MMAPP: Multibranch and Multiscale Adaptive Progressive Pyramid Network for Multispectral Image Pansharpening

Zhiqi Zhang, Chuang Liu, Lu Wei, and Shao Xiang

***Abstract*—Pansharpening is the process of integrating two heterogeneous remote sensing images to obtain high-resolution multispectral images, which is crucial for downstream tasks. Existing methods utilizing advanced deep-learning techniques are able to achieve good sharpening results. However, the heterogeneity between diverse source images is not sufficiently considered, which in turn results in distortions in the sharpening results. Addressing this gap, we have developed a multibranch pyramid structure, which can build bridges between diverse source images at various scales. It contains three distinct branches, including the PAN branch, the MS branch, and the fusion branch, which efficiently and seamlessly integrates the data flow in distinct branches by means of the pyramid structure. Furthermore, in order to retain more advantageous information, we have developed a specialized adaptive extraction and integration module (AEIM) for each branch, namely, the texture shrinkage adaptive module for the PAN branch, the spectral information consistency module for the MS branch, and the adaptive fusion module for the fusion branch. These AEIMs are specifically designed to cater to diverse sources and distinct stages of the pansharpening process. The adaptive weights they generate can be used to extract and fuse more advantageous information. Ultimately, high-fidelity sharpening outcomes are obtained by minimizing the reconstruction errors at various scales in distinct branches. Extensive experiments show that our methodology surpasses that of representative advanced methods, while maintaining a high level of efficiency. All implementations will be published at MMAPP.**

***Index Terms*—Image fusion, multimodal data, multispectral image, panchromatic image (PANI), remote sensing.**

I. Introduction

REMOTE sensing images (RSIs) collected by different imaging sensors have been widely utilized in various fields, such as national defense construction [1], disaster warning [2], and geological survey [3]. The limitations of the physical environment and hardware budget prevent a single sensor from simultaneously acquiring RSIs with fine-grained spatial texture and high-fidelity spectral information. Modern high-resolution optical satellites acquire high spatial resolution (HR) panchromatic image (PANI) and low spatial resolution (LR) multispectral image (MSI) by configuring different sensors. In light of this, the pansharpening technique has been developed as a cost-effective means for integrating information in the above two types of RSIs. Consequently, HR MSIs can be generated, which are crucial for downstream tasks and real-world applications.

The key to pansharpening is to simultaneously retain adequate spatial information in PANI and spectral information in MSI. However, PANI and MSI are captured by different sensors, and there are inherent scale disparities between them. This heterogeneity presents a significant challenge in balancing spatial and spectral preservation. So far, pansharpening methods can be broadly categorized into traditional and deep-learning (DL) methods [4]. In the majority of traditional methods [5], [6], [7], linear transformation is employed to manually extract features from diverse source images, offering the advantages of high efficiency and independence from hardware devices. Nonetheless, the limited representational capabilities of extracted features render it difficult to represent intricate relationships between heterogeneous data, resulting in distortions in the fusion outcomes. Recent years have witnessed a surge in the research on DL, as it has demonstrated great potential in representing the correlations of diverse source data [8]. As depicted in Fig. 1, the design principles of DL methods can be categorized into three distinct categories: single-branch structures incorporating the concept of super-resolution (SR) [9], [10], [11], [12], double-branch structures that accommodate diverse source characteristic [13], [14], and UNet-based multiscale structures that accommodate scale disparities [15], [16], [17]. Despite the fact that these methods surpass traditional methods in quantitative outcomes, the sharpened images still suffer from evident spatial or spectral distortion. There is, on the one hand, a fundamental difference between SR and pansharpening. The former does not necessitate any additional images to improve the spatial resolution, whereas the latter necessitates the utilization of spatial information within the PANI. On the other hand, some DL methods directly migrate advance modules or network structures from other domains to the pansharpening domain. This style has the potential to

Received 23 July 2024; revised 17 September 2024 and 9 October 2024; accepted 31 October 2024. Date of publication 1 November 2024; date of current version 15 November 2024. This work was supported by the National Key R&D Program of China under Grant 2022YFB3902800. *(Corresponding author: Shao Xiang.)*

Zhiqi Zhang is with the School of Computer Science, Hubei University of Technology, Wuhan 430068, China, and also with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: zzq540@hbut.edu.cn).

Chuang Liu is with the School of Computer Science, Hubei University of Technology, Wuhan 430068, China (e-mail: liuchuang@hbut.edu.cn).

Lu Wei is with the School of Information Science and Engineering, Wuchang Shouyi University, Wuhan 430064, China (e-mail: weilu@wsyu.edu.cn).

Shao Xiang is with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: xiangshao@whu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3490755

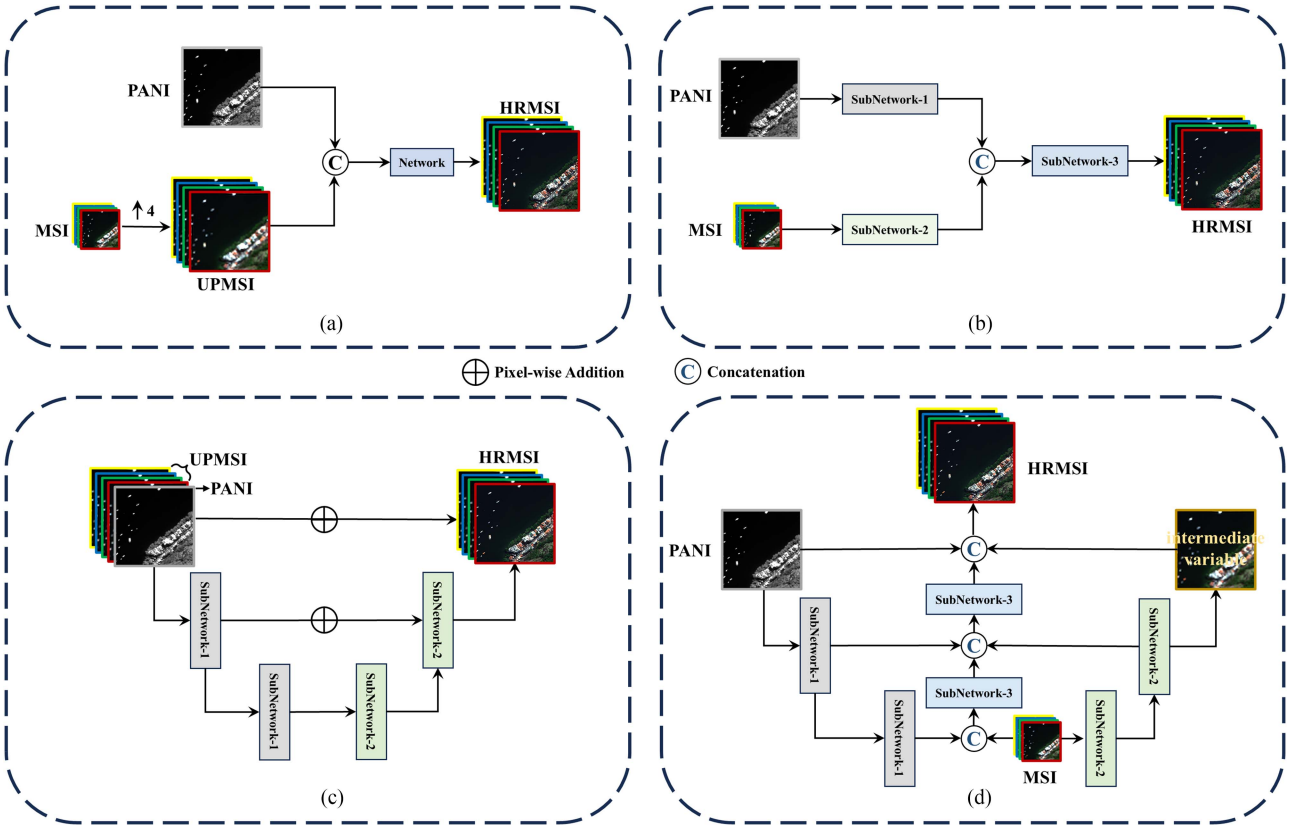

Fig. 1. Design principles of DL-based pansharpening methods. (a) Single-branch structure incorporating the concept of super-resolution. (b) Dual-branch structure. (c) UNet-based multiscale structure. (d) Proposed multibranch pyramid structure.

enhance sharpening performance, however, it lacks targeted research for pansharpening tasks. The heterogeneity of PANI and MSI is neglected by these techniques, resulting in spatial and spectral distortions in fusion outcomes. Furthermore, these incorrect spatial textures and unreasonable spectral distributions contravene the original purpose of RSIs, which are intended to reflect actual geographical characteristics. For the purposes of this article, it is imperative to address three primary concerns: 1) how can we effectively exploit the scale disparities and heterogeneity of diverse source images? 2) how can we enhance the preservation of spatial textures and spectral information in the input image pairs? and 3) how can we obtain sharpened images with higher fidelity and realism?

Regarding the issue of scale inconsistency, it is natural to associate that the bridge between PANI and MSI can be smoothly built using a multiscale structure. This gap can be alleviated with the UNet-based multiscale structure. It first samples two images of different scales directly to the same scale, then performs a concatenation operation in the channel dimension, and finally feeds into the UNet-based multiscale structure. The direct sampling and concatenating procedures overlook the heterogeneity between diverse source images, resulting in distorted fusion results. The dual-branch structure can alleviate this issue; however, it fails to utilize the disparities in scale between diverse source images. Combining a multiscale structure with a dual-branch structure may offer a potential solution. However, the mere amalgamation of the two can be rather blunt, and we require a state that efficiently and seamlessly integrate diverse source features. Therefore, we have developed multibranch pyramid structure to build bridges between diverse source images at various scales. It contains three branches, including the PAN branch, the MS branch, and the fusion branch, which efficiently and seamlessly integrates the data flow in distinct branches by means of the pyramid structure. Comparative to the other three types of sharpened structures, the proposed multibranch pyramid structure is capable of encompassing the heterogeneity and scale differences of PANI and MSI by employing distinct branches to extract and integrate information at various scales from diverse sources. Meanwhile, it establishes the transfer of information at the same scales via long-distance connections, thereby preserving more information.

On this basis, it is imperative to preserve as much of the advantageous information in diverse source images as possible, specifically spatial texture in PANI and spectral information in MSI. Conventional convolution employs context-agnostic convolution kernels, which assign identical weight to distinct local regions, resulting in the loss of crucial texture information. Dynamic convolution is an effective technique that aligns with the notion of attention mechanism, enabling the generation of distinct weights according to diverse inputs. Compared to the attention mechanism, dynamic convolution is more advantageous in terms of efficiency while retaining more information. Inspired by dynamic convolution, we design a specialized adaptive extraction and integration module (AEIM) for each branch, i.e.,

texture shrinkage adaptive module (TSAM) for PAN branch, spectral information consistency module (SICM) for MS branch, and adaptive fusion module (AFM) for fusion branch. These AEIMs are specifically designed to cater to diverse sources and distinct stages of the pansharpening process. They possess the capability to generate adaptive weights based on the inputs, thereby enabling adaptive attention to distinct local regions and preserving more spatial textures and spectral information.

