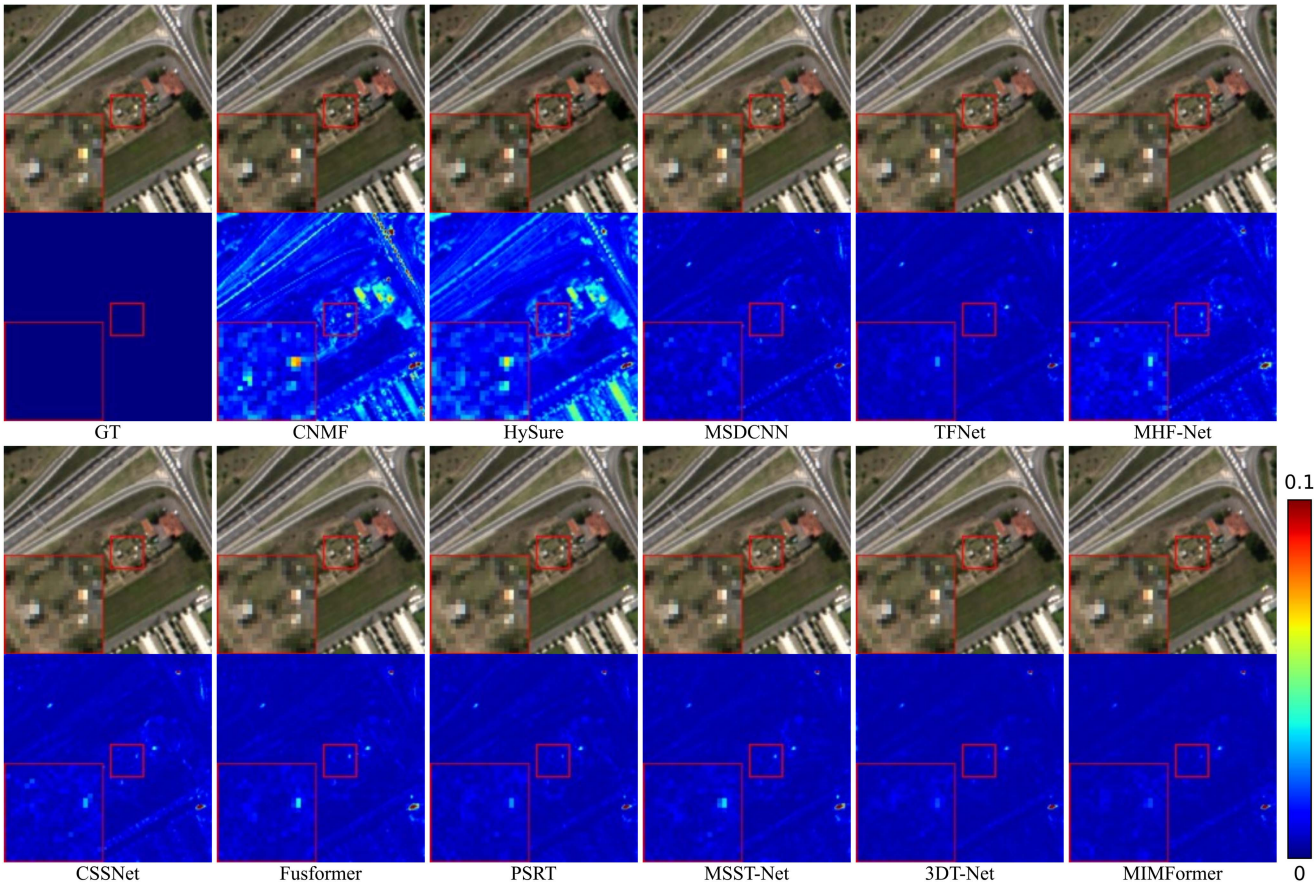


Fig. 7. Fusion results of PU test set. The first row presents a false-color image synthesized from the 29th, 19th, and 9th spectral bands. The second row depicts the error comparison between the fused and the ground truth.

