Enhanced Spatial-Frequency Synergistic Network for Multispectral and Hyperspectral Image Fusion

Meng Xu, *Member, IEEE*, Ziqian Mo, *Graduate Student Member, IEEE*, Xiyou Fu, *Member, IEEE*, and Sen Jia, *Senior Member, IEEE*

***Abstract*— Multispectral and hyperspectral image fusion (MHIF) seeks to combine high-resolution multispectral images (HR-MSIs) with low-resolution hyperspectral images (LR-HSIs) to create high-resolution hyperspectral images (HR-HSIs). Transformer-based architectures have recently become prominent in MHIF tasks due to their effective global self-attention mechanisms. However, the quadratic computational complexity of the global self-attention in Transformers presents significant challenges for practical applications. In this article, we propose an enhanced spatial–frequency synergistic (ESFS) approach that leverages both spatial and frequency-domain features to enhance fusion quality. Our ESFS framework introduces the condensed spatial augmentation module (CSAM), which condenses window features and employs cross-attention to balance extensive contextual understanding and detailed local feature extraction while reducing computational overhead. Additionally, we develop the selective frequency decomposition module (SFDM), which utilizes global filters composed of phase and amplitude information in the frequency domain to retain features, effectively capturing deep frequency-domain characteristics and their interdependencies. Comprehensive experiments on three benchmark MHIF datasets demonstrate that our method achieves superior performance, establishing a new state-of-the-art (SOTA) in both quantitative metrics and visual quality assessments. The code is available at http://szu-hsilab.com/**

***Index Terms*— Deep learning, Fourier transform, multispectral and hyperspectral image fusion (MHIF), spatial–frequency synergistic, Transformer-based method.**

I. Introduction

HYPERSPECTRAL images (HSIs) provide extensive spectral information across hundreds to thousands of narrow bands, capturing the unique characteristics of various materials. The extensive spectral information contained in HSIs renders them indispensable for a wide range of applications, including classification [1], [2], [3], object detection [4], tracking [5], [6], and segmentation [7], [8]. However, the pursuit of high spectral resolution in hyperspectral imaging often compromises spatial resolution, leading to low-resolution HSIs (LR-HSIs).

In contrast, multispectral imaging systems offer high spatial resolution at the expense of spectral details, yielding high-resolution multispectral images (HR-MSIs). Research in multispectral and hyperspectral image fusion (MHIF) aims to combine LR-HSIs with HR-MSIs to produce HSIs with high-resolution hyperspectral images (HR-HSIs). In practice, HR-MSIs provide crucial structural information that aids in reconstructing higher-resolution images. MHIF technology not only utilizes this structural information but also extracts precise spectral information from LR-HSI, resulting in enhanced image richness and accuracy.

Current methodologies for MHIF can be broadly classified into two major categories: traditional methods and deep learning-based approaches. Traditional methods rely on exploiting intrinsic attributes under specific prior knowledge, such as a sparse prior and self-similarity. These include Bayesian-based methods [9] and matrix factorization-based methods [10]. While these techniques have shown decent results in MHIF, they still face significant challenges in efficiently transferring spatial and spectral information.

In recent years, deep learning has achieved notable success in MHIF, demonstrating its potential in various aspects of image super-resolution and fusion [11]. These approaches effectively capture high-level features from input images, generating more accurate and detailed super-resolution images through multiple iterations. CNNs are widely used for image feature extraction in fusion methods [12], [13], [14], [15]. However, the relatively small receptive fields of CNNs restrict their ability to capture global features effectively, which affects the overall performance of CNN-based methods. As an alternative to CNNs, vision Transformer (ViT) [16] has demonstrated impressive performance across a range of computer vision tasks. ViT employs a self-attention mechanism that excels at capturing global interactions by analyzing relationships between tokens.

Recently, swin Transformer [17] has demonstrated significant potential by combining the strengths of both CNNs and Transformers. It employs a window attention mechanism, restricting self-attention computation to nonoverlapping local windows, which enhances computational efficiency. Indeed, swin Transformer for image restoration (SwinIR) [18] builds upon the architecture of the swin Transformer, delivering

Received 19 March 2025; revised 17 June 2025; accepted 9 July 2025. Date of publication 14 July 2025; date of current version 23 July 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 42271336, Grant 62271327, and Grant 42301375; in part by the Natural Science Foundation of Guangdong Province under Grant 2024A1515011079, Grant 2022A1515011290, and Grant 2022A1515110076; in part by the Project of Department of Education of Guangdong Province under Grant 2023KCXTD029; and in part by Shenzhen Science and Technology Program under Grant RCJC20221008092731042, Grant JCYJ20220818100206015, Grant KQTD20200909113951005, and Grant JCYJ20240813141635047. *(Corresponding author: Sen Jia.)*

The authors are with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: m.xu@szu.edu.cn; 2310275043@email.szu.edu.cn; fuxy0623@szu.edu.cn; senjia@szu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2025.3589097

strong performance in various low-level image processing tasks. The three-stage architecture of SwinIR, comprising shallow feature extraction, deep feature extraction, and image reconstruction, has established a foundational framework for numerous MHIF approaches [19], [20].

While recent spatial-domain models, such as multiscale CNNs and transformer architectures, are capable of capturing both local details and global contextual information in HSIs, they fundamentally rely on spatial correlations within the image. These methods excel at modeling spatial structures and interband relationships through localized convolutional operations or attention mechanisms. However, they often implicitly learn spatial patterns without explicitly disentangling different frequency components of the image.

Frequency-domain analysis offers a complementary perspective by explicitly decomposing an image into different frequency components, enabling a more structured understanding of the information content. Specifically, low-frequency components correspond to broad, smooth variations, while high-frequency components capture fine-grained details such as edges and textures. Importantly, in HSIs, subtle but critical spectral variations often manifest differently across frequency bands. Frequency-domain representations can more clearly separate and highlight these variations, which may otherwise be entangled in the spatial domain.

However, when relying solely on frequency-domain methods, some important spatial features may be lost, leading to incomplete image representations. To address this challenge, it is essential to combine both spatial and frequency-domain information, as each domain captures complementary aspects of the image. While spatial-domain methods excel at capturing spatial patterns and contextual features, frequency-domain methods provide complementary benefits by explicitly modeling spectral variations and enhancing discriminative representations. By integrating the strengths of both approaches, it becomes possible to preserve fine-grained spectral details from the frequency domain while simultaneously capturing global dependencies, resulting in a more accurate and comprehensive fusion of LR-HSIs with HR-MSIs.

Compared with existing dual-domain methods such as the spatial–frequency information integration network (SFINet) [21], which applies fixed-window modeling and simple fusion strategies, our method introduces specialized modules tailored for MHIF. Specifically, we propose an enhanced spatial–frequency synergistic (ESFS) network, which integrates spatial and frequency-domain features through two core components: the condensed spatial augmentation module (CSAM) and the selective frequency decomposition module (SFDM). CSAM adaptively captures fine-grained spatial patterns while optimizing computational efficiency by compressing attention operations. SFDM selectively enhances informative frequency components and suppresses noise, enabling more refined and complementary feature representations. These modules are integrated within a residual-guided refinement framework, allowing spatial and frequency cues to interact progressively and synergistically, ultimately enhancing the fusion quality of LR-HSIs and HR-MSIs. In summary, the contributions of this article are as follows.

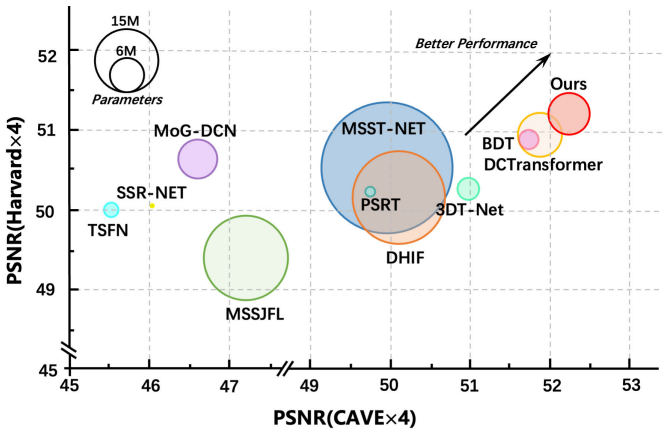

Fig. 1. Comparison of our method and other approaches on the CAVE (×4) and Harvard (×4) datasets. Circles located closer to the top-right corner represent models with better performance, while the circle size corresponds to the number of parameters in each model.

1) The proposed ESFS network effectively integrates the spatial and frequency-domain features to enhance fusion quality by leveraging the advantages of both domains.
2) CSAM achieves a balance between capturing broad contextual information and extracting fine local details. SFDM ensures the retention of essential frequency components while filtering out unnecessary details.
3) Extensive experiments on three benchmark datasets demonstrate that our method exhibits superior visual quality compared to existing techniques and achieves SOTA performance in quantitative metrics. Fig. 1 presents a balanced comparison with other SOTA methods.

II. Related Works

A. Traditional Methods

Traditional MHIF methods can be categorized into three main approaches: matrix factorization-based methods, Bayesian-based methods, and tensor factorization-based methods. Matrix factorization-based methods often decompose 3-D HSIs into 2-D matrices of endmembers and abundances. Yokoya et al. [10] employed coupled nonnegative matrix factorization (CNMF) to independently extract these components from HSIs and MSIs, yielding a fused image with enhanced spectral and spatial resolutions. Other methods include joint unmixing of input images to extract pure reflectance spectra and mixing coefficients [22] or leveraging nonnegative dictionary learning with spatial–spectral sparsity and nonlocal priors to improve reconstruction [23]. Li et al. [24] further refined sparse decomposition using adaptive techniques and iterative optimization to enhance accuracy and adaptability.

Bayesian-based methods offer a probabilistic framework for fusion. Akhtar et al. [25] developed a Bayesian sparse coding approach with dictionaries learned via the Beta process, while Wei et al. [26] proposed the Fast fUsion based on Sylvester Equation (FUSE), combining multiplier alternating

direction methods with block coordinate descent to incorporate problem-specific priors.

Tensor factorization-based methods extend traditional decomposition to high-dimensional tensors, enabling more effective extraction of spatial–spectral information. Coupled tensor approaches address the limitations of matrix factorization-based methods by jointly modeling LR-HSIs and HR-MSIs [27]. Methods such as coupled sparse tensor factorization (CSTF) [28] and low tensor-train rank (LTTR) [29] exploit spectral and nonlocal spatial correlations, laying the groundwork for improving fusion accuracy and robustness. Recent studies, such as Bayesian nonlocal patch tensor factorization (BNPTF) [30], have tackled challenges in rank determination and modeling capacity, achieving improved performance in hyperspectral image fusion.

While these methods have laid a foundation for MHIF, their reliance on assumptions, simplified models, and parameter tuning can result in information loss and limited adaptability. These shortcomings underscore the need for more flexible, data-driven approaches like deep learning, which can better capture the complexity of hyperspectral data.