Likewise, in the fusion branch, we have designed a multistage reconstruction process to restore the HRMSI to its original scale. We calculate reconstruction errors at different scales in both the MS and fusion branches, and subsequently weight these errors to construct multiscale reconstruction constraints. This process is intended to ensure that of more realistic information is maintained. The following contribution can be drawn.

1) Given the diverse sources and scale inconsistencies of PANI and MSI, we propose an efficient multibranch pyramid structure for pansharpening, comprising PAN, MS, and fusion branches. Each branch employs a pyramid structure in order to extract multiple scale features from diverse source images. Due to the highly cohesive design of the three branches, the whole model is capable of achieving high-fidelity sharpening outcomes while maintaining a high level of efficiency. On the one hand, this structure considers the heterogeneity and scale disparities between diverse source data. It is capable of extracting and integrating information of various scales from diverse sources, while establishing the transmission between information of the same scale, thereby achieving the high-fidelity fusion. On the other hand, this structure does not rely on the accumulation of numerous high-complexity modules, ensuring a high level of efficiency.
2) To preserve adequate spatial textures in PANI and spectral information in MSI while effectively balancing them, we have developed three task-specific AEIMs for three distinct branches. These modules can generate adaptive weights based on the characteristics of diverse source images, ensuring that advantageous information can be effectively extracted and integrated.
3) The multiscale loss function is obtained by weighing the reconstruction errors of various scales in distinct branches. This way, more advantageous information can be preserved, leading to a more realistic spatial texture and spectral distribution of the sharpened images.

The rest of this article is organized as follows. Section II outlines the work related to the proposed MMAPP. In Section III, we present a detailed description of the proposed MMAPP. In Section IV, we show the experimental outcomes and conduct a thorough analysis of them. Section V discusses the proposed methodology. Finally, Section VI concludes this article.

II. Related Works

A. Traditional Pansharpening Methods

The traditional methods for pansharpening include component substitution (CS)-based, multiresolution analysis (MRA)-based, and variational optimization (VO)-based methods [13]. The CS methods presuppose that spatial and spectral information in MSI can be separated, which means that spatial details in MSI are directly substituted by PANI. During the process of substitution, these methods produce localized differences due to spectral mismatch, resulting in spectral distortion. Therefore, various improved methods have been proposed, including BT-H [5], GSA [6], and PRACS [7]. Although the improved methods can yield superior rendering effects, spectral distortion still occurs due to the inherent flaws of the design principle of CS. The MRA methods employ the MRA tool to extract spatial features from the PANI, which are then injected into the LRMSI according to specified guidelines. In general, MRA methods outperform the CS, despite the presence of spatial distortion. The MRA methods predominantly encompass AWLP [18] and MTF-GLP [19]. In the VO methods, the pansharpening task is perceived as an ill-posed problem [20], [21], [22]. Palsson et al. [23] designed an observation-model-based sharpening method utilizing prior knowledge and regularized it using the total variation technique, which yielded satisfactory visual outcomes despite its high computational complexity. Based on this, Palsson et al. [24] designed a general-model-based fusion method for two-by-two fusion between PANI, MSI, and hyperspectral image using principle component analysis and wavelets. Vicinanza et al. [25] considered the pansharpening task to be a signal reconstruction problem, which estimates the missing details in MSI from PANI by means of a sparse representation technique. The method was able to achieve excellent quantitative scores on different data, however, it demonstrated poor subjective performance. In general, the sharpening outcomes obtained by VO methods are visually superior, however, they necessitate the artificial setting of hyperparameters and exhibit limited generalization capability in complex scenes. Moreover, guided by the aforementioned classical techniques, certain methods, such as VOGTNet [26], LNM-PS [27], and LRTCP [28], are capable of achieving high-fidelity sharpening outcomes in diverse scenes by introducing nonlinear constraints.

B. DL-Based Pansharpening Methods

Recently, DL has been extensively utilized in the field of pansharpening due to its inherent capability of implicitly acquiring prior knowledge from vast quantities of data. A pansharpening neural network (PNN) [9], influenced by SRCNN, is proposed. Despite the fact that PNN only contains three convolutional layers, it demonstrated a notable sharpened effect, laying the foundation for subsequent DL-based pansharpening methods. Similarly, Cai and Huang [10] proposed a pansharpening network under the influence of the SR processing. To preserve adequate fine-grained spatial details, Yang et al. [13] extracted high-frequency feature from PANI, and retained high-frequency details utilizing the residual structure. Zhang et al. [14] constructed a bidirectional pyramidal network (BDPN) using residual blocks, which is able to fuse the extracted spatial and spectral features through pairwise interactions. The BDPN achieved an excellent sharpening performance, but the stacking of numerous residual blocks results in a significant expansion of the model parameters. Considering the disparities of diverse

source images, Chen et al. [29] developed a novel two-branch structure using guided filtering to extract the high-frequency textures of PANI and MSI, respectively. In order to establish the global spatial–spectral dependence, several methods employed improved attention modules to construct the network, leading to satisfactory sharpened results, such as TANI [30], TRRNet [31], and RSANet [32]. Although these methods can achieve satisfactory results, the stacking of numerous convolutional and attention layers in these methods has a significant impact on the efficacy of pansharpening. Considering the correlation between PAN and MSI, some methods employ a UNet-based multiscale structure to efficiently utilize the information at different scales, such as MUCNN [15] and MMFN [16]. In contrast to the motivation of previous article, we propose a novel multibranch pyramid structure. The design principle of this structure is founded on a comprehensive analysis of the pansharpening task, taking into account the heterogeneity and disparities in scale between diverse source images.

C. Dynamic Convolution

The objects present in distinct local regions in RSIs frequently exhibit distinct characteristics. Due to its context-agnostic, it is challenging to extract fine-grained spatial and spectral information using standard convolution. Dynamic convolution, however, is capable of generating kernels with different weights based on different inputs, which can adapt to different local regions of the same image and thus retain more information. The first work on dynamic convolution was the dynamic filtering network (DFN) [33]. This method can be used to directly obtain the weights of the convolution kernel by means of a filtering network. Wu et al. [34] introduced an attention mechanism to generate an adaptive kernel, which can learn more relevant information from different neighborhoods. Likewise, there were several works that combine convolution with self-attention mechanisms, such as X-volution [35], ACmix [36], and TRRNet [31]. These works can establish global dependencies while reducing the computational burden imposed by self-attention. Moreover, Zhou et al. [37] proposed a decoupled dynamic filtering network (DDF), which was an enhancement of the DFN. Content flexibility can be achieved by decomposing the convolution kernel from the spatial and channel dimensions. Some recent studies have introduced dynamic convolution into pansharpening. Jin et al. [11], inspired by DDF, combined local adaptation with global harmonic deviation to improve feature extraction capability from RSIs. Shi et al. [12] extracted information from PANI and MSI by utilizing two independent branches and then combining them to generate an adaptive discriminant kernel. The discriminant kernel can handle different attributes of diverse source images. The aforementioned works inspired us to introduce dynamic convolution into our endeavors. Under the guidance of the proposed multibranch pyramid structure, we develop three distinct task-specific AEIMs for three branches. These modules are capable of extracting a wide range of contextual information from diverse source data, thereby maximizing the integration of advantageous information in the input image pairs.

III. Methodology

In this section, we furnish a comprehensive description of the proposed methodology. The motivation and overview of MMAPP are presented first. Subsequently, we outline the primary components of the model, including the multibranch pyramid structure, three AEIMs, and multiscale reconstruction constraint.

A. Motivation and Overview

The design principles followed by existing DL methods include single-branch structure, dual-branch structure, and multiscale structure. The utilization of SR by the single-branch structure effectively enhances the spatial resolution of LRMSI. However, it may result in excessive smoothness in certain local areas, which is detrimental to the subsequent interpretation of RSIs. The dual-branch structure employs two branches to extract the robust feature representations of PANI and MSI in parallel before fusion processing, taking into account the heterogeneity between diverse source data. Unfortunately, this structure is responsible for a rigid fusion process and distorted outcomes due to the neglect of the scale disparities between diverse source images. The UNet-based multiscale structure is capable of utilizing the information of various scales to alleviate the distortion issue caused by scale disparities. However, it accomplishes this by sampling diverse source images directly to the same scale, performing the concatenation operation in channel dimension, and finally extracting the information of various scales simultaneously, disregarding the heterogeneity between diverse source data. Combining a multiscale structure with a dual-branch structure may offer a potential solution. Nonetheless, the mere amalgamation of the two can be quite blunt, and it is imperative to establish a state that facilitates the seamless integration of diverse source characteristics. Therefore, we have developed a multibranch pyramid structure to build bridges between diverse source images at various scales. This novel structure contains three distinct branches that efficiently and seamlessly integrates the data flow at various scales by means of the pyramid structure.