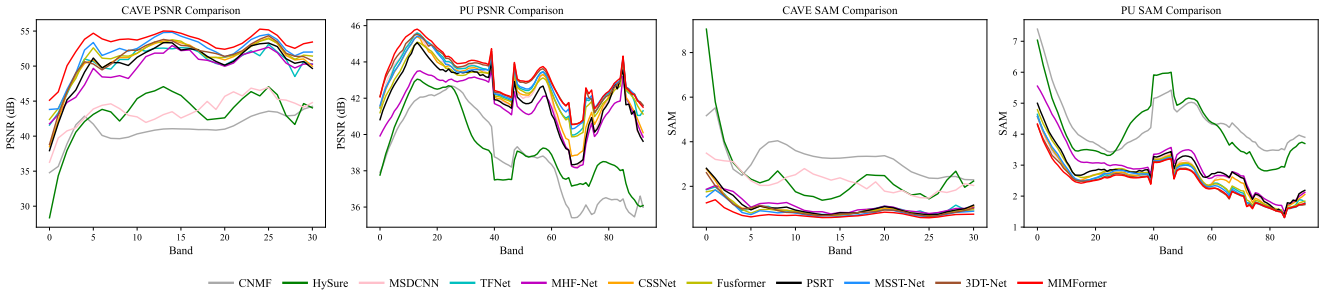


Fig. 8. PSNR and SAM on CAVE and PU datasets of all bands.

transposed convolutions, resulting in a high parameter count. Other methods like CSSNet, PSRT, and 3DT-Net have more moderate FLOPs and parameter counts. For instance, 3DT-Net has FLOPs and parameter counts of 66.122G and 3.455M, respectively. Compared to MIMFormer, 3DT-Net has higher FLOPs but fewer parameters, indicating different optimization strategies in computational burden and parameter usage. Overall, MIMFormer strikes a balance between high performance and reasonable computational burden and parameter count, demonstrating efficiency and resource utilization in practical applications.

In addition, to further evaluate the performance of each method across individual spectral bands of images, we plotted the PSNR and SAM values for different methods across each spectral band on the benchmark datasets CAVE and PU. As shown in Fig. 8, from a spatial perspective, our proposed MIMFormer displays the highest PSNR values in certain bands, indicating optimal performance in terms of spatial information loss; in terms of spectral quality, MIMFormer achieves the lowest SAM values across all bands, indicating minimal spectral information loss. These results demonstrate that our proposed fusion method can generate higher quality fused images.

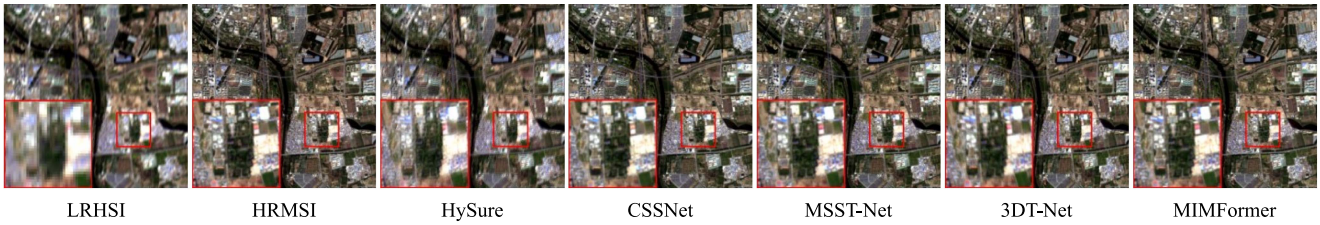


Fig. 9. Fusion results of various methods applied to the ZY1E dataset.