B. Learning-Based Methods

With the advancement of deep learning, it has become evident that this technology excels in capturing the intricate features of HSIs and MSIs. For instance, Wang et al. [31] were among the first to leverage deep residual CNNs to tackle the MHIF problem. DHSIS [32] advanced HSI sharpening for MHIF by leveraging a deep CNN-based residual learning approach to directly learn image priors. Nonetheless, these methods do not fully exploit the potential of deep end-to-end learning. The spatial–spectral reconstruction network (SSRNet) [33] presented a physically intuitive CNN architecture, which employs separate loss functions for optimizing spatial and spectral reconstruction processes. Additionally, model-guided deep convolutional network (MoG-DCN) [34] was introduced as an alternative to ResNet for acquiring a denoising prior, providing a more structured approach to capturing spatial details while effectively reducing noise. In the pursuit of improving the spatial and spectral quality of HSIs, GuidedNet [35] introduced a framework that integrates multiscale high-resolution guidance, effectively reducing network parameters and computational cost through recursive strategies. Similarly, KNLConv [36] proposed an innovative approach by incorporating nonlocal dependencies into the convolutional kernel space, offering a more flexible and global feature extraction mechanism that enhances performance for HSI super-resolution.

The Transformer architecture has shown robust performance in various vision tasks, prompting many researchers to explore its application to MHIF. Fusformer [37] pioneered the use of Transformers for image fusion, achieving impressive results with a lightweight network. The multiscale spatial–spectral Transformer network (MSST-Net) [38] incorporated a self-supervised pre-training strategy designed to enhance the network's performance and generalization capabilities. Pyramid shuffle-and-reshuffle Transformer (PSRT) [19] employs the swin Transformer framework, integrating shuffle-and-reshuffle strategies with multiscale feature extraction, thereby enabling the learning of both local and long-range representations. Additionally, 3DT-Net [39] adapted MoG-DCN by replacing the U-net architecture with a Transformer-based model. DCTransformer [20] captures the interplay between modalities through directional pairwise multihead cross-attention. These approaches showcase the versatility of Transformer-based methods in enhancing MHIF.

Recently, several works have introduced more advanced spatial-domain fusion strategies based on Transformers, including progressive interaction and cross-modality attention mechanisms. For example, MFT-GAN [40] incorporates multiscale spatial feature guidance within a Transformer-based GAN framework to enhance spatial details in an unsupervised manner. The unsupervised hybrid network of Transformer and CNN (uHNTC) [41] combines CNNs and Transformers in a dual-branch architecture to jointly model local spatial textures and global spectral dependencies. The multiscale deep cross-fusion Transformer (MDC-FusFormer) [42] and the unsupervised multilevel spatiospectral fusion Transformer (UMSFT) [43] both propose multiscale deep cross-fertilization modules to enhance spatial–spectral information flow. Additionally, CYformer [44] designs a cyclic cross-modality attention mechanism to iteratively refine intermodal alignment.

While deep learning-based methods have made significant strides in MHIF, they primarily focus on spatial-domain feature extraction and have demonstrated strong capabilities in capturing fine-grained textures and local spatial structures. However, relying solely on spatial-domain features may limit their ability to fully model the spectral correlations and global contextual information inherent in HSIs and MSIs, particularly in complex or multiscale scenarios.

To address such challenges, frequency-domain methods have emerged as a promising complementary direction. By representing signals in the spectral domain, these approaches naturally capture long-range dependencies and preserve global structural consistency. Rather than replacing spatial modeling, frequency-domain analysis offers an alternative and synergistic perspective, highlighting the potential benefits of integrating both spatial and frequency-domain information to enhance fusion quality.

C. Fourier Transform

Fourier transform is widely used for analyzing frequency components in signals, offering a comprehensive view of long-range dependencies [45], [46]. Several researchers have applied it to computer vision applications. For example, the global filter network (GFNet) [47] substitutes self-attention in vision Transformer with a Fourier transform and a learnable filter, enabling long-term spatial dependencies in the frequency domain. In the field of image fusion, the Fourier transform plays a crucial role by providing a way to handle global frequency information, which can complement the local spatial features extracted by spatial-domain methods. The frequency integration and spatial compensation network (FISCNet) [48] integrate frequency-domain phase components

and spatial-domain features to enhance salient object preservation and texture fidelity in infrared and visible image fusion. Similarly, spatial-frequency domain fusion (SFDFusion) [49] proposes a dual-modality approach for infrared and visible image fusion, integrating spatial-domain refinement and frequency-domain information, leveraging FFT to enhance image quality and efficiency. Further advancing these ideas, SFINet [21] integrates both spatial and frequency-domain information for multimodal image fusion, using a dual-branch architecture with spatial convolution and modality-aware deep Fourier transformation, enhancing both local and global feature representations. Similarly, the hierarchical frequency integration network (HFIN) [50] hierarchically decomposes panchromatic images and low-resolution multispectral images into spatial, global Fourier, and local Fourier components, and integrates them to enhance spatial-frequency relationships for pan-sharpening. Additionally, Fourier-enhanced implicit neural fusion network (FeINFN) [51] transforms latent codes into the frequency domain, integrating amplitude and phase representations while enhancing high-frequency details. To leverage frequency-domain priors for image restoration, FFTFormer [52] introduces a frequency-domain self-attention mechanism and a gated feed-forward network to enhance deblurring performance. Similarly, F2Former [53] exploits the fractional Fourier transform to build a unified spatial-frequency representation, enabling more effective frequency-aware attention and feature refinement. Collectively, these studies underscore the effectiveness of the Fourier transform in improving performance, particularly in complex visual tasks such as MHIF, where capturing and integrating frequency information is crucial for accurate image fusion.

However, while Fourier-based methods have shown great promise in the MHIF field, many of them treat the frequency domain and spatial domain as separate entities, without fully exploiting the complementary properties of both. This separation limits the ability to capture the full spectrum of features that are crucial for tasks such as image fusion, where both local fine-grained details and global contextual information are equally important. Frequency-domain methods can effectively capture long-range dependencies and global information, while spatial-domain methods excel in preserving local features, textures, and spatial coherence.

In contrast, our proposed ESFS method integrates both spatial and frequency-domain features, ensuring that the advantages of both domains are leveraged to achieve superior fusion quality. By combining these two domains, ESFS can capture both fine-grained spatial details and long-range frequency dependencies simultaneously. The spatial domain provides detailed local features that are essential for accurate image reconstruction, while the frequency domain captures global structural patterns and dependencies that may not be immediately apparent in the spatial domain. This hybrid approach allows ESFS to effectively merge complementary features from both domains, leading to a more comprehensive understanding of the input data and ultimately improving the performance of MHIF tasks. By integrating these domains rather than treating them separately, ESFS harnesses the complementary strengths of spatial and frequency information, resulting in more accurate, detailed, and robust fusion outputs.

III. Method

In this section, we first present our ESFS approach specifically designed for the MHIF task. Subsequently, we introduce the implementation of the composite modules within the proposed architecture.

A. Overall Network Structure

We commence with a detailed exposition of the overarching structure of our ESFS architecture, as illustrated in Fig. 2. The network processes LR-HSI $\mathbf{X}^{\mathrm{LR}} \in \mathbb{R}^{h\times w\times C}$ and HR-MSI $\mathbf{Y}^{\mathrm{HR}} \in \mathbb{R}^{H\times W\times c}$ as inputs to generate HR-HSI $\mathbf{X}^{\mathrm{HR}} \in \mathbb{R}^{H\times W\times C}$. The term $r = H/h = W/w$ stands for the upsampling scale. Initially, we concatenate the bicubic interpolated $\mathbf{X}^{\mathrm{LR}}_{\mathrm{up}} \in \mathbb{R}^{H\times W\times C}$ and $\mathbf{Y}^{\mathrm{HR}} \in \mathbb{R}^{H\times W\times c}$. Subsequently, a 3×3 convolutional layer is applied to the concatenated input to extract shallow features. The extracted shallow features are then fed into three spatial frequency residual groups (SFRGs) to extract deep features. To capitalize on the varying levels of information extracted by deep neural networks, the module incorporates dense connections [54]. These connections improve feature propagation, promote feature reuse, and mitigate the vanishing-gradient problem. Each SFRG integrates these dense connections to optimize the extraction of deeper features.

CSAM plays a critical role in enhancing spatial relationships by effectively capturing spatial dependencies between the LR-HSI and HR-MSI inputs. Specifically, CSAM utilizes a condensed window multihead self-attention (CW-MSA) mechanism to capture long-range spatial interactions, thereby augmenting the spatial features and providing a more context-aware representation of the input images. This refinement allows the model to capture both local fine-grained details and global spatial patterns, which are essential for accurate fusion in MHIF tasks. Once the spatial features are enhanced by CSAM, they are passed to the SFDM. The reasoning behind this sequential flow is that CSAM first focuses on enhancing the spatial resolution and contextual understanding of the image, which is essential for preserving fine details. In contrast, SFDM processes these refined spatial features by incorporating frequency-domain information. Instead of extracting spectral features directly, SFDM emphasizes frequency-domain dependencies, utilizing frequency-domain transformations to capture global context and interdependencies that are not readily apparent in the spatial domain. This allows the model to handle global structural information and integrate high-frequency content, which complements the spatial features enhanced by CSAM. By combining spatial and frequency-domain processing in a sequential manner, our model leverages the strengths of both domains. CSAM ensures that detailed spatial patterns are well-preserved, while SFDM enriches the model's understanding of global context through frequency-based processing. This integration leads to a more comprehensive fusion of HSIs and MSIs, improving both the spatial and global structural accuracy of the fused image.

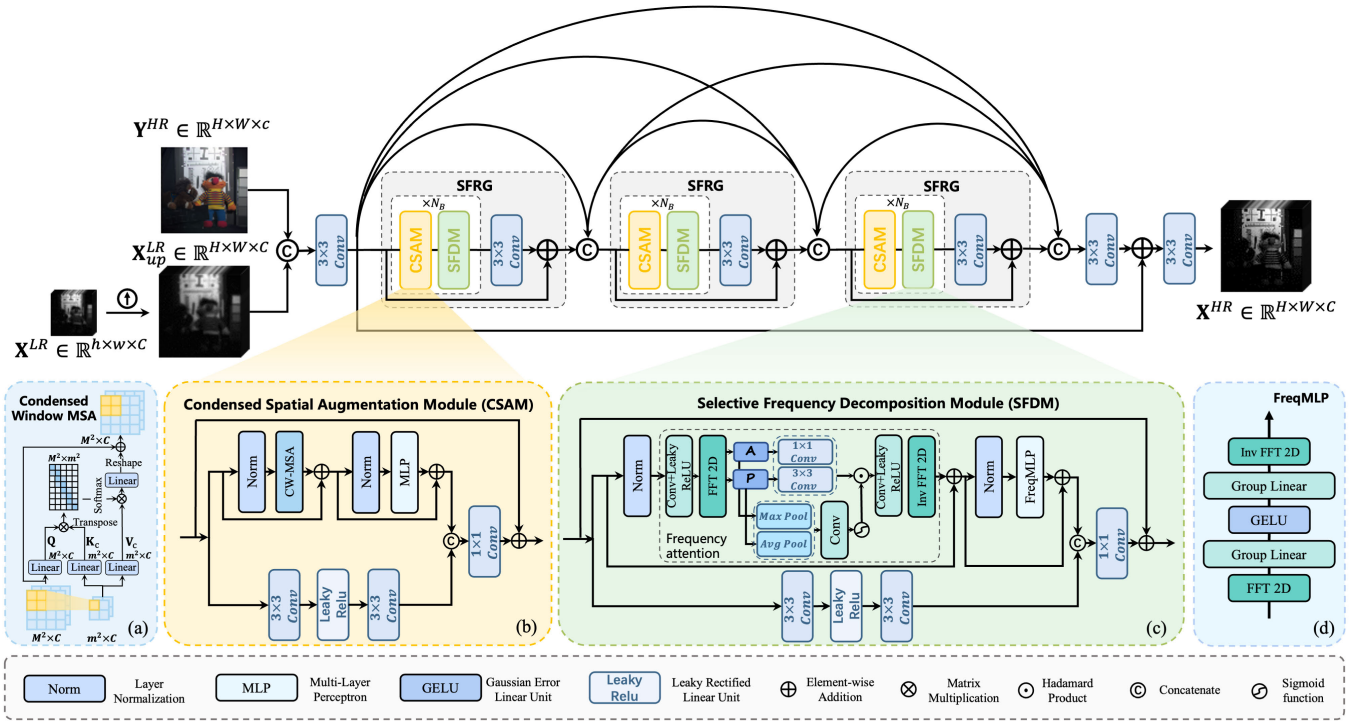

Fig. 2. Architecture of the proposed ESFS network. (a) CW-MSA, which balances global contextual understanding and local feature enhancement by compressing window features and applying multihead cross-attention. (b) CSAM, which enhances spatial relationships between the LR-HSI and HR-MSI by capturing spatial dependencies effectively. (c) SFDM, designed to capture frequency-based dependencies by decomposing features into amplitude and phase components and applying specialized convolutions. (d) FreqMLP, a frequency-domain feed-forward network that leverages group linear layers and GELU activations for advanced frequency manipulation.