As depicted in Fig. 2, it contains three branches, including the PAN branch, the MS branch, and the fusion branch. The MMAPP aims to fuse PANI $\mathcal{P} \in \mathbb{R}^{H\times W\times 1}$ with LRMSI $\mathcal{M} \in \mathbb{R}^{h\times w\times C}$ to obtain the HRMSI $\tilde{\mathcal{M}} \in \mathbb{R}^{H\times W\times C}$. The h/H and w/W represent the height and width, respectively. Besides, C represents the number of bands. Overall, the proposed methodology can be summarized as

$$\tilde{\mathcal{M}} = f_{\mathrm{FB}}\left(\left[f_{\mathrm{PB}}\left(\mathcal{P}\right),\ f_{\mathrm{MB}}\left(\mathcal{M}\right)\right]\right) \tag{1}$$

where f_{PB}, f_{MB}, and f_{FB} represent the PAN, MS, and fusion branches, respectively. $[\cdot]$ represents the concatenation operation. The multibranch structure is developed to exploit the heterogeneity of $\mathcal{P}$ and $\mathcal{M}$. In addition, each branch contains a pyramid structure that efficiently and seamlessly integrates data flows at various scales across distinct branches. In the meantime, we have developed three specialized AEIMs for each branch: the TSAM for the PAN branch, the SICM for the MS branch, and

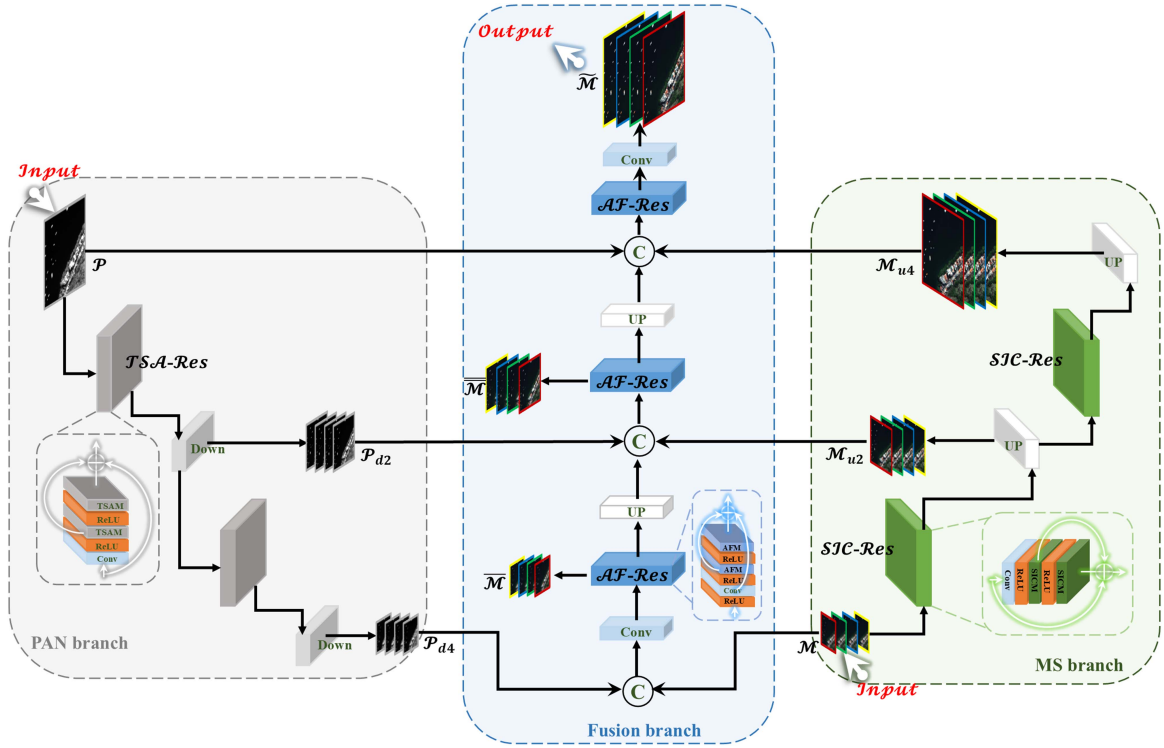

Fig. 2. Overall architecture of the proposed MMAPP. The multibranch structure is developed to exploit the heterogeneity of diverse source images. Each branch contains a pyramid structure that efficiently and seamlessly integrates data flows at various scales across distinct branches. Meanwhile, distinct AEIMs within the pyramid structure of distinct branches are specifically designed to cater to diverse sources and distinct stages of the pansharpening process.

the AFM for the fusion branch. These AEIMs are specifically designed to cater to diverse sources and distinct stages of the pansharpening process, enabling the extraction and integration of more advantageous information. Finally, multiscale constraints are developed to obtain high-fidelity fused images.

B. Multibranch Pyramid Structure

We develop a multibranch pyramid structure that explores the various scale information of the heterogeneous $\mathcal{P}$ and $\mathcal{M}$. Specifically, we develop distinct feature extraction branches for $\mathcal{P}$ and $\mathcal{M}$, referred to as PAN and MS branches, instead of utilizing a single-branch structure or a UNet-based multiscale structure to directly stitch two diverse source images as input. Moreover, an additional branch is dedicated to achieving feature fusion and reconstruction, commonly referred to as the fusion branch. The notion of pyramid is used in each of the three branches to preset three different scales. In the meantime, the connection between different scale features is established through long range residual learning and concatenation learning, preserving more spatial textures and spectral information.

In the PAN branch, spatial features at various scales can be extracted by progressively downsampling $\mathcal{P}$ by a factor of 2 and 4, respectively, by means of a pyramid structure. The process can be quantified as follows:

$$\mathcal{P}, \mathcal{P}_{d2}, \mathcal{P}_{d4} = f_{\mathrm{PB}}(\mathcal{P}) \tag{2}$$

where $\mathcal{P}_{d2}$ and $\mathcal{P}_{d4}$ are the results of the downsampling of $\mathcal{P}$ by the factor of 2 and 4 in the PAN branch, respectively. In a similar manner, $\mathcal{M}$ is gradually upsampled by factor of 2 and 4, respectively. The process can be quantified as follows:

$$\mathcal{M}, \mathcal{M}_{u2}, \mathcal{M}_{u4} = f_{\mathrm{MB}}(\mathcal{M}) \tag{3}$$

where $\mathcal{M}_{u2}$ and $\mathcal{M}_{u4}$ are the results of the upsampling of $\mathcal{M}$ by the factor of 2 and 4 in the MS branch, respectively. Fig. 2 illustrates that the PAN and MS branches extract different scale features from opposite directions, which could enhance the interaction between the two branches and enhance the performance of feature extraction.

Subsequently, $\mathcal{P}_{d4}$ and $\mathcal{M}$ are combined in the spectral dimension, and forwarded to the fusion branch. Therefore, the $\tilde{\mathrm{M}}$ in (1) is remodeled as follows:

$$\bar{\mathcal{M}}, \bar{\bar{\mathcal{M}}}, \tilde{\mathcal{M}} = f_{\mathrm{FB}}([\mathcal{P}_{d4}, \mathcal{M}]) \tag{4}$$

where $\bar{\mathcal{M}}$, $\bar{\bar{\mathcal{M}}}$, and $\tilde{\mathcal{M}}$ denote the outputs at various scales, which are utilized to construct the multiscale reconstruction constrains. The fusion branch is able to obtain $\tilde{\mathcal{M}}$ by fusing and reconstructing the multiscale spatial features in the PAN branch and the multiscale spectral features in the MS branch.

C. Adaptive Extraction and Integration Modules for Heterogeneous Features

In order to extract and integrate the multiscale spatial features in the PAN branch and the multiscale spectral features in the

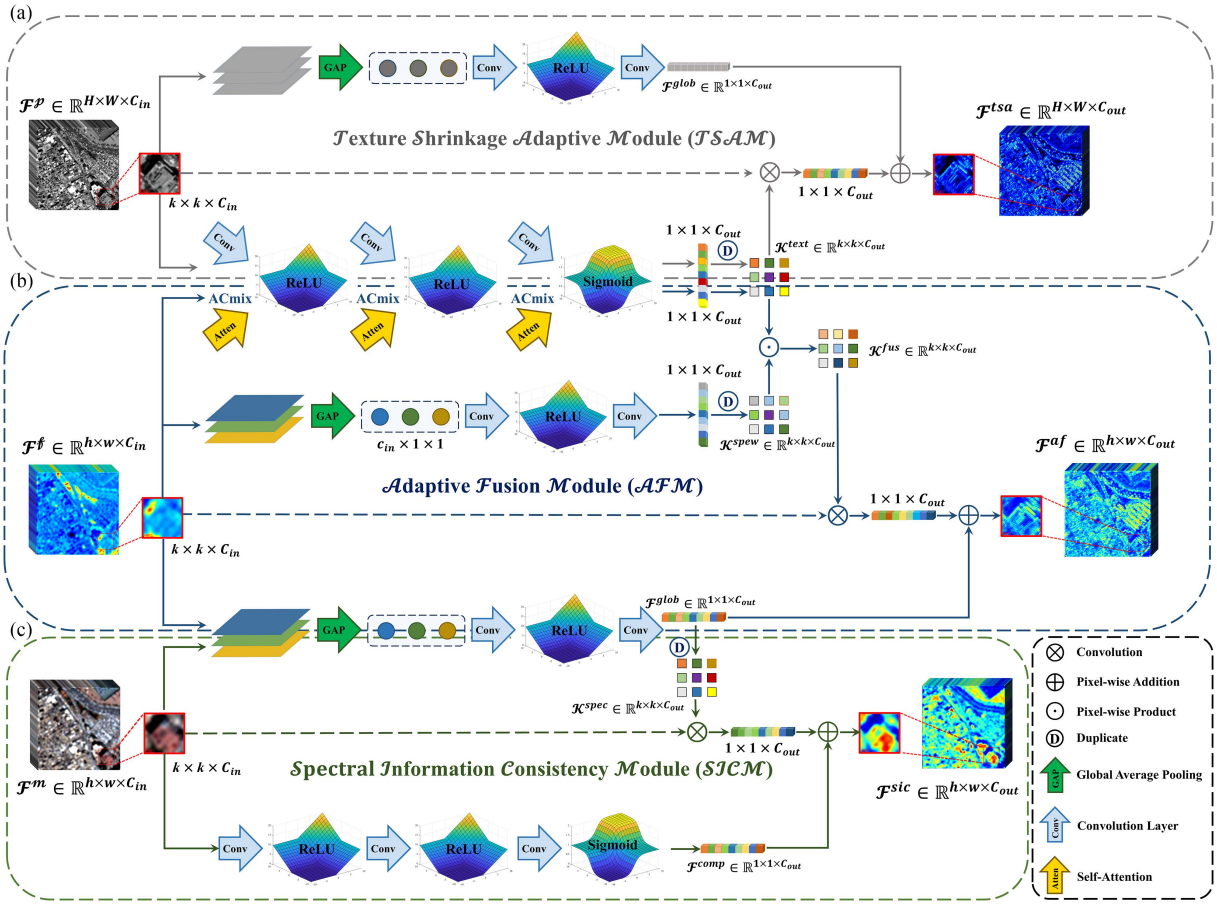

Fig. 3. Schematic of three AEIMs developed for three different branches. (a) TSAM for the PAN branch. (b) AFM for the fusion branch. (c) SICM for the MS branch.

MS branch, we develop a specific AEIM for each branch, comprising TSAM for the PAN branch, SICM for the MS branch, and AFM for the fusion branch. These AEIMs are specifically engineered to cater to diverse sources and distinct stages of the pansharpening process, enabling the extraction and integration of more advantageous information. They possess the capability to generate adaptive weights based on heterogeneous inputs, thereby enabling adaptive attention to distinct local regions and preserving more spatial textures and spectral information.