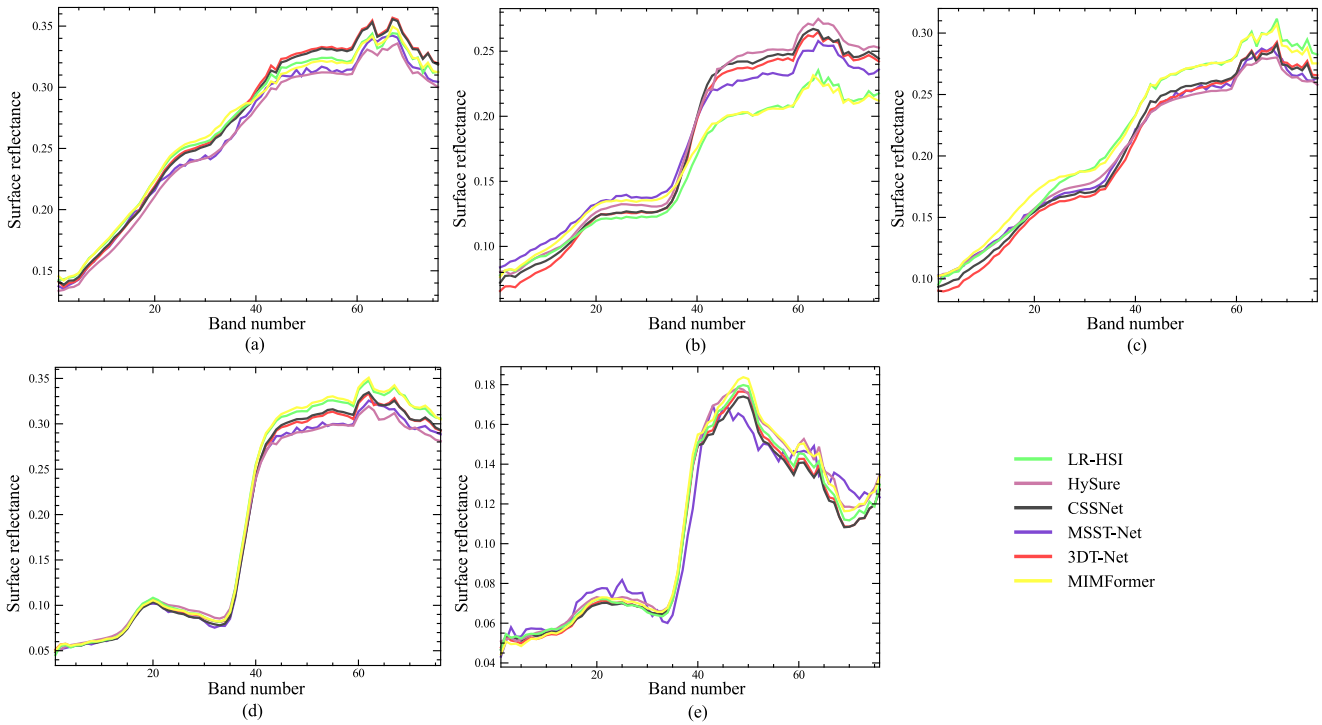


Fig. 10. Spectral contrast of diffferent objects in the ZY1E dataset. (a) Bright roof building. (b) Highway. (c) Playground. (d) Cultivated land. (e) Lake.

D. Experiments With Real Data

To validate the robustness of our method on real data, we created a real paired LR-HSI and HR-MSI dataset, named ZY1E. Fig. 9 displays the fused images from the ZY1E dataset, which include diverse scenes, such as water bodies, buildings, roads, and farmland, generated by different methods. We selected a 100×100 region within the red box and enlarged it in the bottom left corner of the fused image. A clear comparison reveals several key observations: the traditional HySure method exhibits severe distortion and spatial aliasing; CSSNet produces images with varying degrees of color distortion; and while MSST-Net and 3DT-Net show some improvement in spatial details, they still fall short. In contrast, MIMFormer not only achieves a color accuracy closer to LR-HSI but also significantly enhances spatial details, demonstrating superior performance overall.

Fig. 10 illustrates the spectral curves of different objects after fusion by various methods from the ZY1E dataset, compared to the original LR-HSI:

1) bright roof building;
2) highway;
3) playground;
4) cultivated land;
5) lake.

In these plots, our proposed MIMFormer is represented in yellow, while the reference LR-HSI curves are in green. It is evident that the spectral curves of MIMFormer, despite minor discrepancies, closely match those of all objects in LR-HSI, achieving the smallest error. The traditional method HySure overall performs the worst, particularly showing significant deviations from the LR-HSI spectral curves in the highway and cultivated land scenarios. MSST-Net and 3DT-Net exhibit smaller spectral curve errors in the bright roof building and lake scenario. Severe deviations in the highway and lake scenarios. In addition, we employed the QNR image quality assessment metric to evaluate the fusion results of all methods. As shown in Table V, the proposed method achieved the highest score, indicating that MIMFormer outperformed its competitors in terms of image quality. Overall, our method effectively addresses feature redundancy through the spectral splitting mechanism, ensuring superior spectral quality in the fused images. Furthermore, by integrating the strengths of CNNs and transformers, we were

TABLE V
NO-REFERENCE INDEXES FOR THE FUSION RESULTS OF EACH METHOD ON THE ZY1E DATASET

Method	HySure	CSSNet	MSST-Net	3DT-Net	MIMFormer
QNR	0.932	0.951	0.968	0.969	**0.974**

The best indicators are displayed in bold.

able to extract comprehensive spatial and spectral information, resulting in significantly enhanced image quality.