B. Condensed Spatial Augmentation Module

The CSAM is designed to enhance spatial feature extraction by capturing spatial dependencies between LR-HSI and HR-MSI. CSAM introduces a convolutional branch that combines the results of the CW-MSA with convolutional operations. This dual approach is aimed at preserving global information captured by the attention mechanism while refining local features through convolution, ensuring that spatial relationships are both preserved and enhanced.

The convolutional branch is included to counteract the potential loss of fine-grained spatial details that can occur when relying solely on the CW-MSA mechanism. While CW-MSA excels at capturing global context and long-range dependencies, it may not fully capture fine local spatial features, which are crucial for accurate fusion in MHIF tasks. The convolutional branch is therefore designed to refine the spatial features locally, ensuring that small-scale details are preserved while the global dependencies are maintained.

By integrating both the attention-based and convolutional features, CSAM effectively balances the need for global context and detailed local feature enhancement. This hybrid approach strengthens the model's ability to capture complex spatial interactions, ensuring that both long-range dependencies and fine-grained spatial details are optimally fused. Ultimately, this dual approach improves the model's ability to generate more accurate and coherent fused images, enhancing the overall performance of the ESFS network.

Our proposed CW-MSA combines the global feature extraction capability of window-based attention with the local feature enhancement of convolution. By compressing window features through convolution, CW-MSA preserves crucial global information while reducing computational complexity. It effectively balances the need for extensive contextual understanding and detailed local feature enhancement, which is critical in MHIF. By capturing the intricate spectral–spatial relationships unique to these modalities, CW-MSA enhances the fusion process, leading to more accurate and representative feature extraction. This not only optimizes computational efficiency but also significantly improves the fidelity and quality of the fused images, ensuring that the complementary information from both HSI and MSI sources is fully leveraged.

CW-MSA initially condenses the original window features into a representative feature map, which intuitively aggregates the information of the entire window. As depicted in Fig. 2(a), given an input feature $\mathbf{X}_{\text{in}} \in \mathbb{R}^{H\times W\times C}$, it is first partitioned into (HW/M^2) local windows, with the size of each window $M \times M$. For a local window feature $\mathbf{X}_w \in \mathbb{R}^{M^2\times C}$, we first reduce the spatial dimensions through iterative depthwise convolutions [55], transitioning from the original dimension $M \times M$ to a smaller dimension $m \times m$. This reduction is achieved through multiple iterations of depthwise convolutions, each configured with a kernel size of 2×2 and a stride of 2. In the first iteration, the spatial dimensions are reduced from $M \times M$ to $(M/2) \times (M/2)$, and with further iterations, the dimensions are progressively reduced until they reach $m \times m$. The resulting feature map is a coarsely condensed aggregation map. This coarse map is then refined using depthwise separable convolutions [55], resulting in a condensed refined representation $\mathbf{X}_c \in \mathbb{R}^{m^2\times C}$ while preserving the channel dimensions to maintain the expressive capacity of

the attention maps generated by each attention head. Cross-attention is then performed using the $\mathbf{Q} \in \mathbb{R}^{M^2 \times d}$ (query) generated from the original window features, along with the $\mathbf{K}_c \in \mathbb{R}^{m^2 \times d}$ (key) and $\mathbf{V}_c \in \mathbb{R}^{m^2 \times d}$ (value) derived from the representative feature map. The attention matrix is computed based on the dot-product interaction between the query and the key. The mathematical formulation of the proposed CW-MSA is presented as follows:

$$\mathbf{Q} = \mathbf{X}_w \mathbf{W}_Q, \quad \mathbf{K}_c = \mathbf{X}_c \mathbf{W}_K$$
$$\mathbf{V}_c = \mathbf{X}_c \mathbf{W}_V \tag{1}$$

$$\text{CW-MSA}(\mathbf{Q}, \mathbf{K}_c, \mathbf{V}_c) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}_c^{\mathrm{T}}}{\sqrt{d}} + B\right)\mathbf{V}_c \tag{2}$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V \in \mathbb{R}^{C \times d}$ are the projection matrices and $B \in \mathbb{R}^{M^2 \times m^2}$ represents an aligned relative position embedding, obtained by interpolating the original embedding defined in [17], as the window size of $\mathbf{Q}$ differs from that of $\mathbf{K}_c$. The term $\sqrt{d}$ is a scalar as defined in [16]. We execute the cross-attention function h times concurrently and then concatenate the outcomes to implement multihead cross-attention. Consistent with the method described in [17], we apply shifted window partitioning over two successive CW-MSA.

To quantify the computational benefits, we analyze the complexity of CW-MSA and compare it to the original window-based multi-head self-attention (W-MSA) [17]. Given an input feature with the size of $H \times W$, the image is divided into nonoverlapping windows with the size of $M \times M$, the original W-MSA performs three distinct linear projections to obtain the query, key, and value, followed by an additional linear projection after the attention mechanism. This process incurs a computational complexity of $\Omega(4M^2C^2)$. Additionally, the complexity associated with the attention computation itself is $\Omega(2M^4C)$. The computational complexity for each individual window is expressed as

$$\Omega(\text{Window}) = 4M^2C^2 + 2M^4C. \tag{3}$$

Thus, the total computational complexity of W-MSA can be expressed as

$$\Omega(\text{W-MSA}) = 4HWC^2 + 2M^2HWC. \tag{4}$$

In contrast, as the spatial dimensions are condensed to $h \times w$, the size of the condensed window feature also reduces to $m \times m$. The proposed CW-MSA reduces the complexity of the linear projections to $\Omega(2M^2C^2 + 2m^2C^2)$. The attention computation complexity is $\Omega(2M^2m^2C)$, as the attention mechanism is performed over both the original and condensed window features. The computational complexity for each condensed window can be represented as

$$\Omega(\text{Condensed Window}) = 2M^2C^2 + 2m^2C^2 + 2M^2m^2C. \tag{5}$$

The overall complexity of CW-MSA is

$$\Omega(\text{CW-MSA}) = 2HWC^2 + 2hwC^2 + 2m^2HWC. \tag{6}$$

Given that $m \ll M$, the proposed CW-MSA offers significant advantages in addressing the computational challenges associated with MHIF tasks. By reducing the computational complexity, CW-MSA is particularly well-suited for handling high-dimensional HSIs, where the burden of processing large spatial dimensions can be substantial. This reduction in complexity not only accelerates the processing speed but also makes the approach more scalable and efficient. Consequently, CW-MSA enhances the feasibility of applying advanced attention mechanisms to large-scale MHIF tasks.

C. Selective Frequency Decomposition Module

The architecture of the SFDM is illustrated in Fig. 2(c). The module is designed to capture frequency-based dependencies by selectively decomposing the input features into frequency domains. SFDM captures intricate frequency-based dependencies that are often overlooked in spatial-based models. By decomposing the input features into distinct frequency domains, the module can more effectively capture and leverage spatial frequency characteristics, which are crucial for enhancing the performance of MHIF tasks.

Although the input features comprise multiple spectral bands, SFDM primarily targets the spatial dimensions during frequency decomposition. Specifically, the spectral channels are jointly processed as separate feature maps, and the Fourier transform is applied independently along the spatial axes of each channel. This design ensures that the spectral structure is preserved throughout the frequency operations, avoiding disruption of interband correlations.

The frequency-domain features are extracted using the 2D-FFT, enabling efficient transformation and analysis of spatial frequency characteristics. This approach enhances the module's capacity to handle the high-dimensional data of HSIs by focusing on frequency-specific patterns. By leveraging 2D-FFT, the SFDM effectively captures and integrates intricate spatial frequency features, contributing to improved performance and precision in MHIF tasks.

Given a feature $\mathbf{X}_{\text{fea}} \in \mathbb{R}^{H \times W \times C}$, we adopt 2D-FFT to obtain the corresponding frequency representations

$$\mathbf{X}_F(u, v) = \mathcal{F}(\mathbf{X}_{\text{fea}}) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} \mathbf{X}_{\text{fea}}(h, w) e^{-2\pi i(uh+vw)} \tag{7}$$

where $\mathcal{F}(\cdot)$ represents 2D-FFT, with u and v indicating specific horizontal and vertical spatial frequencies within the Fourier spectrum $\mathbf{X}_F$. $\mathbf{X}_F$ comprises complex values, expressed as $\mathbf{X}_F = x_f^{\text{re}} + x_f^{\text{im}} \cdot i$, where x_f^{re} and x_f^{im} are the real and imaginary components, respectively. Due to the conjugate symmetry property, the FFT-transformed features retain only half of the spatial dimensions. Therefore, this property also reduces the computational complexity of the network.