1) TSAM: The objective of this AEIM is to extract the multiscale spatial features from $\mathcal{P}$, which corresponds to the issue solved by (2). The TSAM has been specifically designed for the PAN branch, which focuses on extracting the precise multiscale spatial textures. The objects present in distinct local regions in RSIs frequently exhibit distinct characteristics. Conventional convolution employs context-agnostic convolution kernels, which assign identical weight to distinct local regions, resulting in the loss of crucial texture information, particularly during the process of feature scale shrinkage. In the contrary, TSAM focuses on different local regions in the same spatial dimension and generates adaptive weights for different local regions. In addition, since $\mathcal{P}$ is mapped to multiple channel dimensions during feature extraction, we introduce global spatial indexes for different channel dimensions to extract finer spatial textures by means of residual learning. As illustrated in Fig. 3(a), stable local features are extracted by a series of convolution operations, which are then subjected to a series of transformations to generate adaptive weights for the corresponding local regions. A series of transformations includes deformation, updimensionality, and duplication operations, which we ignore for ease of understanding in the following equations. Subsequently, the generated weights are utilized to execute a convolution operation on the corresponding locations in the input feature map, while the global feature representation in this region is extracted. Finally, the two are subjected to an additive operation to obtain a precise representation of spatial texture. The process can be quantified as follows:

$$\mathcal{K}^{\text{text}}_{(i,j)} = \text{Sig}\left(\widehat{\text{Conv}}\left(\mathcal{F}^{\mathcal{P}}_{(i,j)}\right)\right) \tag{5}$$

$$\mathcal{F}^{\text{glob}}_{(i,j)} = \widehat{\text{Conv}}\left(\text{GAP}\left(\mathcal{F}^{\mathcal{P}}_{(i,j)}\right)\right) \tag{6}$$

$$\mathcal{F}^{\text{tsa}} = \mathcal{K}^{\text{text}} \otimes \mathcal{F}^{\mathcal{P}} \oplus \mathcal{F}^{\text{glob}} \tag{7}$$

where $\mathcal{F}^{\mathcal{P}}_{(i,j)}$ represents the features in the PAN branch with spatial coordinates (i, j). $\widehat{\text{Conv}}$ represents a series of convolutional and ReLU layers. Sig and GAP represent sigmoid and global average pooling, respectively. $\mathcal{K}^{\text{text}}$ and $\mathcal{F}^{\text{glob}}$ represent

the convolution kernel with extracted adaptive weights and the extracted global feature representation, respectively. Besides, $\otimes$ and $\oplus$ represent the convolution and addition operations, respectively. The operation above is defined as $\mathcal{TSA}$, i.e., feature extraction by TSAM. $\mathcal{F}^{\text{tsa}}$ is the output of the $\mathcal{TSA}$ operation. Subsequently, we proceed to follow the residual structure depicted in the grey dashed box in Fig. 2 to construct the residual block pertaining to TSAM, referred to as $\mathcal{TSA}$-$\mathcal{R}$es, while the operation is defined as $\widehat{\mathcal{TSA}}$. The initial convolution block in the residual structure is employed to increase the dimensionality of the input image, whereas the subsequent residual blocks are employed to extract robust spatial texture characteristics. With $\mathcal{TSA}$-$\mathcal{R}$es, it is possible to extract the spatial texture from $\mathcal{P}$ at various shrinkage scales. As a result, the $\mathcal{P}_{d2}$ and $\mathcal{P}_{d4}$ in (2) are remodeled as follows:

$$\mathcal{P}_{d2} = \text{MP}\left(\widehat{\mathcal{TSA}}\left(\mathcal{P}\right)\right) \tag{8}$$

$$\mathcal{P}_{d4} = \text{MP}\left(\widehat{\mathcal{TSA}}\left(\mathcal{P}_{d2}\right)\right) \tag{9}$$

where MP denotes the MaxPooling operation.

In this way, three kinds of spatial texture features with different shrinkage scales are obtained by progressively executing $\mathcal{TSA}$ operation on $\mathcal{P}$. The $\mathcal{TSA}$ operation generates the kernel weights of this local region according to the distinct positions of the input features, which can maximize the preservation of the differences between distinct local regions in RSIs. The extracted fine-grained spatial textures lay the foundation of the subsequent spatial–spectral feature integration.

2) SICM: This AEIM attempts to extract the multiscale spectral features from $\mathcal{M}$, which corresponds to the issue resolved by (3). The SICM has been specifically designed for the MS branch, which focuses on extracting the high-fidelity multiscale spectral information. Different targets in RSIs exhibit distinct spectral curves. Unlike TSAM, SICM focuses its domain on the spectral dimension and generates adaptive weights for distinct spectral bands within the same spatial region. It is well known that $\mathcal{P}$ and $\mathcal{M}$ are imaged by different sensors, while radiometric discrepancy between them exists. This means that the HRMSI contains spatial details that are absent in $\mathcal{P}$ but present in $\mathcal{M}$ [7]. For this reason, we simultaneously extract the spatial information within the same spatial region in order to obtain spatial information complementary to $\mathcal{P}$. As shown in Fig. 3(c), the attention is focused on the spectral dimension by GAP and a series of convolution operations are used to extract the robust spectral features. A series of deformation, updimensionality, and duplication operations are then used to generate the spectral adaptive weights. Simultaneously, convolution operations are employed to extract spatial feature representations in the corresponding spatial region that are complementary to $\mathcal{P}$. Finally, the convolution kernels with spectral adaptive weights are convolved with the input features and the corresponding complementary spatial feature representations are added to obtain high-fidelity spectral features. The process can be quantified as follows:

$$\mathcal{K}^{\text{spec}}_{(i,j)} = \widehat{\text{Conv}}\left(\text{GAP}\left(\mathcal{F}^{m}_{(i,j)}\right)\right) \tag{10}$$

$$\mathcal{F}^{\text{comp}}_{(i,j)} = \text{Sig}\left(\widehat{\text{Conv}}\left(\mathcal{F}^{m}_{(i,j)}\right)\right) \tag{11}$$

$$\mathcal{F}^{\text{sic}} = \mathcal{K}^{\text{spec}} \otimes \mathcal{F}^{m} \oplus \mathcal{F}^{\text{comp}} \tag{12}$$

where $\mathcal{F}^{m}_{(i,j)}$ represents the features in the MS branch with spatial coordinates (i, j). $\mathcal{K}^{\text{spec}}$ and $\mathcal{F}^{\text{comp}}$ represent the convolution kernel with extracted spectral adaptive weights and the extracted complementary spatial features, respectively. The operation above is defined as $\mathcal{SIC}$, i.e., feature extraction by SICM. $\mathcal{F}^{\text{sic}}$ is the output of the $\mathcal{SIC}$ operation. Similar to $\mathcal{TSA}$-$\mathcal{R}$es, we continue to construct the residual block associated with SICM, referred to as $\mathcal{SIC}$-$\mathcal{R}$es, while the operation is defined as $\widehat{\mathcal{SIC}}$. With $\mathcal{SIC}$-$\mathcal{R}$es, it is feasible to obtain consistent spectral information from $\mathcal{M}$ while performing upsampling. As a result, the $\mathcal{M}_{u2}$ and $\mathcal{M}_{u4}$ in (3) are remodeled as follows:

$$\mathcal{M}_{u2} = \text{UP}\left(\widehat{\mathcal{SIC}}\left(\mathcal{M}\right)\right) \tag{13}$$

$$\mathcal{M}_{u4} = \text{UP}\left(\widehat{\mathcal{SIC}}\left(\mathcal{M}_{u2}\right)\right) \tag{14}$$

where UP denotes the upsampling operation.

In this way, three kinds of spectral consistency features with various scales are obtained by progressively executing $\mathcal{SIC}$ operation on $\mathcal{M}$. By focusing on the spectral dimension, $\mathcal{SIC}$ possesses the capability to obtain rich multiscale spectral information. Moreover, this information comprises spatial information that is complementary to $\mathcal{P}$, which can lead to the formation of a more reasonable spectral distribution. These serve as the foundation for subsequent high spatial–spectral fidelity fusion.

3) AFM: This AEIM aims to integrate the output of the PAN branch with the output of the MS branch. The AFM has been specifically designed for the fusion branch, which focuses on integrating the multiscale spatial features in the PAN branch and the multiscale spectral features in the MS branch. AFM combines the advantages of TSAM and SICM, which focus on both spatial and spectral dimensions to seamlessly integrate advantageous information in heterogeneous features. However, both ordinary convolution and dynamic convolution possess limited receptive fields and are incapable of capturing global dependencies. The correlation between different local regions is important to prevent distortion caused by overattention to a certain local area when fusing heterogeneous features from different branches. Self-attention mechanism has the potential to address these issues, however, it will come with computational burdens. Given the above factors, we introduce a convolutional and self-attention combination module, ACmix [36], which has been demonstrated to successfully combine the advantages of convolution and self-attention. As shown in Fig. 3(b), the AFM establishes global dependencies through ACmix when generating adaptive fusion weights for different local regions. Subsequently, we adopt the identical spectral weight generation scheme as in SICM and execute pixelwise dot-product of the spatial adaptive weights with the corresponding spectral adaptive weights. This approach enables us to obtain adaptive fusion weights for heterogeneous features, thereby facilitating their seamless integration. Additionally, similar to TSAM, we introduce global spatial features for different spectral bands and

integrate spatial–spectral features with fine-graininess by means of residual learning. The process can be quantified as follows:

$$\mathcal{K}^{\text{spaw}}_{(i,j)} = \text{Sig}\left(\widehat{\text{ACmix}}\left(\mathcal{F}^{\natural}_{(i,j)}\right)\right) \tag{15}$$

$$\mathcal{K}^{\text{spew}}_{(i,j)} = \widehat{\text{Conv}}\left(\text{GAP}\left(\mathcal{F}^{\natural}_{(i,j)}\right)\right) \tag{16}$$

$$\mathcal{K}^{\text{fus}}_{(i,j)} = \mathcal{K}^{\text{spaw}}_{(i,j)} \odot \mathcal{K}^{\text{spew}}_{(i,j)} \tag{17}$$

$$\mathcal{F}^{\text{af}} = \mathcal{K}^{\text{fus}} \otimes \mathcal{F}^{\natural} \oplus \mathcal{F}^{\text{glob}} \tag{18}$$

where $F^{f}_{(i,j)}$ represents the input of the fusion branch with spatial coordinates (i, j). $\widehat{\text{ACmix}}$ represents a series of ACmix and ReLU layers. $\mathcal{K}^{\text{spaw}}$, $\mathcal{K}^{\text{spew}}$, and $\mathcal{K}^{\text{fus}}$ represent the convolution kernel with spatial adaptive weights, corresponding spectral adaptive weights, and corresponding adaptive fusion weights, respectively. $\odot$ represents the pixelwise dot-product operation. Besides, $\mathcal{F}^{\text{glob}}$ is obtained in a similar way as in (6) with the input $\mathcal{F}^{\natural}_{(i,j)}$. The operation above is defined as $\mathcal{AF}$, i.e., feature integration by AFM. $\mathcal{F}^{\text{af}}$ is the output of the $\mathcal{AF}$ operation. Subsequently, we proceed to follow the residual structure depicted in the blue dashed box in Fig. 2 to construct the residual block pertaining to AFM, referred to as $\mathcal{AF}$-$\mathcal{R}$es, while the operation is defined as $\widehat{\mathcal{AF}}$. The initial ReLU layer in the residual structure is used to reduce the redundancy of heterogeneous features. With $\mathcal{AF}$-$\mathcal{R}es$, RSIs with fine-grained spatial textures and high-fidelity spectral information can be reconstructed progressively. As a result, the $\bar{\mathcal{M}}$, $\bar{\bar{\mathcal{M}}}$, and $\tilde{\mathcal{M}}$ in (4) are remodeled as follows:

$$\bar{\mathcal{M}} = \widehat{\mathcal{AF}}\left(\text{Conv}\left(\left[\mathcal{P}_{d4}, \mathcal{M}\right]\right)\right) \tag{19}$$

$$\bar{\bar{\mathcal{M}}} = \widehat{\mathcal{AF}}\left(\left[\mathcal{P}_{d2}, \text{UP}\left(\bar{\mathcal{M}}\right), \mathcal{M}_{u2}\right]\right) \tag{20}$$

$$\tilde{\mathcal{M}} = \widehat{\mathcal{AF}}\left(\left[\mathcal{P}, \text{UP}\left(\bar{\bar{\mathcal{M}}}\right), \mathcal{M}_{u4}\right]\right). \tag{21}$$