V. DISCUSSION

The key advantage of MIMFormer lies in the introduction of the spectral splitting mechanism, which enhances the control and utilization of different spectral channel characteristics, effectively modeling and extracting spatial–spectral information dispersed across various bands. However, the dim setting within the spectral splitting mechanism primarily depends on the network's C value (number of feature maps) and is based on empirical settings, presenting a limitation of our network. In the future, we plan to adopt methods, such as genetic algorithms and particle swarm optimization to automatically select or explore other approaches for optimal band selection, thereby further optimizing network performance.

In addition, our current network primarily targets well-aligned images. Moving forward, we aim to optimize the MIMFormer network and investigate advanced image alignment techniques to enhance its performance under varying alignment conditions, such as changes in posture and illumination. This will improve its applicability and generalization in practical applications. Finally, we plan to explore cross-dataset generalization strategies to address the generalization issues arising from differences in the number of bands and image features. Through these improvements, our goal is to promote the widespread application of fusion networks, enabling them to demonstrate outstanding performance across various hyperspectral datasets.

VI. CONCLUSION

This article presents a multiscale hybrid network (MIMFormer) based on the Inception structure, combining CNN and transformer, effectively addressing the shortcomings of existing hybrid architectures in modeling local and global features, as well as the lack of flexibility in information processing. By designing the MST and ISSM modules, MIMFormer can capture and fuse local and global information of LR-HSIs and HR-MSIs at different scales, significantly enhancing the spatial and spectral details of the fused images. In addition, the introduction of the spectral splitting mechanism allows MIMFormer to more effectively control and utilize the characteristics of different spectral channels, modeling, and extracting spatial–spectral information dispersed across different bands. Experimental results demonstrate that MIMFormer performs excellently on both benchmark and real-world datasets, maintaining the integrity of spatial and spectral information while being efficient and accurate in processing spectral and spatial information. In the future, we will continue to optimize this network and explore more advanced image alignment and cross-dataset generalization strategies to further enhance its performance in various application scenarios.

REFERENCES

[1] D. Landgrebe, "Hyperspectral image data analysis," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 17–28, Jan. 2002.

[2] E. M. Middleton et al., "The Earth observing one (EO-1) satellite mission: Over a decade in space," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 2, pp. 243–256, Apr. 2013.

[3] C. Niu et al., "Radiometric cross-calibration of the ZY1-02D hyperspectral imager using the GF-5 AHSI imager," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5519612.

[4] Y. Peng et al., "Estimation of soil nutrient content using hyperspectral data," *Agriculture*, vol. 11, no. 11, 2021, Art. no. 1129. [Online]. Available: https://www.mdpi.com/2077-0472/11/11/1129

[5] A. Hennessy, K. Clarke, and M. Lewis, "Hyperspectral classification of plants: A review of waveband selection generalisability," *Remote Sens.*, vol. 12, no. 1, 2020, Art. no. 113. [Online]. Available: https://www.mdpi.com/2072-4292/12/1/113

[6] J. Benediktsson, J. Palmason, and J. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 480–491, Mar. 2005.

[7] A. R. Gillespie, A. B. Kahle, and R. E. Walker, "Color enhancement of highly correlated images. II. Channel ratio and "chromaticity" transformation techniques," *Remote Sens. Environ.*, vol. 22, no. 3, pp. 343–365, 1987.

[8] L. Gao, J. Li, K. Zheng, and X. Jia, "Enhanced autoencoders with attention-embedded degradation learning for unsupervised hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5509417.

[9] 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, 2021.

[10] D. Sara, A. K. Mandava, A. Kumar, S. Duela, and A. Jude, "Hyperspectral and multispectral image fusion techniques for high resolution applications: A review," *Earth Sci. Informat.*, vol. 14, no. 4, pp. 1685–1705, 2021.

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

[12] 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, 2024.

[13] X. Qin, H. Song, J. Fan, and K. Zhang, "Spatio-spectral cross-attention transformer for hyperspectral image and multispectral image fusion," *Remote Sens. Lett.*, vol. 14, no. 12, pp. 1303–1314, 2023.

[14] X. Cao, Y. Lian, K. Wang, C. Ma, and X. Xu, "Unsupervised hybrid network of transformer and CNN for blind hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5507615.

[15] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12116–12128.

[16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.

[17] Z. Huang, Q. Chen, Q. Chen, and X. Liu, "Variational pansharpening for hyperspectral imagery constrained by spectral shape and Gram–Schmidt transformation," *Sensors*, vol. 18, no. 12, 2018, Art. no. 4330.