In our approach, we explicitly separate the amplitude and phase information by computing their closed-form expressions from the real and imaginary parts, rather than learning this separation implicitly within the network. The amplitude and phase components are then extracted from the spectrum, which can be expressed as follows:

$$\mathcal{A}(\mathbf{X}_F) = \sqrt{\left(x_f^{\text{re}}\right)^2 + \left(x_f^{\text{im}}\right)^2}, \quad \mathcal{P}(\mathbf{X}_F) = \arctan\left[\frac{x_f^{\text{im}}}{x_f^{\text{re}}}\right]. \tag{8}$$

TABLE I
COMPARISON OF METHODS ON CAVE ×4 AND CAVE ×8 DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD, AND THE SECOND-BEST RESULTS ARE UNDERLINED. "M" DENOTES MILLION

Ratio	Methods	CAVE						
		PSNR (↑)	RMSE (↓)	RASE (↓)	SAM (↓)	ERGAS (↓)	#params	#FLOPs (10^{12})
4	CNMF IGARSS'11 [10]	36.62	0.0139	11.051	6.313	5.8879	-	-
	SSR-NET TGRS'21 [33]	46.09	0.0053	5.020	1.607	1.2550	**0.026M**	**0.006**
	TSFN TCSVT'21 [12]	45.64	0.0058	5.309	1.699	1.3274	1.074M	1.964
	MoG-DCN TIP'21 [34]	46.55	0.0054	5.118	1.582	1.2588	7.070M	0.516
	MSSJFL DependSys'21 [13]	47.27	0.0046	4.520	1.415	1.1300	20.725M	0.304
	DHIF-Net TCI'21 [57]	50.20	0.0034	3.287	1.039	0.8219	22.669M	3.507
	PSRT TGRS'23 [19]	49.77	0.0035	3.423	1.070	0.8558	0.247M	0.067
	MSST-NET IF'23 [38]	49.94	0.0034	3.311	1.049	0.8279	34.402M	2.972
	3DT-Net IF'23 [39]	51.35	0.0029	2.934	0.889	0.7000	3.455M	4.236
	BDT IJCAI'23 [58]	51.69	0.0028	2.768	0.862	0.6798	2.657M	0.117
	DCTransformer IF'24 [20]	51.95	0.0027	2.628	0.833	0.6571	8.121M	4.916
	ESFS (Ours)	**52.23**	**0.0026**	**2.513**	**0.812**	**0.6378**	7.876M	3.493
8	CNMF IGARSS'11 [10]	35.97	0.0149	11.695	6.786	6.3013	-	-
	SSR-NET TGRS'21 [33]	44.89	0.0062	5.894	1.866	1.4736	**0.026M**	**0.006**
	TSFN TCSVT'21 [12]	44.94	0.0061	5.871	1.815	1.4678	1.074M	1.964
	MoG-DCN TIP'21 [34]	45.54	0.0058	5.035	1.763	1.4012	7.070M	0.516
	MSSJFL DependSys'21 [13]	45.71	0.0055	5.445	1.719	1.3612	20.725M	0.304
	DHIF-Net TCI'21 [57]	48.91	0.9983	3.927	1.229	0.0041	22.669M	3.507
	PSRT TGRS'23 [19]	48.05	0.0043	4.272	1.338	1.0682	0.247M	0.067
	MSST-NET IF'23 [38]	48.35	0.0042	4.070	1.277	1.0176	34.402M	2.972
	3DT-Net IF'23 [39]	49.71	0.0036	3.443	1.089	0.8654	3.455M	4.236
	BDT IJCAI'23 [58]	50.51	0.0033	3.287	0.986	0.7844	2.657M	0.117
	DCTransformer IF'24 [20]	51.09	0.0031	3.015	0.925	0.7335	8.121M	4.916
	ESFS (Ours)	**51.22**	**0.0030**	**2.927**	**0.914**	**0.7291**	7.876M	3.493

The amplitude component $\mathcal{A}(\mathbf{X}_F) \in \mathbb{R}^{H\times\lceil(W+1)/2\rceil\times C}$ captures the structural information of the image, while the phase component $\mathcal{P}(\mathbf{X}_F) \in \mathbb{R}^{H\times\lceil(W+1)/2\rceil\times C}$ encodes high-frequency details, such as textures and fine-grained variations. Accurately merging information across different spectral bands is crucial in MHIF tasks. Therefore, this separation of amplitude and phase proves particularly advantageous.

Since directly applying 3 × 3 convolutions to the amplitude component may result in spectral leakage and channel misalignment [51], we employ pointwise convolutions to preserve the spectral fidelity. Pointwise convolutions can confine their operations to single spatial positions in the frequency domain and effectively prevent the overlap that could otherwise compromise the structural coherence across channels. This approach allows for the accurate extraction and integration of spectral information across different bands, minimizing the risk of artifacts and ensuring the integrity of the fused data. Conversely, the phase component, which encodes texture details and other fine-grained information, requires 3 × 3 convolutions to effectively capture the spatial information. It can ensure that the edge sharpness and textural consistency across spectral bands are accurately represented. After feature extraction, in order to reconstruct the complete frequency-domain representation, we combine the processed amplitude and phase results by utilizing the polar coordinate transformation.

The frequency domain has a significant property: multiplication in the frequency domain is equivalent to circular convolution in the spatial domain, known as the Convolution Theorem [56]. Leveraging this property, we construct a global filter by applying max pooling to the amplitude and average pooling to the phase. Subsequently, we enhance the global filter through convolution operations to improve its effectiveness. By multiplying this global filter with the processed features, we can selectively retain critical frequency information while suppressing irrelevant details. Finally, the frequency-domain features are transformed back to the image domain using the inverse Fourier transform. The transform above is formulated as follows:

$$\tilde{\mathbf{X}}_F = \mathrm{Conv}_{1\times1}(\mathcal{A}(\mathbf{X}_F)) \cdot e^{\mathrm{Conv}_{3\times3}(\mathcal{P}(\mathbf{X}_F))} \tag{9}$$

$$\mathcal{G}(\mathbf{X}_F) = \mathrm{Conv}_{3\times3}\big(\mathrm{Maxpool}(\mathcal{A}(\mathbf{X}_F)) \cdot e^{\mathrm{Avgpool}(\mathcal{P}(\mathbf{X}_F))}\big) \tag{10}$$

$$\mathbf{X}_{\mathrm{out}} = \mathcal{F}^{-1}\big(\tilde{\mathbf{X}}_F \odot \mathcal{G}(\mathbf{X}_F)\big) \tag{11}$$

where $\tilde{\mathbf{X}}_F$ denotes the feature component resulting from the convolutional integration of amplitude and phase information, and $\mathcal{G}(\mathbf{X}_F)$ represents the global filter. The term $\mathcal{F}^{-1}$ corresponds to the inverse Fourier transform. The symbol $\odot$ denotes element-wise multiplication. The output $\mathbf{X}_{\mathrm{out}}$ is the selective decomposition feature map. The selective bifurcation process effectively captures deep frequency-domain characteristics and their interdependencies, enhancing the overall image representation.

In addition to the frequency attention mechanism, SFDM introduces a parallel convolutional branch to further enhance the frequency-domain feature extraction process. The convolutional branch operates in parallel with the frequency-domain attention mechanism, and while the attention mechanism captures the global frequency dependencies, the convolutional branch helps preserve the model's ability to capture local

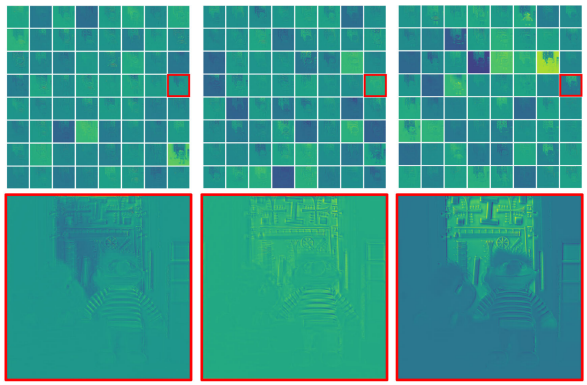

Fig. 3. (First row) visualization of 64 feature maps out of the total 180 channels after each SFRG stage in ESFS. (Second row) Zoomed-in map of the 95th channel across each stage, highlighting the evolution of features.

frequency details. This parallel structure ensures that the model can process global frequency information through the attention mechanism, while the convolutional branch supports the network in maintaining its capacity to process finer, localized frequency features.

By introducing the convolutional branch, SFDM effectively balances the need for capturing global frequency dependencies and maintaining the ability to extract local frequency features, enhancing the overall fusion process. This hybrid approach ensures that SFDM can leverage both global and local frequency-domain information, which is essential for accurate image fusion, and results in a more precise and robust fusion output.

We have also developed a frequency-domain feed-forward network, referred to as frequency multilayer perceptron (FreqMLP). As shown in Fig. 2(d), a filtering structure is constructed in the frequency domain by combining group linear layers with GELU activations.

The initial step in FreqMLP involves applying the FFT to the input signal and converting it into the frequency domain. The transformed signal then undergoes a group linear layer, which performs a grouping and linear transformation analogous to frequency decomposition. This is followed by a GELU activation, which introduces nonlinearity and acts as a nonlinear filter. The second group linear layer further refines the frequency components by re-combining and fine-tuning them. This architecture functions like a frequency-domain filter, selectively enhancing or suppressing different frequency components. As a result, it effectively manipulates the signal's frequency bands to extract nuanced features and improve the overall representation.

The use of group linear transformation in FreqMLP introduces a significant distinction from traditional fully connected linear layers. In a standard linear layer, the model applies the same linear transformation to all input features simultaneously. While this is effective in many scenarios, it may not be the most optimal method when dealing with frequency-domain data, where the characteristics of different frequency components can vary significantly. By introducing a group-wise structure, the group linear layer allows the model to apply linear transformations within smaller, more specialized groups of features, rather than treating the entire feature set uniformly. This method ensures that the frequency-domain features are processed more efficiently and effectively, allowing the model to focus on local patterns within each frequency group while preserving the global dependencies across groups.

This approach enables FreqMLP to better capture the complex nature of frequency-domain data, which typically exhibits diverse patterns across different frequency bands. The grouping mechanism allows the model to learn more specialized transformations for each frequency group, which improves the processing of both low-frequency and high-frequency components in a manner that a traditional linear layer could not achieve.

To address the differing requirements of frequency manipulation in the frequency attention and FreqMLP, we apply a separate FFT and inverse FFT in each module. In the frequency attention module, the initial FFT is used to decompose the signal into amplitude and phase components, which are then processed independently to preserve their individual frequency characteristics. In contrast, the FreqMLP module applies another round of FFT to further refine the frequency components, enabling a more sophisticated and detailed frequency-domain filtering process. This separation ensures that both modules can process frequency information in ways that best suit their respective tasks, with the inverse FFT used at the end of the FreqMLP module to transform the manipulated features back to the spatial domain.

By operating in the frequency domain, FreqMLP can leverage the inherent properties of the Fourier-transformed signal, offering a sophisticated mechanism for handling complex frequency information.