In this way, fused images of three different scales are reconstructed by progressively executing $\mathcal{AF}$ operation on multiscale heterogeneous features from the different branches. By focusing on both spatial and spectral dimensions, the $\mathcal{AF}$ operation is capable of generating adaptive fusion weights that correspond to different regions, thereby efficiently and seamlessly integrating multiscale heterogeneous features. With the assistance of AFM, the fusion branch can concentrate on integrating the advantageous information from the PAN and MS branches, thus balancing spatial and spectral preservation and generating fused images with high spatial–spectral fidelity.

D. Multiscale Reconstruction Constraint

To ensure that fused images contain more realistic information, we impose reconstruction constraints on outputs of various scales in different branches. The reconstruction constraints can be specifically quantified as follows:

$$\mathcal{L}_{d4} = \frac{\sum_{m=1}^{H}\sum_{n=1}^{W}\sum_{l=1}^{C}\|\bar{\mathcal{M}}(m,n,l) - \bar{\mathcal{G}}(m,n,l)\|_F}{\text{HWC}} \tag{22}$$

$$\mathcal{L}_{d2} = \frac{\sum_{m=1}^{H}\sum_{n=1}^{W}\sum_{l=1}^{C}\|\bar{\bar{\mathcal{M}}}(m,n,l) - \bar{\bar{\mathcal{G}}}(m,n,l)\|_F}{\text{HWC}} \tag{23}$$

$$\mathcal{L}_{\text{ori}} = \frac{\sum_{m=1}^{H}\sum_{n=1}^{W}\sum_{l=1}^{C}\|\tilde{\mathcal{M}}(m,n,l) - \mathcal{G}(m,n,l)\|_F}{\text{HWC}} \tag{24}$$

$$\mathcal{L}_{\text{spe}} = \frac{\sum_{m=1}^{H}\sum_{n=1}^{W}\sum_{l=1}^{C}\|\mathcal{M}_{u4}(m,n,l) - \mathcal{M}^{U}(m,n,l)\|_F}{\text{HWC}} \tag{25}$$

$$\mathcal{L}_{\text{Total}} = \lambda_1 \cdot \mathcal{L}_{d4} + \lambda_2 \cdot \mathcal{L}_{d2} + \lambda_3 \cdot \mathcal{L}_{\text{ori}} + \lambda_4 \cdot \mathcal{L}_{\text{spe}} \tag{26}$$

where $\mathcal{G} \in \mathbb{R}^{H\times W\times C}$ is the ground truth (GT). $\bar{\bar{\mathcal{G}}}$ and $\bar{\mathcal{G}}$ are the results after downsampling $\mathcal{G}$ by a factor of two and four, respectively. Besides, $\mathcal{M}^{U} \in \mathbb{R}^{H\times W\times C}$ is the results after upsampling $\mathcal{M}$ by a factor of four. The objectives of $\mathcal{L}_{d4}$, $\mathcal{L}_{d2}$, and $\mathcal{L}_{\text{ori}}$ are to provide guidance to the fusion branch in obtaining precise multiscale reconstructed images, while $\mathcal{L}_{\text{spe}}$ aims to enable the MS branch to concentrate more on learning spectral distributions that are consistent with $\mathcal{M}$. Besides, λ_1, λ_2, λ_3, and λ_4 are four proportion coefficients that have been empirically set to 0.2, 0.3, 0.5, and 0.01, which will be discussed in Section IV-G. In particular, the initial learning rate, epochs, and batch size are set to 3e−4, 600, and 32, respectively. The Adam is chosen as the optimizer and the learning rate is decayed by 10% every 100 epochs.

IV. Experiments

A. Datasets

To comprehensively evaluate the efficacy of the proposed MMAPP, we select three publicly available datasets from three different satellites in PanCollection[1] [38]. Table I shows the basic information of each dataset in detail.

1) **GaoFen-2 (GF2) dataset** contains image data captured by the GF2 satellite in Guangzhou. The size of MSI is 6907 × 7300, the 1000 pixels on the right side of the image are used for testing, and the rest of the image is cut into 22 010 64 × 64 image patches for training and validation.
2) **QuickBird (QB) dataset** mainly contains data captured by the QB satellite in Indianapolis. The size of MSI is 4096 × 4096, and like the GF2 dataset, 1000 pixels on the right side of the image were used for testing, and the rest of the image was cut into 19 044 64 × 64 image patches for training and validation.
3) **WorldView-3 (WV3) dataset** mainly contains data captured by the WV3 satellite in Rio and Tripoli. The size of MSI of 2811 × 3408 and 1871 × 2020 in Rio and Tripoli, respectively. Among them, the right 1/4 part of the images taken in Rio region in May and Tripoli region in August are cut out for testing, and the rest is cut into 10 794 64 × 64 image patches for training and validation.

[1][Online]. Available: https://github.com/liangjiandeng/PanCollection

TABLE I
BASIC INFORMATION OF EACH DATASET

		Bit depth	Band	Resolution(m)	Train set		Valid set		RST set		FST set		Scene
					Size	Number	Size	Number	Size	Number	Size	Number	
GF2	PANI	10	1	0.8	64×64	19 809	64×64	2201	256×256	20	512×512	20	coasts, vegetation, buildings, urban
	MSI		4	3.2	16×16		16×16		64×64		128×128		
QB	PANI	11	1	0.61	64×64	17 139	64×64	1905	256×256	20	512×512	20	
	MSI		4	2.44	16×16		16×16		64×64		128×128		
WV3	PANI	11	1	0.3	64×64	9714	64×64	1080	256×256	20	512×512	20	
	MSI		8	1.2	16×16		16×16		64×64		128×128		

All three datasets are publicly available, which can be accessed in published articles and corresponding websites. It is worth noting that the experiment mainly consists of reduced-scale test (RST) and full-scale test (FST), wherein the RST mainly evaluates the fitting capability of the model, while the FST mainly evaluates the generalization capability in real-world scenarios. There are no identical images between the two distinct test sets from each of the three datasets, even though they are extracted from the same city.

B. Evaluation Metrics

We use five well-known metrics exactly evaluate the fusion performance in RST, namely the relative dimensionless global error in synthesis (ERGAS) [39], the spectral angle mapper (SAM) [40], the spatial correlation coefficient (sCC) [41], the *Q2n* [42], and the relative average spectral error (RASE) [41]. Among them, ERGAS and sCC are capable of assessing the spatial fidelity of the sharpened image, whereas SAM, *Q2n*, and RASE primarily assess the spectral fidelity. In addition, the closer the values of ERGAS, SAM, and RASE are to 0, the better, while the closer the values of sCC and *Q2n* are to 1, the better. Similarly, we employ five nonreference metrics to evaluate the performance in FST, namely the spatial distortion index (D_s) [43], the spectral distortion index (D_λ) [43], the spectral distortion index from Khan's protocol (D_λ^F) [44], the quality with no reference (QNR) [43] and the hybrid quality with no reference (HQNR) [44]. The less distorted the sharpened image becomes, the closer the values of D_s, D_λ, and D_λ^F are to 0; similarly, the closer the values of QNR and HQNR are to 1, the better the balance between spatial and spectral preservation of the sharpened image becomes.

C. Comparative Methods

To evaluate the efficacy of MMAPP, we choose 22 advanced methods for comparison experiments, including EXP [45], PRACS [7], BT-H [5], BDSD-PC [46], AWLP [18], MTF-GLP-HPM-R (HPM-R) [47], MTF-GLP-FS (FS) [48], TV [23], PWMBF [24], SR-D [25], PanNet [13], BDPN [14], SRPPNN [10], MUCNN [15], LAGConv [11], ADKNet [12], TANI [30], TRRNet [31], MMFN [16], RSANet [32], MSSTNet [49], and CSTNet [29]. Among them, EXP is an upsampling method, also called 23-tap polynomial interpolation. PRACS, BT-H, and BDSD-PC are CS-based methods. AWLP, HPM-R, and FS are MRA-based methods. TV, SR-D, and PWMBF are VO-based methods. These nine methods all belong to the traditional method. Besides, the rest are DL-based methods. In particular, the design principle of PanNet, SRPPNN, LAGConv, ADKNet, and RSANet belongs to single-branch structure, while the design principle of BDPN, TANI, TRRNet, and CSTNet belongs to dual-branch structure, while the design principle of MUCNN, MSSTNet, and MMFN belongs to UNet-based multiscale structure. For a fair comparison, we use the authors' officially released code and the setup described in the original paper. All of the codes are executed on a computer that is equipped with an i5-11600 CPU and two GTX-3060 GPUs. In addition, for the sake of reproducibility, all implementations will be published at MMAPP.