[18] J. Duran, A. Buades, B. Coll, and C. Sbert, "A nonlocal variational model for pansharpening image fusion," *SIAM J. Imag. Sci.*, vol. 7, no. 2, pp. 761–796, 2014.

[19] G. Vivone et al., "A critical comparison among pansharpening algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2565–2586, May 2015.

[20] B. Huang, H. Song, H. Cui, J. Peng, and Z. Xu, "Spatial and spectral image fusion using sparse matrix factorization," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 3, pp. 1693–1704, Mar. 2014.

[21] M. Simoes, J. Bioucas-Dias, L. B. Almeida, and J. Chanussot, "Hyperspectral image superresolution: An edge-preserving convex formulation," in *Proc. IEEE Int. Conf. Image Process.*, 2014, pp. 4166–4170.

[22] L. Zhang, J. Nie, W. Wei, Y. Li, and Y. Zhang, "Deep blind hyperspectral image super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2388–2400, Jun. 2021.

[23] R. B. Gomez, A. Jazaeri, and M. Kafatos, "Wavelet-based hyperspectral and multispectral image fusion," in *Geo-Spatial Image and Data Exploitation II*, vol. 4383. Bellingham, WA, USA: SPIE, 2001, pp. 36–42.

[24] P. Kwarteng and A. Chavez, "Extracting spectral contrast in landsat thematic mapper image data using selective principal component analysis," *Photogrammetric Eng. Remote Sens.*, vol. 55, no. 1, pp. 339–348, 1989.

[25] F. Palsson, J. R. Sveinsson, M. O. Ulfarsson, and J. A. Benediktsson, "Model-based fusion of multi-and hyperspectral images using PCA and wavelets," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2652–2663, May 2015.

[26] T.-M. Tu, S.-C. Su, H.-C. Shyu, and P. S. Huang, "A new look at IHS-like image fusion methods," *Inf. Fusion*, vol. 2, no. 3, pp. 177–186, 2001.

[27] C. A. Laben and B. V. Brower, "Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening," U.S. Patent 6,011,875, Jan. 4, 2000.

[28] 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.

[29] M. R. Vicinanza, R. Restaino, G. Vivone, M. D. Mura, and J. Chanussot, "A pansharpening method based on the sparse representation of injected details," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 1, pp. 180–184, Jan. 2015.

[30] P. Chavez, S. C. Sides, and J. A. Anderson, "Comparison of three different methods to merge multiresolution and multispectral data-landsat TM and SPOT panchromatic," *Photogrammetric Eng. Remote Sens.*, vol. 57, no. 3, pp. 295–303, 1991.

[31] M. J. Shensa, "The discrete wavelet transform: Wedding the a trous and mallat algorithms," *IEEE Trans. Signal Process.*, vol. 40, no. 10, pp. 2464–2482, Oct. 1992.

[32] K. Zhang, M. Wang, S. Yang, and L. Jiao, "Spatial–spectral-graph-regularized low-rank tensor decomposition for multispectral and hyperspectral image fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 4, pp. 1030–1040, Apr. 2018.

[33] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled non-negative matrix factorization (CNMF) for hyperspectral and multispectral data fusion: Application to pasture classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2011, pp. 1779–1782.

[34] T. Xu et al., "A coupled tensor double-factor method for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5515417.

[35] A. Karami, M. Yazdi, and G. Mercier, "Compression of hyperspectral images using discerete wavelet transform and tucker decomposition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 2, pp. 444–450, Apr. 2012.

[36] Y. Zhang, S. D. Backer, and P. Scheunders, "Noise-resistant wavelet-based Bayesian fusion of multispectral and hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 11, pp. 3834–3843, Nov. 2009.

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

[38] M. Shafiq and Z. Gu, "Deep residual learning for image recognition: A survey," *Appl. Sci.*, vol. 12, no. 18, 2022, Art. no. 8972.

[39] C. Si, W. Yu, P. Zhou, Y. Zhou, X. Wang, and S. Yan, "Inception transformer," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 23495–23509.

[40] R. Dian, S. Li, A. Guo, and L. Fang, "Deep hyperspectral image sharpening," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5345–5355, Nov. 2018.

[41] X.-H. Han, Y. Zheng, and Y.-W. Chen, "Multi-level and multi-scale spatial and spectral fusion CNN for hyperspectral image super-resolution," in *2019 IEEE/CVF Int. Conf. Comput. Vis. Workshop*, 2019, pp. 4330–4339.

[42] 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.