IV. Experiments

A. Datasets

We thoroughly assessed our model using three prominent MHIF benchmark datasets: the Columbia Imaging and Vision Laboratory (CAVE) dataset [59], Harvard dataset [60], and Washington DC Mall (WDCM) dataset [61]. The CAVE dataset includes 32 HSIs of indoor scenes, each measuring 512×512 pixels and comprising 31 spectral bands. These images were captured at intervals of 10 nm, covering a spectral range from 400 to 700 nm. For our evaluation, we employed the first 22 HSIs for training, designated five HSIs for validation, and reserved the remaining five HSIs for testing. The Harvard dataset features 50 HSIs that depict both indoor and outdoor scenes with various objects under natural daylight. Each HSI in this dataset consists of 31 spectral bands, ranging from 420 to 720 nm, with a resolution of 1040×1392 pixels. We used the first 34 HSIs for training, designated eight HSIs for validation, and allocated the remaining eight HSIs for testing. The HSIs from the WDCM dataset comprise 191 spectral bands, covering wavelengths from 400 to 2400 nm, with a spatial resolution of 2.5 m. The dataset has spatial dimensions of 1280×307 pixels. For validation and testing, two subimages of 128×128 pixels were extracted from the lower-left corner of the image, while

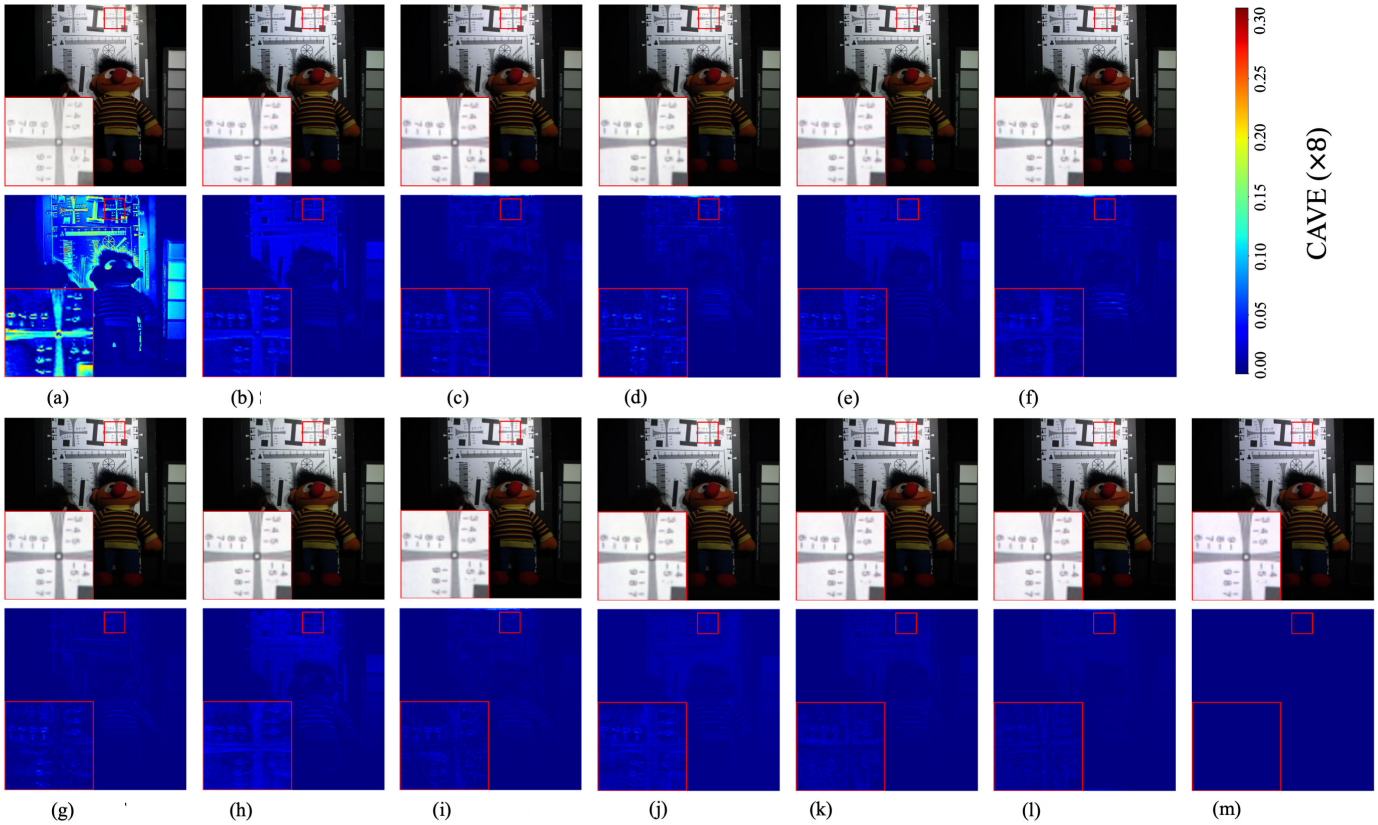

Fig. 4. (First and third rows) results from the CAVE dataset's "Chart and Stuffed Toy" scene are presented using false-color representations. Red rectangles highlight specific areas for close-up examination. (Second and fourth rows) Residual differences between the ground truth (GT) and the fusion outcomes. (a) CNMF. (b) SSR-NET. (c) TSFN. (d) MoG-DCN. (e) MSSJFL. (f) DHIF-NET. (g) PSRT. (h) MSST-NET. (i) 3DT-Net. (j) BDT. (k) DCTransformer. (l) ESFS (Ours). (m) Ground truth.

the remaining regions were designated for training the model. These datasets serve as a solid foundation for examining the effectiveness and generalization capabilities of our proposed approach.

B. Implementation Details

Before implementing our proposed method on the CAVE, Harvard, and WDCM datasets, we adhered to Wald's protocol [62] to simulate LR-HSIs and HR-MSIs from the HR-HSIs. Initially, Gaussian filtering was applied to the HR-HSIs in these datasets, resulting in blurred images. To simulate different spatial resolutions, these blurred HSIs were then downsampled with reduction factors of 4 and 8, thereby producing LR-HSIs. HR-MSIs comprising three spectral bands were generated using the spectral response matrix of the Nikon D700 camera [33], [63]. For the WDCM dataset, the HR-MSI comprising ten bands was created based on the spectral response matrix of the Sentinel-2A instrument [61]. Subsequently, we implemented our ESFS network using PyTorch 1.12.1 in a Python 3.9 environment and trained it on an NVIDIA A40 GPU. The network was optimized using the Adam optimizer to train the network for 200 epochs, with a batch size of 4. The learning rate is initialized to 0.0001 and will decay by a factor of 2 when it reaches 100, 150, 175, 190, and 195 epochs. The mean absolute error (MAE) loss function was employed to guide the optimization process.

C. Benchmark

To evaluate the performance of our ESFS network, we compared it with several MHIF methods, incorporating a diverse range of approaches. Specifically, we included CNMF [10], a matrix factorization-based method; CNN-based methods, SSR-NET [33], TSFN [12], MoG-DCN [34], MSSJFL [13], and DHIF-Net [57]; as well as Transformer-based methods including PSRT [19], MSST-NET [38], 3DT-Net [39], BDT [58], and DCTransformer [20]. These methods were selected based on their demonstrated effectiveness in addressing the MHIF problem and their prominence within the field. All deep learning approaches are trained using the same input pairs to ensure a fair comparison. Additionally, the relevant hyperparameters are chosen in alignment with those specified in the original papers.

D. Evaluation Metrics

Four quality indicators (QIs) were employed to evaluate the performance of the different methods: peak signal-to-noise ratio (PSNR), root mean square error (RMSE), relative absolute spectral error (RASE), spectral angle mapper (SAM), and error relative global dimensionless synthesis (ERGAS).

The PSNR measures the difference between the maximum signal value and the background noise in an image. A higher PSNR value indicates a lower level of noise and less distortion, thus suggesting better image quality. The formula for its

TABLE II
COMPARISON OF METHODS ON HARVARD ×4 AND HARVARD ×8 DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD, AND THE SECOND-BEST RESULTS ARE UNDERLINED. "M" DENOTES MILLION

Ratio	Methods	Harvard						
		PSNR(↑)	RMSE(↓)	RASE(↓)	SAM(↓)	ERGAS(↓)	#params	#FLOPs (10^{12})
4	CNMF IGARSS'11 [10]	45.48	0.0044	5.471	2.094	1.5739	-	-
	SSR-NET TGRS'21 [33]	50.11	0.0032	3.916	1.653	0.9790	**0.026M**	**0.027**
	TSFN TCSVT'21 [12]	50.01	0.0032	3.951	1.666	0.9878	1.074M	2.064
	MoG-DCN TIP'21 [34]	50.67	0.0030	3.619	1.527	0.9048	7.070M	17.58
	MSSJFL DependSys'21 [13]	49.49	0.0034	4.202	1.769	1.0505	20.725M	1.218
	DHIF-Net TCI'21 [57]	50.22	0.0034	3.843	1.623	0.9609	22.669M	14.02
	PSRT TGRS'23 [19]	50.22	0.0031	3.842	1.623	0.9606	0.247M	0.268
	MSST-NET IF'23 [38]	50.69	0.0030	3.616	1.527	0.9042	34.402M	11.89
	3DT-Net IF'23 [39]	50.27	0.0031	3.836	1.597	0.9281	3.455M	4.236
	BDT IJCAI'23 [58]	50.87	0.0030	3.587	1.491	0.8827	2.657M	0.468
	DCTransformer IF'24 [20]	50.96	0.0029	3.476	1.469	0.8692	8.121M	4.915
	ESFS (Ours)	**51.14**	**0.0028**	**3.234**	**1.454**	**0.8689**	7.876M	3.465
8	CNMF IGARSS'11 [10]	44.55	0.0051	6.258	2.588	1.8546	-	-
	SSR-NET TGRS'21 [33]	49.77	0.0034	4.093	1.672	0.9901	**0.026M**	**0.027**
	TSFN TCSVT'21 [12]	48.69	0.0038	4.650	1.874	1.1249	1.074M	2.064
	MoG-DCN TIP'21 [34]	50.09	0.0032	3.912	1.598	0.9472	7.070M	17.58
	MSSJFL DependSys'21 [13]	49.02	0.0037	4.502	1.856	1.1057	20.725M	1.218
	DHIF-Net TCI'21 [57]	50.05	0.0033	3.934	1.628	0.9646	22.669M	14.02
	PSRT TGRS'23 [19]	49.66	0.0034	4.123	1.700	1.0077	0.247M	0.268
	MSST-NET IF'23 [38]	50.13	0.0032	3.883	1.612	0.8572	34.402M	11.89
	3DT-Net IF'23 [39]	49.88	0.0032	4.007	1.680	0.9765	3.455M	4.236
	BDT IJCAI'23 [58]	50.38	0.0031	3.792	1.586	0.9408	2.657M	0.468
	DCTransformer IF'24 [20]	50.51	**0.0030**	3.696	1.561	0.9241	8.121M	4.915
	ESFS (Ours)	**50.74**	**0.0030**	**3.562**	**1.520**	**0.9152**	7.876M	3.465

calculation is given as follows:

$$\mathrm{PSNR}(\mathbf{X},\hat{\mathbf{X}}) = 10\lg\left(\frac{\max(\mathbf{X}_k)^2}{\frac{1}{HW}\left\|\mathbf{X}_k-\hat{\mathbf{X}}_k\right\|^2}\right) \tag{12}$$

where the function $\max(\cdot)$ denotes the maximum value of the image, $\hat{\mathbf{X}}$ represents the estimated HR-HSI, $\mathbf{X}$ denotes the ground-truth HR-HSI, with $\mathbf{X}_k$ and $\hat{\mathbf{X}}_k$ denoting the kth band of the reference and the estimated HR-HSI, respectively.

The RMSE calculates the average difference between the predicted and ground-truth values, offering a measure of the model's accuracy. The RMSE is computed as follows:

$$\mathrm{RMSE}(\mathbf{X},\hat{\mathbf{X}}) = \sqrt{\frac{\sum_{k=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(\mathbf{X}_k(i,j)-\hat{\mathbf{X}}_k(i,j)\right)^2}{HWC}} \tag{13}$$

where $\mathbf{X}_k(i,j)$ and $\hat{\mathbf{X}}_k(i,j)$ represent the element values at position (i,j) in the kth band of the reference and the estimated HR-HSI, respectively.

The RASE evaluates how well the spectral information is maintained by calculating the absolute error in each band and normalizing it to the reference image. A smaller RASE value indicates that the reconstructed image is closer to the reference image in terms of spectral accuracy. The RASE is calculated as follows:

$$\mathrm{RASE}(\mathbf{X},\hat{\mathbf{X}}) = \frac{100}{HWC}\sum_{k=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W}\frac{\left|\mathbf{X}_k(i,j)-\hat{\mathbf{X}}_k(i,j)\right|}{|\mathbf{X}_k(i,j)|} \tag{14}$$

where the numerator is the spectral error at each pixel position and the denominator is the spectral value of the reference image at the corresponding position, which is multiplied by 100 to obtain the error in percentage form.