D. Experiments on GF2 Dataset

Fig. 4 shows the RST results on the GF2 dataset. The red box in the lower left corner is the result of enlarging the local area. Despite its poor results in terms of spatial quality, the fusion results obtained by EXP can be used as the spectral reference, presumably due to the absence of the PANI. Furthermore, it is apparent that all CS-based and MRA-based methods exhibit spectral distortion, yet they maintain fine spatial texture in the zoom-in zone. For the VO-based methods, TV demonstrates good visual effects, SR-D demonstrates severe spatial distortions, and PWMBF demonstrates significant spectral distortion. In the methods employing the single-branch structure, PanNet and ADKNet show significant spectral distortion, while SRPPNN, LAGConv, and RSANet exhibit too much smoothing in the zoomed-in region. In contrast, ADKNet retains more texture information. In the methods using the dual-branch structure, BDPN performs poorly in both spatial and spectral retention. TANI and CSTNet demonstrate slight spectral distortion and significant spatial distortion, while TRRNet demonstrates better visual effects although some loss of detail occurs. In the methods using multiscale structures, MUCNN demonstrates significant spatial blurring. MSSTNet demonstrates slight spectral distortion and significant spatial distortion. MMFN demonstrates good spectral fidelity but it loses some of its spatial details, as evidenced by the white part of the edges of the zoomed-in region. Overall, only ADKNet, TRRNet, and the proposed MMAPP exhibit a balance between spatial and spectral preservation. In contrast, MMAPP exhibits the closest spatial texture and spectral distribution to GT. The corresponding absolute error maps (AEMs) are provided to illustrate the difference between the different methods, as shown in Fig. 5. It is clear that the proposed MMAPP exhibits the least residuals, which further validates that MMAPP can achieve the best qualitative results.

Further, the qualitative results on FST are shown in Fig. 6, revealing a number of resemblances. Three CS-based methods all maintain good spatial texture but lose spectral information.

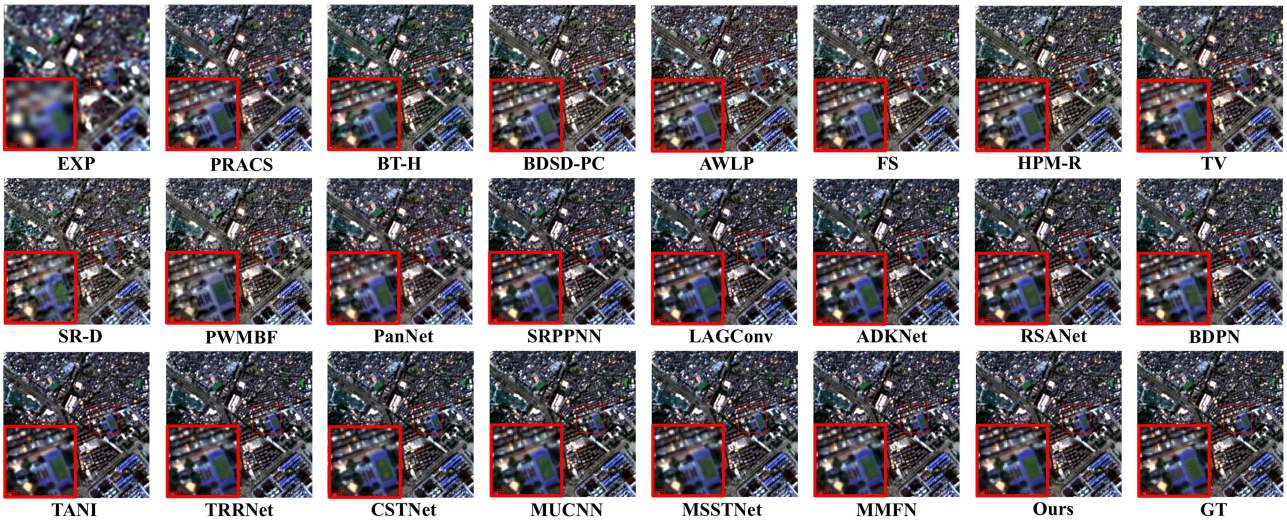

Fig. 4. Qualitative comparison under the RST on GF2 dataset (selected bands: red, green, and blue).

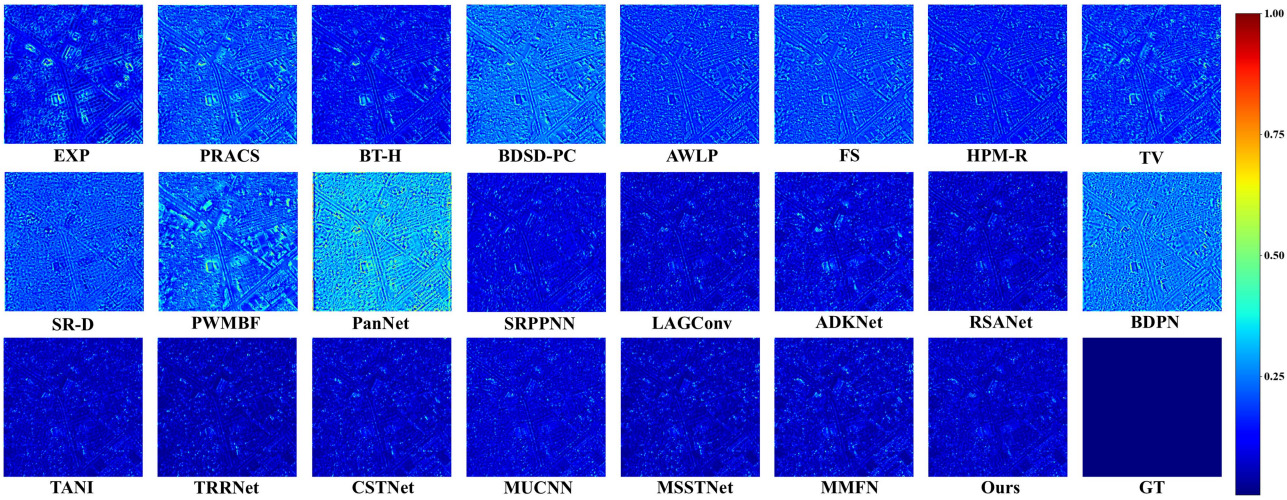

Fig. 5. AEMs that correspond to each method are depicted in Fig. 4.

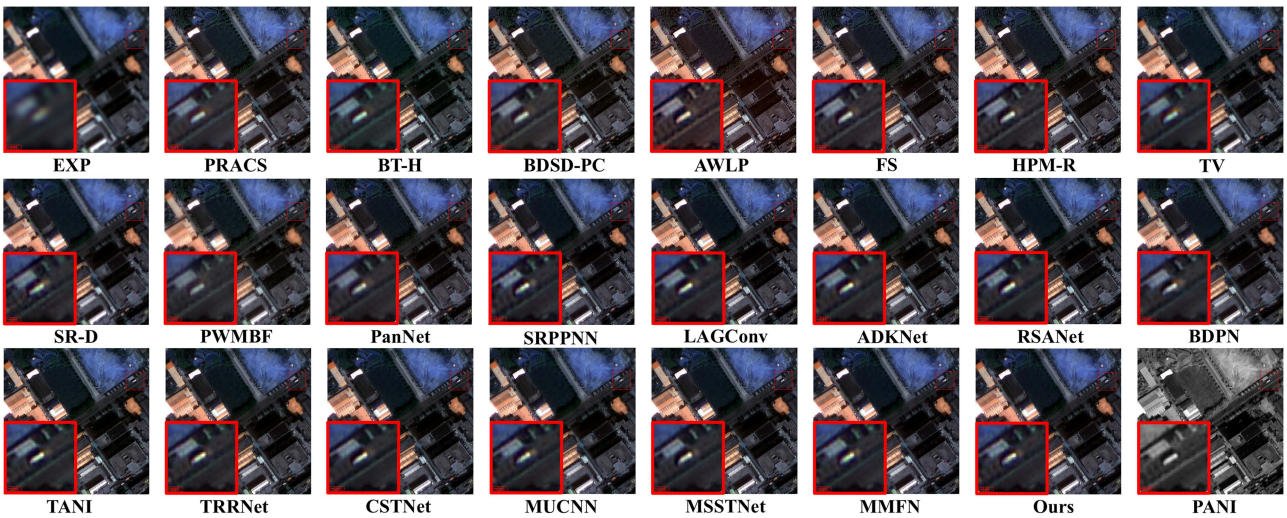

Fig. 6. Qualitative comparison under the FST on GF2 dataset (selected bands: red, green, and blue).