[43] X. Wu et al., "Multi-task multi-objective evolutionary network for hyperspectral image classification and pansharpening," *Inf. Fusion*, vol. 108, 2024, Art. no. 102383. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253524001611

[44] X. Wu, J. Feng, R. Shang, X. Zhang, and L. Jiao, "Multiobjective guided divide-and-conquer network for hyperspectral pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5525317.

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

[46] 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.

[47] 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, 2022, Art6 no. 012305.

[48] S. Jia, Z. Min, and X. Fu, "Multiscale spatial–spectral transformer network for hyperspectral and multispectral image fusion," *Inf. Fusion*, vol. 96, pp. 117–129, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253523000921

[49] J. Fang, J. Yang, A. Khader, and L. Xiao, "MIMO-SST: Multi-input multi-output spatial-spectral transformer for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–20, 2024.

[50] H. Wu, C. Wang, C. Lu, and T. Zhan, "HCT: A hybrid CNN and transformer network for hyperspectral image super-resolution," *Multimedia Syst.*, vol. 30, no. 4, 2024, Art. no. 185.

[51] J. Li, H. Xing, Z. Ao, H. Wang, W. Liu, and A. Zhang, "Convolution-transformer adaptive fusion network for hyperspectral image classification," *Appl. Sci.*, vol. 13, no. 1, 2022, Art. no. 492.

[52] 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, 2023, Art. no. 101907.

[53] Y. Liu, J. Hu, X. Kang, J. Luo, and S. Fan, "Interactformer: Interactive transformer and CNN for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5531715.

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

[55] Q. Yuan, Y. Wei, X. Meng, H. Shen, and L. Zhang, "A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 3, pp. 978–989, Mar. 2018.

[56] X. Liu, Q. Liu, and Y. Wang, "Remote sensing image fusion based on two-stream fusion network," *Inf. Fusion*, vol. 55, pp. 1–15, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253517308060

[57] 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.

[58] 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.

[59] S. Jia, Z. Min, and X. Fu, "Multiscale spatial–spectral transformer network for hyperspectral and multispectral image fusion," *Inf. Fusion*, vol. 96, pp. 117–129, 2023.

[60] G. Vivone, "Multispectral and hyperspectral image fusion in remote sensing: A survey," *Inf. Fusion*, vol. 89, pp. 405–417, 2023.

[61] 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.

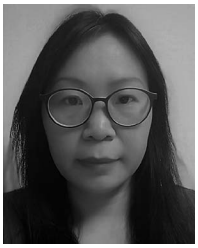
[62] N. Yokoya, C. Grohnfeldt, and J. Chanussot, "Hyperspectral and multispectral data fusion: A comparative review of the recent literature," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 2, pp. 29–56, Jun. 2017.

[63] Y. Xu et al., "Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1709–1724, Jun. 2019.

[64] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogrammetric Eng. Remote Sens.*, vol. 63, no. 6, pp. 691–699, 1997.

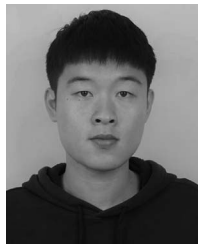
Rumei Li received the bachelor's degree in information management and information system from the Chongqing University of Technology, Chongqing, China, in 2022. She is currently working toward the master's degree in hydraulic engineering with Capital Normal University, Beijing, China.

Her current research interest is in deep learning in remote sensing.



Liyan Zhang received the B.S. degree from the Shandong University of Technology, Zibo, China, in 2000, the M.S. degree from Beijing University of Technology, Beijing, China, in 2004, and the Ph.D. degree from the Beijing Institute of Technology, Beijing, China, in 2007, all in mechanical engineering.

She is currently a Lecturer with the Capital Normal University, Beijing, China. She was a Postdoctoral Research in Geography with the Capital Normal University in 2019. Her current research interests include hyperspectral imagery and its application.



Zun Wang received the bachelor's degree in computer science and technology from Henan Normal University, Henan, China, in 2023. He is currently working toward the master's degree in surveying and mapping engineering with Capital Normal University, Beijing, China.

His research interests include hyperspectral image super-resolution.

Xiaojuan Li received the Ph.D. degree in geoinformatics from the Institute of Remote Sensing Applications, Chinese Academy of Science, Beijing, China, in 1999.

She is currently a Professor with the Beijing Advanced Innovation Center for Imaging Technology and Key Laboratory of 3D Information Acquisition and Application, College of Resource Environment and Tourism, Capital Normal University, Beijing, China. Her research interest includes remote sensing of environment and geographical information science.