The SAM evaluates the spectral similarity between two images, measuring the angle between the two vectors in the spectral space. The SAM is calculated as follows:

$$\mathrm{SAM}(\mathbf{X},\hat{\mathbf{X}}) = \frac{1}{HW}\sum_{k=1}^{HW}\cos^{-1}\left(\frac{\mathbf{x}_k^{\mathrm{T}}\hat{\mathbf{x}}_k}{\|\mathbf{x}_k\|_2\|\hat{\mathbf{x}}_k\|_2}\right) \tag{15}$$

where $\cos^{-1}$ denotes the arccosine function, and $\mathbf{x}_i$ and $\hat{\mathbf{x}}_i$ represent the spectra of the ith pixel of the reference and estimated HR-HSI, respectively.

The ERGAS assesses the relative error between the predicted and ground-truth images, providing a global measure of the model's performance. Thus, we have the following:

$$\mathrm{ERGAS}(\mathbf{X},\hat{\mathbf{X}}) = \frac{100}{r}\sqrt{\frac{1}{C}\sum_{k=1}^{C}\frac{\mathrm{MSE}(\mathbf{X}_k,\hat{\mathbf{X}}_k)}{\mu}} \tag{16}$$

where r is the downsampling ratio and μ denotes the mean value of each band of the estimated HR-HSI, and $\mathrm{MSE}(\mathbf{X}_k,\hat{\mathbf{X}}_k)$ represents the mean square error between the ith band of the reference and the estimated HR-HSI.

These metrics collectively offer a comprehensive evaluation of the methods' performance, encompassing aspects of image quality, spectral similarity, and overall global accuracy.

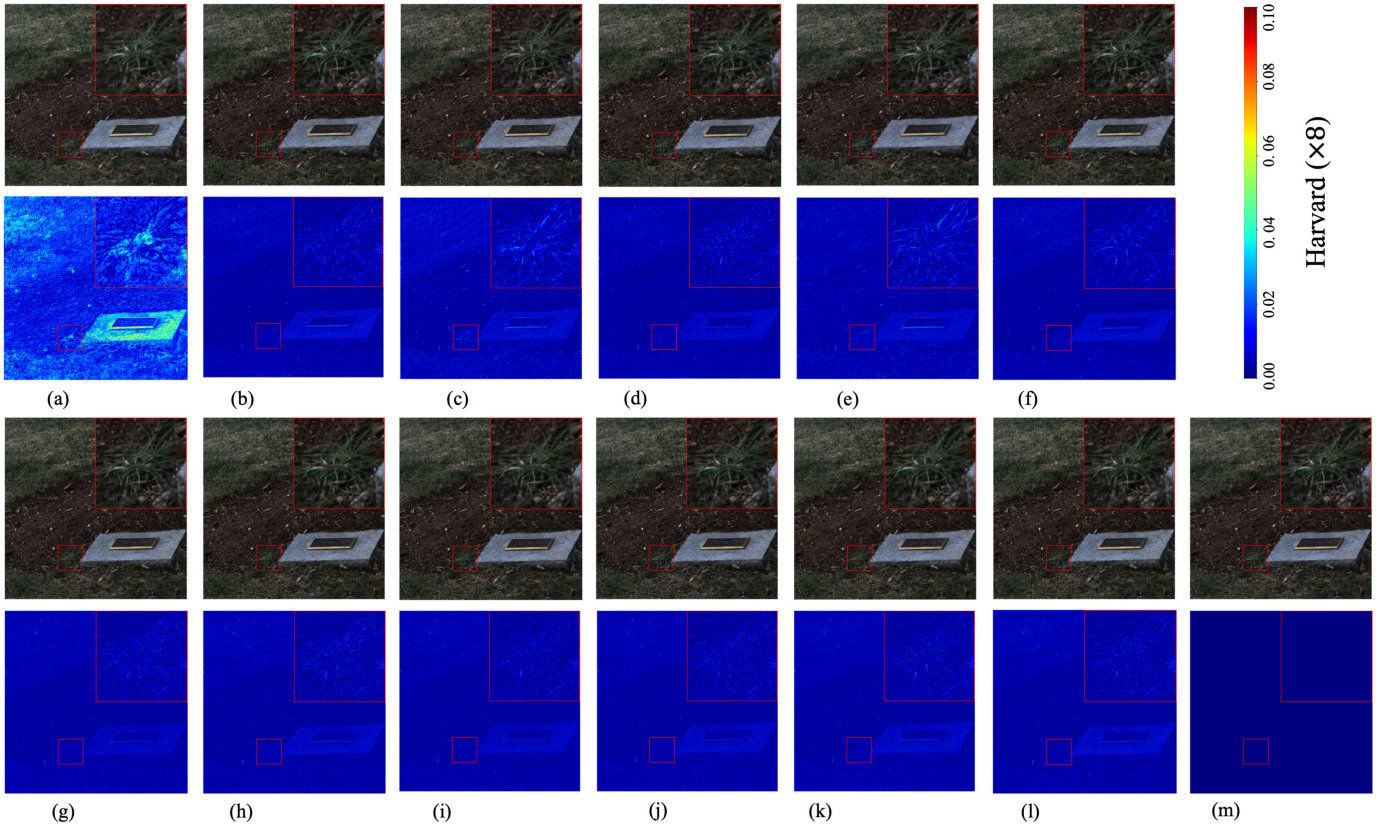

Fig. 5. (First and third rows) results from the Harvard dataset's "Gravestone" scene are presented using false-color representations. Red rectangles highlight specific areas for close-up examination. (Second and fourth rows) Residual differences between the ground truth (GT) and the fusion outcomes. (a) CNMF. (b) SSR-NET. (c) TSFN. (d) MoG-DCN. (e) MSSJFL. (f) DHIF-NET. (g) PSRT. (h) MSST-NET. (i) 3DT-Net. (j) BDT. (k) DCTransformer. (l) ESFS(Ours). (m) Ground truth.

E. Results on CAVE Dataset

We test our ESFS network on the CAVE dataset with upscaling factors of 4 and 8. The results are summarized in Table I. Our experimental results demonstrate that deep learning-based methods significantly outperform traditional model-based approaches. Furthermore, Transformer-based models generally surpass CNN-based methods. Focusing on the performance of our ESFS model, we observed that it outperforms other methods across all four QIs. Specifically, in the CAVE $\times 4$ dataset, our ESFS model achieved a PSNR improvement of 0.88, 0.54, and 0.28 dB over 3DT-Net, BDT, and DCTransformer, respectively. In the CAVE $\times 8$ dataset, it achieved with increases of 1.51, 0.71, and 0.13 dB over the same methods. Before discussing the perceptual quality, we visualize feature maps at each SFRG stage in ESFS, as shown in Fig. 3. These visualizations illustrate how the network progressively refines spatial and frequency features, contributing to the superior performance observed in our model. Specifically, at each stage, we selectively display 64 channel features, including the first 16 channels, the middle 32 channels, and the last 16 channels of the total 180 channels. Additionally, we provide a zoomed-in view of the feature maps from the 95th channel across the three stages to offer a detailed view of how features evolve as they pass through the network. These visualizations demonstrate the network's ability to capture and enhance critical spatial and frequency information as it progresses through each SFRG stage.

To demonstrate the perceptual quality of diverse methods, we provide visual comparisons in Fig. 4 for the CAVE $\times 8$ dataset. These comparisons, including detailed close-ups and error maps, clearly demonstrate that our fusion results closely resemble the ground truth, achieving the highest visual quality. In the error maps, lower brightness consistently indicates greater similarity, with our ESFS model producing reconstructions closest to the all-zero map, reflecting minimal error and accurate detail preservation.

F. Results on Harvard Dataset

Our ESFS network was also evaluated on the Harvard dataset with upscaling factors of 4 and 8, and the results are summarized in Table II. Despite the different nature of the dataset, the ESFS model consistently exhibited superior performance across all four QIs. Unlike the CAVE dataset, where the focus was primarily on indoor scenes, the Harvard dataset includes a mix of indoor and outdoor scenes, challenging the model to generalize across varied contexts. Nevertheless, our ESFS model maintained its advantage, delivering higher accuracy and better visual quality compared to the competing methods. This consistent outperformance demonstrates the robustness of our approach, not only in terms of four QIs but also in the preservation of intricate details across a wider range

TABLE III
COMPARISON OF METHODS ON WDCM ×4 AND WDCM ×8 DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD, AND THE SECOND-BEST RESULTS ARE UNDERLINED. "M" DENOTES MILLION

Ratio	Methods	WDCM						
		PSNR(↑)	RMSE(↓)	RASE(↓)	SAM(↓)	ERGAS(↓)	#params	#FLOPs (10^{12})
4	CNMF IGARSS'11 [10]	36.83	0.0089	9.784	4.843	3.3375	-	-
	SSR-NET TGRS'21 [33]	41.46	0.0084	3.991	1.730	0.9978	0.985M	0.004
	TSFN TCSVT'21 [12]	39.96	0.0100	4.742	2.043	1.1856	1.263M	0.005
	MoG-DCN TIP'21 [34]	41.26	0.0086	3.844	1.769	1.0205	19.386M	0.157
	MSSJFL DependSys'21 [13]	40.86	0.0090	4.275	1.851	1.0688	16.008M	0.009
	DHIF-Net TCI'21 [57]	44.68	0.0058	2.755	1.193	0.6889	81.243M	0.998
	PSRT TGRS'23 [19]	40.64	0.0092	4.386	1.901	1.0966	**0.341M**	**0.002**
	MSST-NET IF'23 [38]	41.87	0.0080	3.803	1.649	0.9513	147.296M	1.115
	3DT-Net IF'23 [39]	45.72	0.0051	2.443	1.058	0.6109	15.845M	0.131
	BDT IJCAI'23 [58]	45.76	0.0051	2.438	1.036	0.6029	3.905M	0.007
	DCTransformer IF'24 [20]	47.62	0.0041	1.948	0.844	0.4870	8.643M	0.079
	ESFS (Ours)	**48.29**	**0.0038**	**1.819**	**0.788**	**0.4547**	8.025M	0.056
8	CNMF IGARSS'11 [10]	34.19	0.0121	10.674	6.651	4.5247	-	-
	SSR-NET TGRS'21 [33]	40.48	0.0094	4.466	1.936	1.1166	0.985M	0.004
	TSFN TCSVT'21 [12]	39.01	0.0112	5.293	2.295	1.3233	1.263M	0.005
	MoG-DCN TIP'21 [34]	38.05	0.0125	4.082	2.561	1.4772	19.386M	0.157
	MSSJFL DependSys'21 [13]	38.59	0.0117	5.557	2.403	1.3892	16.008M	0.009
	DHIF-Net TCI'21 [57]	43.78	0.0064	3.056	1.319	0.7640	81.243M	0.998
	PSRT TGRS'23 [19]	39.59	0.0104	4.947	2.125	1.2369	**0.341M**	**0.002**
	MSST-NET IF'23 [38]	40.82	0.0090	4.292	1.854	1.0738	147.296M	1.115
	3DT-Net IF'23 [39]	44.78	0.0057	2.722	1.180	0.6806	15.845M	0.131
	BDT IJCAI'23 [58]	44.03	0.0062	3.125	1.272	0.7355	3.905M	0.007
	DCTransformer IF'24 [20]	46.15	0.0048	2.305	0.999	0.5764	8.643M	0.079
	ESFS (Ours)	**46.76**	**0.0045**	**2.168**	**0.937**	**0.5420**	8.025M	0.056

of scenes. The visual comparison in Fig. 5 further supports these findings, showing that our ESFS model provides superior reconstruction fidelity, especially in challenging regions with complex textures and spectral variations. This reinforces the effectiveness of our method in delivering high-quality image fusion results, establishing it as a leading approach in the field.