TABLE II
QUANTITATIVE COMPARISON ON GF2 DATASET

Comp.	Quantitative evaluation in RST					Quantitative evaluation in FST				
	ERGAS↓(±std)	SAM↓(±std)	sCC↑(±std)	$Q2n$↑(±std)	RASE↓(±std)	D_s↓(±std)	D_λ↓(±std)	D_λ^F↓(±std)	QNR↑(±std)	HQNR↑(±std)
EXP	2.4094±0.4647	1.8531±0.3459	0.9542±0.0203	0.7971±0.0430	8.5984±1.7342	0.0263±0.0176	0.0000±0.0000	0.0140±0.0048	0.9737±0.0176	0.9601±0.0187
CS-based methods										
PRACS	1.6482±0.3425	1.7135±0.3120	0.9625±0.0143	0.8975±0.0270	5.9004±1.2328	0.0469±0.0172	0.0125±0.0086	0.0618±0.0191	0.9413±0.0227	0.8943±0.0264
BT-H	1.5526±0.3548	1.6819±0.3084	0.9690±0.0125	0.9089±0.0284	5.5194±1.2519	0.0555±0.0179	0.0216±0.0114	0.0639±0.0232	0.9242±0.0262	0.8841±0.0296
BDSD-PC	1.6954±0.3896	1.7243±0.3118	0.9689±0.0119	0.8847±0.0300	6.1165±1.4205	0.0559±0.0196	0.0130±0.0095	0.0810±0.0286	0.9319±0.0260	0.8678±0.0368
MRA-based methods										
AWLP	1.7241±0.3574	1.9687±0.4955	0.9329±0.0311	0.8610±0.0328	6.8346±1.7655	0.0491±0.0188	0.0154±0.0147	0.0398±0.0162	0.9364±0.0300	0.9131±0.0266
FS	1.6201±0.3526	1.6807±0.3394	0.9685±0.0118	0.8904±0.0250	5.8813±1.3128	0.0524±0.0173	0.0236±0.0134	0.0375±0.0130	0.9254±0.0273	0.9121±0.0210
HPM-R	1.6197±0.3632	1.6759±0.3465	0.9697±0.0110	0.8934±0.0245	5.8878±1.3675	0.0526±0.0176	0.0220±0.0131	0.0364±0.0120	0.9267±0.0275	0.9129±0.0209
VO-based methods										
TV	1.7767±0.4074	1.9190±0.3876	0.9612±0.0176	0.9051±0.0267	6.0435±1.3495	0.0478±0.0165	0.0217±0.0132	0.0584±0.0413	0.9317±0.0263	0.8967±0.0451
SR-D	1.6739±0.3169	1.6541±0.3107	0.9713±0.0126	0.8989±0.0240	6.0506±1.2045	0.0379±0.0108	0.0158±0.0105	0.0493±0.0208	0.9470±0.0243	0.9147±0.0212
PWMBF	1.8480±0.3275	1.9768±0.3307	0.9401±0.0187	0.8561±0.0251	6.8597±1.2207	0.0346±0.0109	0.0204±0.0095	0.0704±0.0255	0.9457±0.0179	0.8974±0.0286
Single-branch structures										
PanNet	1.3438±0.1944	1.5439±0.2339	0.9751±0.0097	0.9154±0.0352	4.8840±0.7304	0.0356±0.0127	0.0121±0.0095	0.0336±0.0262	0.9528±0.0193	0.9320±0.0296
SRPPNN	0.7299±0.1088	0.8250±0.1387	0.9933±0.0028	0.9778±0.0085	2.6163±0.4054	**0.0276±0.0111**	**0.0058±0.0056**	**0.0198±0.0079**	**0.9668±0.0149**	**0.9531±0.0135**
LAGConv	**0.7223±0.0963**	**0.8099±0.1282**	**0.9933±0.0027**	**0.9785±0.0096**	**2.5852±0.3656**	0.0397±0.0138	0.0113±0.0085	0.0296±0.0089	0.9495±0.0192	0.9319±0.0161
ADKNet	0.8215±0.1149	0.8830±0.1509	0.9915±0.0036	0.9721±0.0097	2.9927±0.4391	**0.0253±0.0100**	0.0086±0.0071	0.0218±0.0091	**0.9663±0.0152**	**0.9513±0.0137**
RSANet	0.8011±0.1248	0.8846±0.1514	0.9918±0.0034	0.9734±0.0103	2.8742±0.4696	0.0378±0.0133	0.0099±0.0083	0.0297±0.0100	0.9528±0.0184	0.9337±0.0160
Dual-branch structures										
BDPN	1.5190±0.3684	1.4359±0.2739	0.9765±0.0110	0.9234±0.0217	5.1054±1.3084	0.0364±0.0127	0.0117±0.0096	0.0357±0.0198	0.9524±0.0194	0.9292±0.0242
TANI	1.1494±0.2541	1.1067±0.2118	0.9859±0.0064	0.9498±0.0204	4.1662±0.9110	0.0419±0.0157	0.0105±0.0083	0.0409±0.0207	0.9481±0.0205	0.9190±0.0272
TRRNet	0.8246±0.1153	0.8858±0.1240	0.9927±0.0030	0.9685±0.0138	2.9839±0.4240	0.0310±0.0118	0.0074±0.0067	0.0394±0.0229	0.9619±0.0152	0.9308±0.0241
CSTNet	0.8151±0.1393	0.8958±0.1633	0.9923±0.0032	0.9734±0.0092	2.9048±0.5028	0.0373±0.0130	0.0102±0.0080	0.0274±0.0088	0.9530±0.0182	0.9364±0.0143
UNet-based multiscale structures										
MUCNN	0.8287±0.1512	0.8952±0.1722	0.9920±0.0035	0.9737±0.0078	2.9732±0.5574	0.0316±0.0113	0.0085±0.0065	**0.0202±0.0084**	0.9602±0.0157	0.9488±0.0128
MSSTNet	0.9358±0.1764	0.9641±0.1796	0.9896±0.0046	0.9676±0.0088	3.3735±0.6521	0.0375±0.0132	0.0100±0.0081	0.0266±0.0093	0.9529±0.0184	0.9369±0.0152
MMFN	**0.7167±0.1331**	**0.7710±0.1486**	**0.9940±0.0028**	**0.9796±0.0070**	**2.5714±0.4952**	0.0295±0.0111	**0.0066±0.0061**	0.0211±0.0085	0.9641±0.0151	0.9499±0.0124
The proposed multibranch pyramid structure										
Ours	**0.6715±0.1010**	**0.7587±0.1290**	**0.9943±0.0025**	**0.9809±0.0081**	**2.4206±0.3847**	**0.0272±0.0106**	**0.0060±0.0057**	**0.0221±0.0077**	**0.9670±0.0144**	**0.9534±0.0119**

The red, green, and blue highlights indicate the best, the second-best, and the third-best values, respectively.

The fusion outcomes of MRA-based methods exhibit varying degrees of spectral distortion and detail blurring, with AWLP exhibiting the lowest level of spectral fidelity. As with the results in RST, the VO-based methods SR-D and PWMBF exhibit significant distortion. For methods based on single-branch structures, PanNet and ADKNet demonstrate spectral distortion. The fusion results of SRPPNN demonstrate spatial distortion. LAGConv and RSANet balance spatial and spectral preservation. The fusion results of all four methods based on the dual-branch structure show slight spectral distortion. For the methods employing multiscale structures, the fusion results of MUCNN exhibit slight spectral distortion. The fusion results of MSSTNet are similar to those of CSTNet, exhibiting excessive smoothing in local areas. MMFN demonstrates good visual results. The proposed MMAPP exhibits the closest spatial texture to PANI, specifically in the pavement portion of the zoomed-in area. Furthermore, unlike other methods, MMAPP does not introduce spectral information that is not present in MSI, especially in the road tooth region. This supports the ability of the proposed MMAPP to achieve more realistic sharpening results.

The quantitative outcomes are presented in Table II. The top three ranked methods are highlighted in red, green, and blue, respectively. The ranking does not take EXP into account because it is a sampling method and not a pansharpening one. The outcomes show that DL-based methods outperform traditional method on the metrics. Specifically, the proposed MMAPP achieves the best scores on all reference metrics. In addition, on the nonreference metrics, the MMAPP does not perform the best on the spatial (D_s) and spectral distortion metrics (D_λ and D_λ^F), but it did perform the best on the two comprehensive metrics (QNR and HQNR), which further demonstrates that the proposed methodology is able to best balance spatial and spectral preservation.

E. Experiments on QB Dataset

Fig. 7 shows the RST results on the QB dataset. Except for PRACS and TV, all other traditional methods exhibit significant spatial and spectral distortions. In the methods employing the single-branch structure, PanNet exhibits severe spatial and spectral distortions. The fusion results of SRPPNN, ADKNet,

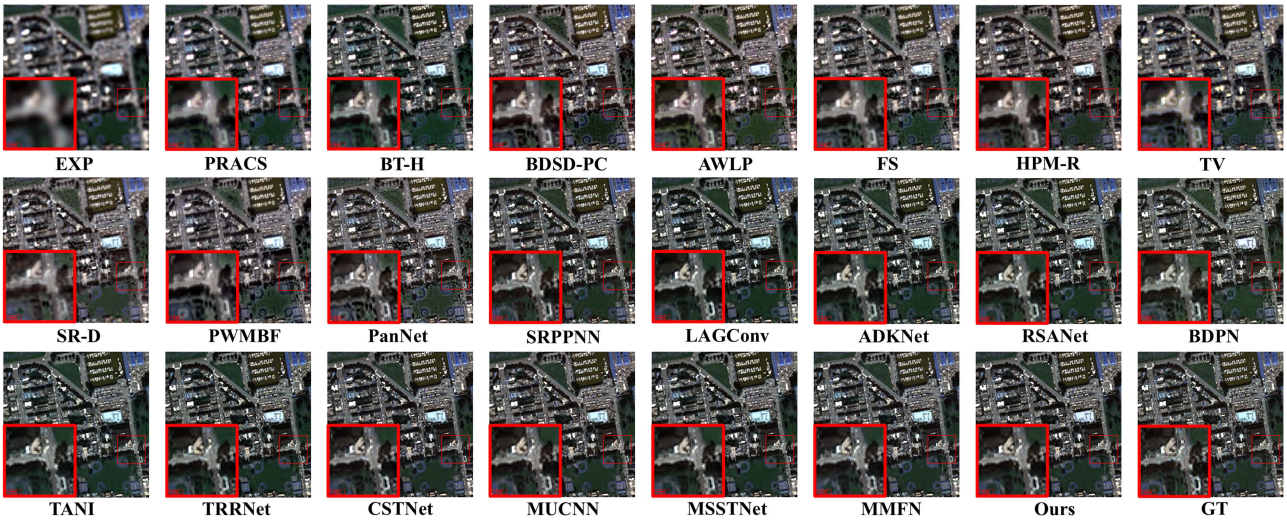

Fig. 7. Qualitative comparison under the RST on QB dataset (selected bands: red, green, and blue).

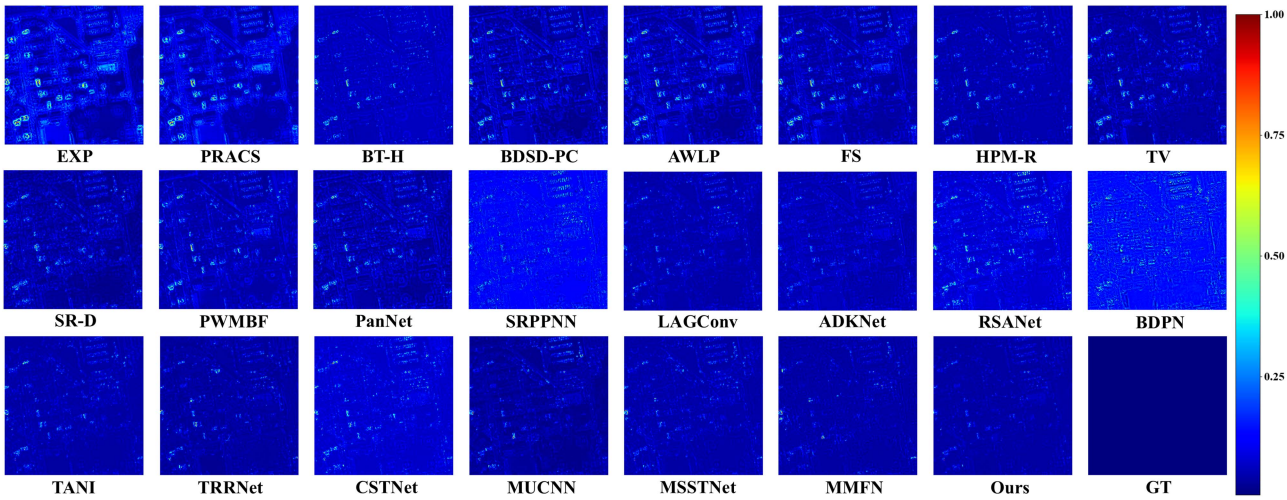

Fig. 8. AEMs that correspond to each method are depicted in Fig. 7.

and RSANet show slight spectral distortion. The fusion results of LAGConv exhibit good visual results. In the methods employing the dual-branch structure, the fusion results of BDPN demonstrate significant spatial and spectral distortions. TANI demonstrates slight spatial distortions, while CSTNet demonstrates slight spectral distortions. In the methods employing the multiscale structure, MUCNN exhibits slight spectral distortion. In general, LAGConv, TRRNet, MSSTNet, and MMFN demonstrate satisfactory subjective performance. In contrast, the proposed MMAPP exhibits a spatial texture and spectral distribution more consistent with that of GT, as shown in the lower left corner of the zoomed-in region. The AEMs in Fig. 8 show the differences between the methods more clearly and it can be seen that the proposed MMAPP possesses the least residuals.