G. Results on WDCM Dataset

Finally, we evaluated our ESFS network on the WDCM dataset with upscaling factors of 4 and 8, with results summarized in Table III. The ESFS model consistently outperformed the other methods across all four QIs, demonstrating its robustness and generalization capabilities. The WDCM dataset, with its unique spectral range and spatial resolution, presents a distinct challenge compared to the CAVE and Harvard datasets. Despite these differences, our ESFS model maintained its superior performance, highlighting its adaptability and effectiveness in handling diverse datasets. The visual comparison in Fig. 6 further illustrates the exceptional quality of our fusion results, showcasing the network's ability to accurately reconstruct complex scenes with high fidelity. These visual comparisons, coupled with the quantitative results, underscore the superior performance of our ESFS model, establishing it as a leading approach for MHIF tasks across a wide range of datasets.

H. Ablation Studies

1) Effectiveness of ESFS: To evaluate the effectiveness of our ESFS network, we conducted an ablation study on the CAVE ×4 dataset to assess the impact of the CSAM and SFDM modules. Additionally, we further analyzed the contribution of the FreqMLP, which is a sub-module within SFDM. By systematically removing or isolating these components, we aimed to quantify their individual contributions to the overall performance of the network. This analysis allowed us to gain a deeper understanding of how each component enhances the fusion quality, both independently and synergistically. Table IV highlights the significant role that both CSAM and SFDM play in improving the fusion quality. Their inclusion leads to superior performance, validating the design choices and justifying their integration into the ESFS network.

As shown in Table IV, the inclusion of CSAM and SFDM significantly improves the performance metrics across all quality indexes. Specifically, when the CSAM module is removed, a notable drop in spectral and spatial consistency is observed, highlighting its critical role in capturing global and local contextual information. Similarly, removing the SFDM module, including its FreqMLP sub-module, results in a marked decline in overall fusion quality, especially in spectral fidelity, as the network struggles to effectively exploit frequency-based dependencies.

Further analysis of FreqMLP reveals its distinct contribution within SFDM. The absence of FreqMLP leads to diminished performance, particularly in high-frequency reconstruction tasks, underscoring its importance in refining frequency-specific details. These results validate the modular design of the ESFS network, demonstrating that the combination of CSAM and SFDM, with its embedded FreqMLP, provides

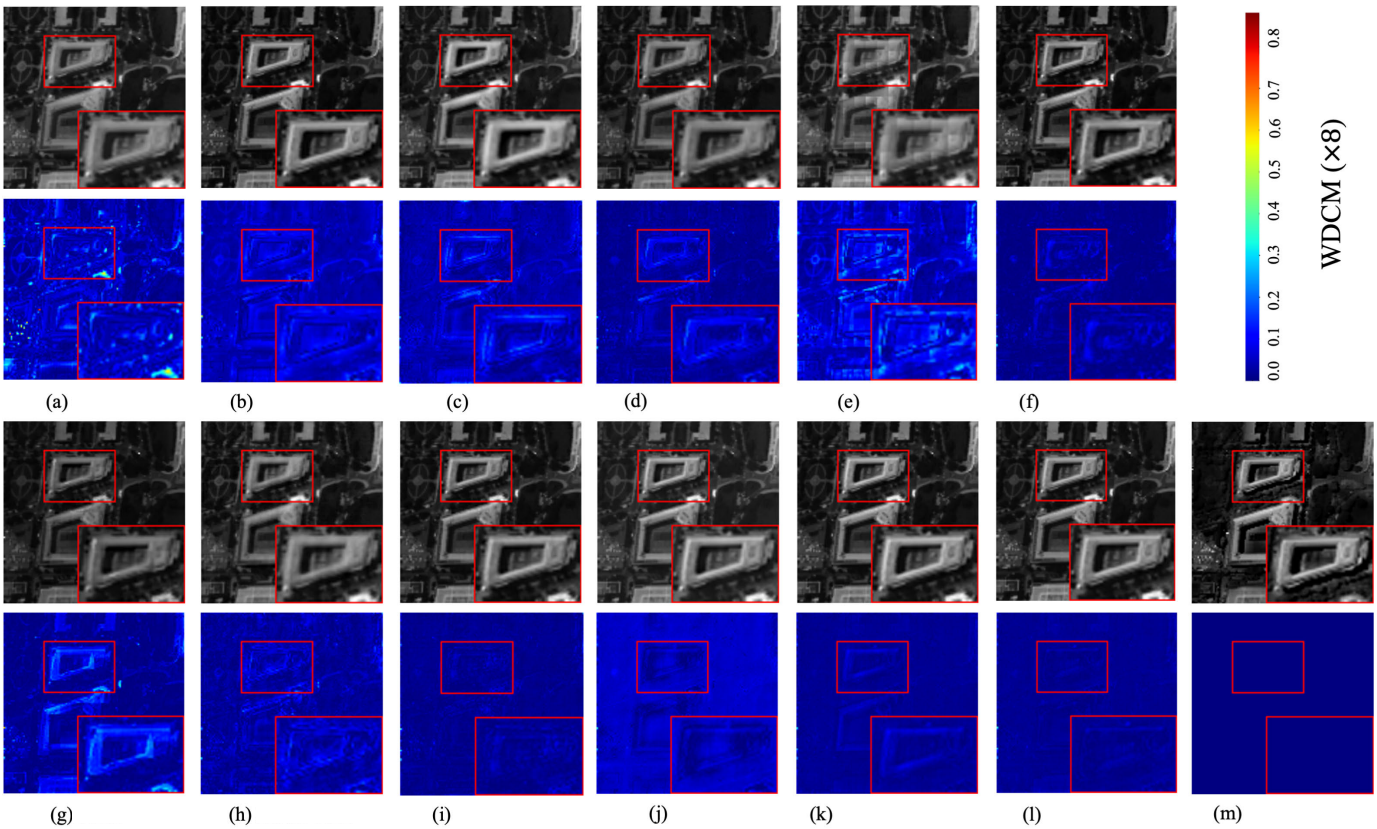

Fig. 6. (First and third rows) results from the WDCM datasets are presented using gray representations. Red rectangles highlight specific areas for close-up examination. (Second and fourth rows) Residual differences between the ground truth (GT) and the fusion outcomes. (a) CNMF. (b) SSR-NET. (c) TSFN. (d) MoG-DCN. (e) MSSJFL. (f) DHIF-NET. (g) PSRT. (h) MSST-NET. (i) 3DT-Net. (j) BDT. (k) DCTransformer. (l) ESFS(Ours). (m) Ground truth.

TABLE IV

ABLATION STUDY ON THE EFFECTIVENESS OF VARIOUS COMPONENTS IN THE ESFS NETWORK, EVALUATED ON THE CAVE ×4 DATASET

CSAM (w/o CNN)	SFDM (w/o CNN)	CSAM	SFDM	FreqMLP	Dense Connection	PSNR(↑)	RMSE(↓)	RASE(↓)	SAM(↓)	ERGAS(↓)	#params	#FLOPs (10^{12})
✗	✗	✓	✗	✗	✓	51.87	0.0028	2.662	0.843	0.6657	6.190M	2.630
✗	✗	✗	✓	✗	✓	51.49	0.0028	2.763	0.876	0.6907	5.486M	**2.101**
✗	✗	✗	✓	✓	✓	51.73	0.0028	2.690	0.852	0.6725	5.684M	2.257
✓	✓	✗	✗	✓	✓	51.55	0.0029	2.762	0.874	0.6906	6.898M	2.765
✗	✗	✓	✓	✓	✗	51.01	0.0030	2.932	0.930	0.7532	**5.052M**	2.689
✗	✗	✓	✓	✓	✓	**52.23**	**0.0026**	**2.513**	**0.812**	**0.6378**	7.876M	3.493

complementary benefits that collectively enhance the fusion quality.

In addition to the removal of individual modules, we also conducted experiments where we tested combinations of modules, such as CSAM w/o CNN and SFDM w/o CNN, to evaluate the impact of convolutional branches. The results further emphasize the importance of these components, showing that the convolutional branches are crucial for capturing fine-grained spatial information. Moreover, we also investigated the effect of dense connections, which were shown to enhance feature propagation and improve model robustness, and help mitigate the performance drop, demonstrating their role in facilitating more efficient information flow within the network. These experiments provide a more comprehensive understanding of how the modules and their components interact, confirming the necessity of their integration in the ESFS network. The ablation study confirms the robustness of our approach and highlights the necessity of integrating these components to achieve state-of-the-art (SOTA) performance.

TABLE V

ABLATION STUDY ON THE IMPACT OF DIFFERENT CONDENSED WINDOW SIZES m IN CW-MSA, EVALUATED ON THE CAVE ×4 DATASET

Config.	PSNR(↑)	RMSE(↓)	RASE(↓)	SAM(↓)	ERGAS(↓)	#params	#FLOPs (10^{12})
$m = M$ **(W-MSA)**	51.56	0.0028	2.712	0.882	0.6970	8.354M	3.643
$m = \frac{M}{2}$	51.73	0.0028	2.682	0.853	0.6905	8.128M	3.565
$m = \frac{M}{4}$ **(Ours)**	**52.23**	**0.0026**	**2.513**	**0.812**	**0.6378**	7.876M	3.493
$m = \frac{M}{8}$	51.55	0.0029	2.765	0.876	0.6934	**7.677M**	**3.354**

2) Impact of CW-MSA on Efficiency and Performance: To further evaluate the efficiency and performance of the CW-MSA component, we tested different values of the

TABLE VI
ABLATION STUDY BY REPLACING THE CSAM MODULE IN THE ESFS NETWORK WITH DIFFERENT TRANSFORMER-BASED ALTERNATIVES, EVALUATED ON THE CAVE ×4 DATASET

Method	PSNR(↑)	RMSE(↓)	RASE(↓)	SAM(↓)	ERGAS(↓)	#Params	#FLOPs (10^{12})
ESFS (w/ Restormer) [64]	51.82	0.0028	2.671	0.847	0.6678	7.965M	3.683
ESFS (w/ DRSformer) [65]	51.90	0.0027	2.642	0.837	0.6607	8.168M	4.189
ESFS (w/ CSAM)	**52.23**	**0.0026**	**2.513**	**0.812**	**0.6378**	**7.876M**	**3.493**

condensed window size m on the CAVE ×4 dataset. Notably, when $m = M$, CW-MSA is equivalent to the original W-MSA. Our experiments identified $m = (M/4)$ as the optimal configuration, which effectively balances computational complexity reduction with superior fusion quality. However, when the number of iterations increases, the value of m becomes too small, which leads to excessive compression of the window's spatial features and results in the loss of critical spatial information. This degradation in spatial representation results in reduced fusion performance. The results, as shown in Table V, reveal that CW-MSA effectively reduces computational demands and achieves superior fusion quality compared to the original W-MSA.

The improved performance highlights CW-MSA's ability to balance global contextual understanding and local feature refinement, making it a vital enhancement to the network. Moreover, the efficiency gain demonstrates its practicality for real-world applications requiring high-speed processing.

However, the study also indicates that although CW-MSA contributes to a reduction in parameters, its effectiveness in this regard is substantially constrained by the recurrent utilization of convolutional branches within each SFRG module, which mitigates the overall parameter savings. These convolutional components, while critical for extracting deep spatial and frequency features, add to the overall network complexity, limiting the net reduction in parameters. Despite this, the trade-off between efficiency and accuracy firmly justifies the inclusion of CW-MSA, as it contributes significantly to the optimized performance of the ESFS network.

3) Impact of Replacing CSAM With Existing Transformer-Based Modules: To assess the effectiveness of the CSAM module in our ESFS network, we conducted an ablation study by replacing CSAM with two representative Transformer-based blocks: Restormer [64] and DRSformer [65]. These modules are known for their capabilities in capturing long-range dependencies and have demonstrated strong performance in various image restoration tasks. The results of this substitution experiment on the CAVE ×4 dataset are summarized in Table VI.

The findings show that CSAM achieves better performance while maintaining lower computational cost. While Restormer and DRSformer are designed for general restoration tasks and emphasize spatial context modeling, they do not explicitly address the trade-off between modeling precision and efficiency under the constraints of multisource fusion. In contrast, CSAM uses a fixed-ratio compressed attention design to capture structure-aware features critical to MHIF, leading to more accurate alignment between LR-HSIs and HR-MSIs. This task-specific adaptation enables CSAM to outperform these alternatives in both accuracy and efficiency.

V. CONCLUSION

In this study, we introduce an ESFS network, a novel framework designed to address the challenges in MHIF. ESFS leverages a synergistic integration of spatial and frequency-domain processing to enhance the fusion process, capturing both detailed spatial features and critical frequency information. By employing CW-MSA mechanisms and selective frequency transforms, the proposed network effectively preserves spectral fidelity while improving the overall representation of fused images. Our experimental results demonstrate that ESFS outperforms existing methods, offering a significant advancement in MHIF and contributing to more robust and reliable image analysis in remote sensing and computer vision applications. As the proposed method has not yet been validated on real LR-HSI and HR-MSI datasets, there are certain constraints on assessing its overall effectiveness. Therefore, future research should focus on conducting comprehensive experiments with real-world datasets to further refine and enhance the method's performance.

REFERENCES

[1] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2020.
[2] N. Li, S. Jiang, J. Xue, S. Ye, and S. Jia, "Texture-aware self-attention model for hyperspectral tree species classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5502215.
[3] B. Tu, W. He, Q. Li, Y. Peng, and A. Plaza, "A new context-aware framework for defending against adversarial attacks in hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5505114.
[4] Z. Li, W. An, G. Guo, L. Wang, Y. Wang, and Z. Lin, "SpecDETR: A transformer-based hyperspectral point object detection network," 2024, *arXiv:2405.10148*.
[5] X. Wu, D. Hong, and J. Chanussot, "UIU-net: U-net in U-net for infrared small object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 364–376, 2023.
[6] H. Van Nguyen, A. Banerjee, and R. Chellappa, "Tracking via object reflectance using a hyperspectral video camera," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops*, Jun. 2010, pp. 44–51.
[7] Z. Zhou et al., "Swin-spectral transformer for cholangiocarcinoma hyperspectral image segmentation," in *Proc. 14th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2021, pp. 1–6.
[8] X. Zhang, N. Yokoya, X. Gu, Q. Tian, and L. Bruzzone, "Local-to-global cross-modal attention-aware fusion for HSI-X semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5531817.
[9] F. Ye, Z. Wu, Y. Xu, H. Liu, and Z. Wei, "Bayesian hyperspectral image super-resolution in the presence of spectral variability," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5545613.
[10] 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.*, Jul. 2011, pp. 1779–1782.
[11] 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.
[12] X. Wang, J. Chen, Q. Wei, and C. Richard, "Hyperspectral image super-resolution via deep prior regularization with parameter estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 1708–1723, Apr. 2022.

[13] Z. Min, Y. Wang, and S. Jia, "Multiscale spatial–spectral joint feature learning for multispectral and hyperspectral image fusion," in *Proc. IEEE 23rd Int. Conf. High Perform. Comput. Commun.; 7th Int. Conf. Data Sci. Syst.; 19th Int. Conf. Smart City; 7th Int. Conf. Dependability Sensor, Cloud Big Data Syst. Appl. (HPCC/DSS/SmartCity/DependSys)*, Dec. 2021, pp. 1265–1270.

[14] J. Li, K. Zheng, W. Liu, Z. Li, H. Yu, and L. Ni, "Model-guided Coarse-to-Fine fusion network for unsupervised hyperspectral image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[15] 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.

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

[17] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[18] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.

[19] 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.

[20] 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.

[21] M. Zhou et al., "A general spatial-frequency learning framework for multimodal image fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 7, pp. 5281–5298, Jul. 2025.

[22] 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.

[23] W. Dong et al., "Hyperspectral image super-resolution via non-negative structured sparse representation," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2337–2352, May 2016.

[24] X. Li, Y. Zhang, Z. Ge, G. Cao, H. Shi, and P. Fu, "Adaptive non-negative sparse representation for hyperspectral image super-resolution," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4267–4283, 2021.

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

[26] 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.

[27] 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.

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

[29] R. Dian, S. Li, and L. Fang, "Learning a low tensor-train rank representation for hyperspectral image super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2672–2683, Sep. 2019.

[30] F. Ye, Z. Wu, X. Jia, J. Chanussot, Y. Xu, and Z. Wei, "Bayesian nonlocal patch tensor factorization for hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 32, pp. 5877–5892, 2023.

[31] C. Wang, Y. Liu, X. Bai, W. Tang, P. Lei, and J. Zhou, "Deep residual convolutional neural network for hyperspectral image super-resolution," in *Proc. 9th Int. Conf. Image Graphics*, Shanghai, China, Jan. 2017, pp. 370–380.

[32] 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.

[33] 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.

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

[35] R. Ran, L.-J. Deng, T.-X. Jiang, J.-F. Hu, J. Chanussot, and G. Vivone, "GuidedNet: A general CNN fusion framework via high-resolution guidance for hyperspectral image super-resolution," *IEEE Trans. Cybern.*, vol. 53, no. 7, pp. 4148–4161, Jul. 2023.

[36] 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.

[37] 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.

[38] 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, Aug. 2023.

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

[40] Y. Shang, J. Liu, J. Zhang, and Z. Wu, "MFT-GAN: A multi-scale feature-guided transformer network for unsupervised hyperspectral pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5518516.

[41] 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.

[42] L. Sun et al., "MDC-FusFormer: Multiscale deep cross-fusion transformer network for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5528914.

[43] X. Cao, Y. Lian, J. Li, K. Wang, and C. Ma, "Unsupervised multi-level spatio-spectral fusion transformer for hyperspectral image super-resolution," *Opt. Laser Technol.*, vol. 176, Sep. 2024, Art. no. 111032.

[44] S. Chen, L. Zhang, and L. Zhang, "Cyclic cross-modality interaction for hyperspectral and multispectral image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 1, pp. 741–753, Jan. 2025.

[45] J. Lee-Thorp, J. Ainslie, I. Eckstein, and S. Ontanon, "FNet: Mixing tokens with Fourier transforms," 2021, *arXiv:2105.03824*.

[46] N. Sevim, E. Ozan Özyedek, F. Öahinuç, and A. Koç, "Fast-FNet: Accelerating transformer encoder models via efficient Fourier layers," 2022, *arXiv:2209.12816*.

[47] Y. Rao, W. Zhao, Z. Zhu, J. Lu, and J. Zhou, "Global filter networks for image classification," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2021, pp. 980–993.

[48] N. Zheng, M. Zhou, J. Huang, and F. Zhao, "Frequency integration and spatial compensation network for infrared and visible image fusion," *Inf. Fusion*, vol. 109, Sep. 2024, Art. no. 102359.

[49] K. Hu, Q. Zhang, M. Yuan, and Y. Zhang, "SFDFusion: An efficient spatial-frequency domain fusion network for infrared and visible image fusion," in *Proc. ECAI*, Oct. 2024, pp. 482–489.

[50] J. Tan et al., "Revisiting spatial-frequency information integration from a hierarchical perspective for panchromatic and multi-spectral image fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 25922–25931.

[51] Y.-J. Liang, Z. Cao, L.-J. Deng, and X. Wu, "Fourier-enhanced implicit neural fusion network for multispectral and hyperspectral image fusion," 2024, *arXiv:2404.15174*.

[52] L. Kong, J. Dong, J. Ge, M. Li, and J. Pan, "Efficient frequency domain-based transformers for high-quality image deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5886–5895.

[53] S. Paul, S. Kumawat, A. Gupta, and D. Mishra, "F2former: When fractional Fourier meets deep Wiener deconvolution and selective frequency transformer for image deblurring," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Feb. 2025, pp. 9457–9467.

[54] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 4700–4708.

[55] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[56] J. G. Proakis, *Digital Signal Processing: Principles, Algorithms, and Applications, 4/E*. Upper Saddle River, NJ, USA: Pearson Education, 2007.

[57] T. Huang, W. Dong, J. Wu, L. Li, X. Li, and G. Shi, "Deep hyperspectral image fusion network with iterative spatio-spectral regularization," *IEEE Trans. Comput. Imag.*, vol. 8, pp. 201–214, 2022.

[58] S. Deng, L.-J. Deng, X. Wu, R. Ran, and R. Wen, "Bidirectional dilation transformer for multispectral and hyperspectral image fusion," in *Proc. 32nd Int. Joint Conf. Artif. Intell.*, Aug. 2023, pp. 3633–3641.

[59] 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.

[60] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 193–200.

[61] 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.

[62] 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.

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

[64] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5728–5739.

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

Meng Xu (Member, IEEE) received the B.S. and M.E. degrees in electrical engineering from the Ocean University of China, Qingdao, China, in 2011 and 2013, respectively, and the Ph.D. degree from the University of New South Wales, Canberra, ACT, Australia, in 2017.

She is currently an Associate Professor with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. Her research interests include cloud removal and remote sensing image processing.

Ziqian Mo (Graduate Student Member, IEEE) received the B.S. degree in computer science from Hubei University of Technology, Wuhan, China, in 2023. He is currently pursuing the M.S. degree in artificial intelligence at Shenzhen University, Shenzhen, China.

His research interests include hyperspectral image processing and image super-resolution.

Xiyou Fu (Member, IEEE) received the B.S. degree from Wuhan University, Wuhan, China, in 2012, and the M.S. and Ph.D. degrees from the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China, in 2015 and 2019, respectively.

He is currently an Assistant Professor with Shenzhen University, Shenzhen, China. His research interests include hyperspectral image restoration, anomaly detection, and super-resolution.

Sen Jia (Senior Member, IEEE) received the B.S. and Ph.D. degrees from the College of Computer Science, Zhejiang University, Hangzhou, China, in 2002 and 2007, respectively. Since 2008, he has been with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China, where he is currently a Full Professor. His research interests include hyperspectral image processing, signal and image processing, and machine learning.