Likewise, the results of the FST are as shown in Fig. 9. The performance of most of the methods is degraded, as evidenced by the fact that all traditional and DL methods show more or less spectral distortion. Among traditional methods, the results of BT-H exhibit the most pronounced spectral distortion, whereas SR-D exhibits the most severe spatial distortion, owing to the inability of the linear a priori assumptions to be applied to intricate scenarios. In the methods employing the single-branch structure, only RSANet displays a spectral fidelity similar to EXP. In the methods employing the dual-branch structure, it appears that BDPN and TRRNet exhibit severe spectral distortions, whereas TANI exhibits slight spectral distortion, in addition to the spatial distortion associated with the fusion outcome of CSTNet. Furthermore, MUCNN exhibits evident spectral distortion. The fusion results of MSSTNet, MMFN, and the proposed MMAPP are similar, with MSSTNet being the closest to EXP in terms of spectral fidelity, and MMFN and MMAPP in terms of spatial texture. Overall, MMAPP also achieves satisfactory results in FST, with no obvious spectral or spatial distortion.

The quantitative outcomes are presented in Table III. In RST, akin to the findings obtained from the GF2 dataset, our MMAPP

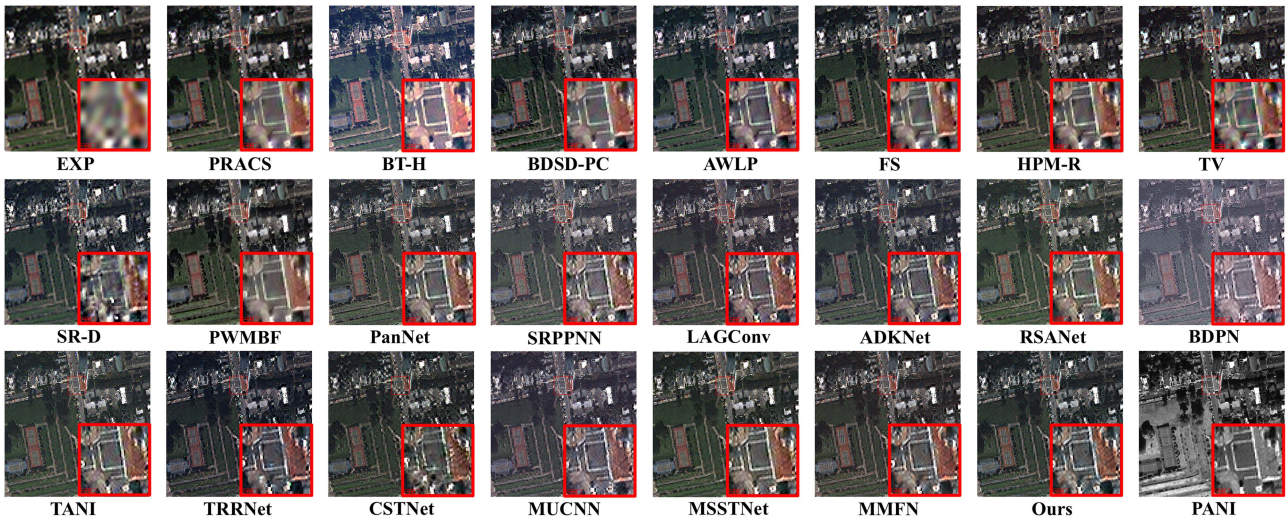

Fig. 9. Qualitative comparison under the FST on QB dataset (selected bands: red, green, and blue).

TABLE III
QUANTITATIVE COMPARISON ON QB DATASET

Comp.	Quantitative evaluation in RST					Quantitative evaluation in FST				
	ERGAS↓(±std)	SAM↓(±std)	sCC↑(±std)	$Q2n$↑(±std)	RASE↓(±std)	D_s↓(±std)	D_λ↓(±std)	D_λ^F↓(±std)	QNR↑(±std)	HQNR↑(±std)
EXP	12.0189±1.5432	8.5575±1.7025	0.8701±0.0264	0.5782±0.0774	46.6474±8.3320	0.0529±0.0156	0.0001±0.0000	0.0348±0.0071	0.9470±0.0156	0.9142±0.0190
CS-based methods										
PRACS	8.4567±0.7392	8.2863±1.7793	0.9269±0.0147	0.7861±0.1029	32.8178±4.6692	0.1399±0.0249	0.0262±0.0116	0.1218±0.0230	0.8375±0.0253	0.7556±0.0362
BT-H	7.4939±0.5895	7.2981±1.3674	0.9416±0.0104	0.8295±0.0963	29.1453±3.8422	0.1244±0.0506	0.1457±0.0458	0.2344±0.0738	0.7489±0.0690	0.6731±0.0863
BDSD-PC	7.6084±0.5744	8.1813±1.7780	0.9324±0.0131	0.8287±0.0966	29.6034±3.9402	0.1415±0.0306	**0.0261±0.0142**	0.1948±0.0334	0.8362±0.0346	0.6919±0.0464
MRA-based methods										
AWLP	7.7451±0.7005	8.3039±1.8186	0.9266±0.0204	0.8237±0.0982	29.9031±4.1927	0.0975±0.0290	0.0467±0.0203	**0.0398±0.0100**	0.8605±0.0356	0.8668±0.0333
FS	7.4454±0.5512	7.8662±1.6282	0.9378±0.0124	0.8336±0.0932	29.0174±3.7189	0.1146±0.0225	0.0342±0.0199	**0.0454±0.0149**	0.8554±0.0356	0.8454±0.0303
HPM-R	7.7172±1.8120	7.8345±1.6667	0.9424±0.0107	0.8398±0.1007	29.4088±3.6581	0.1052±0.0239	0.0319±0.0209	0.0462±0.0145	0.8663±0.0324	0.8536±0.0303
VO-based methods										
TV	7.7524±0.6749	7.5653±1.4959	0.9194±0.0201	0.8204±0.0876	30.6846±4.4723	0.0939±0.0213	0.0303±0.0120	0.0565±0.0140	0.8788±0.0289	0.8551±0.0296
SR-D	8.3004±1.3787	8.4884±1.7668	0.9237±0.0266	0.8038±0.1080	32.3252±6.6068	0.0260±0.0112	**0.0260±0.0251**	**0.0284±0.0101**	**0.9520±0.0280**	**0.9496±0.0148**
PWMBF	8.4010±0.6818	8.9367±1.8737	0.9066±0.0169	0.7880±0.0970	32.7361±4.3937	0.1223±0.0190	0.0353±0.0165	0.1348±0.0320	0.8469±0.0297	0.7596±0.0368
Single-branch structures										
PanNet	6.9987±0.5552	7.9971±1.4499	0.9320±0.0155	0.8413±0.1040	27.3428±3.4054	0.1262±0.1265	0.0563±0.0659	0.0763±0.0263	0.8321±0.1465	0.8073±0.1213
SRPPNN	4.3333±0.2882	5.1869±0.8108	0.9789±0.0040	0.9140±0.1125	16.5415±1.3385	0.0279±0.0193	0.0384±0.0327	0.0601±0.0204	0.9353±0.0465	0.9140±0.0339
LAGConv	**3.8610±0.3068**	**4.7172±0.7511**	**0.9837±0.0035**	**0.9315±0.0895**	**14.9851±1.5373**	**0.0201±0.0098**	0.0406±0.0312	0.0648±0.0173	0.9402±0.0362	0.9164±0.0224
ADKNet	3.9416±0.3213	4.9035±0.8194	0.9820±0.0039	0.9299±0.0880	15.2746±1.6387	0.0293±0.0195	0.0296±0.0285	0.0791±0.0102	0.9423±0.0413	0.8941±0.0247
RSANet	3.8686±0.2722	4.7659±0.7969	0.9834±0.0037	0.9301±0.0916	15.0549±1.6040	0.0240±0.0141	0.0348±0.0235	0.0620±0.0129	0.9422±0.0355	0.9194±0.0276
Dual-branch structures										
BDPN	5.4729±0.6364	6.2399±1.1646	0.9609±0.0093	0.8965±0.0939	21.2544±3.1819	0.0273±0.0147	0.0588±0.0546	0.0721±0.0348	0.9162±0.0638	0.9029±0.0438
TANI	5.2219±0.3829	5.3832±0.9007	0.9749±0.0046	0.9039±0.0961	20.1670±2.2983	0.0920±0.0262	0.0268±0.0251	0.0730±0.0179	0.8841±0.0452	0.8420±0.0366
TRRNet	4.8495±0.9964	5.0151±0.9622	0.9786±0.0078	0.9146±0.0866	18.8421±4.3051	0.0304±0.0103	0.0458±0.0389	0.0788±0.0174	0.9253±0.0406	0.8933±0.0214
CSTNet	**3.8246±0.2780**	**4.6744±0.7626**	**0.9849±0.0033**	**0.9317±0.0901**	**14.9135±1.5062**	0.0495±0.0178	0.0489±0.0415	0.0498±0.0114	0.9038±0.0385	0.9031±0.0174
UNet-based multiscale structures										
MUCNN	5.3620±1.0684	5.2130±1.0031	0.9740±0.0094	0.9021±0.0930	20.7546±4.6108	0.0300±0.0140	0.0348±0.0344	0.0838±0.0196	0.9366±0.0450	0.8889±0.0288
MSSTNet	5.5252±0.8150	5.3956±0.9889	0.9715±0.0079	0.8972±0.0959	21.3443±3.8320	0.0860±0.0318	**0.0200±0.0193**	0.0822±0.0200	0.8963±0.0473	0.8322±0.0362
MMFN	5.7966±1.1541	5.5140±1.0623	0.9720±0.0107	0.8995±0.0858	22.5006±5.1659	**0.0156±0.0081**	0.0333±0.0250	0.0466±0.0250	**0.9517±0.0293**	**0.9386±0.0285**
The proposed multibranch pyramid structure										
Ours	**3.7467±0.2896**	**4.5658±0.7554**	**0.9864±0.0030**	**0.9332±0.0911**	**14.5086±1.5903**	**0.0224±0.0187**	0.0358±0.0194	0.0474±0.0150	**0.9429±0.0325**	**0.9314±0.0271**

The red, green, and blue highlights indicate the best, the second-best, and the third-best values, respectively.

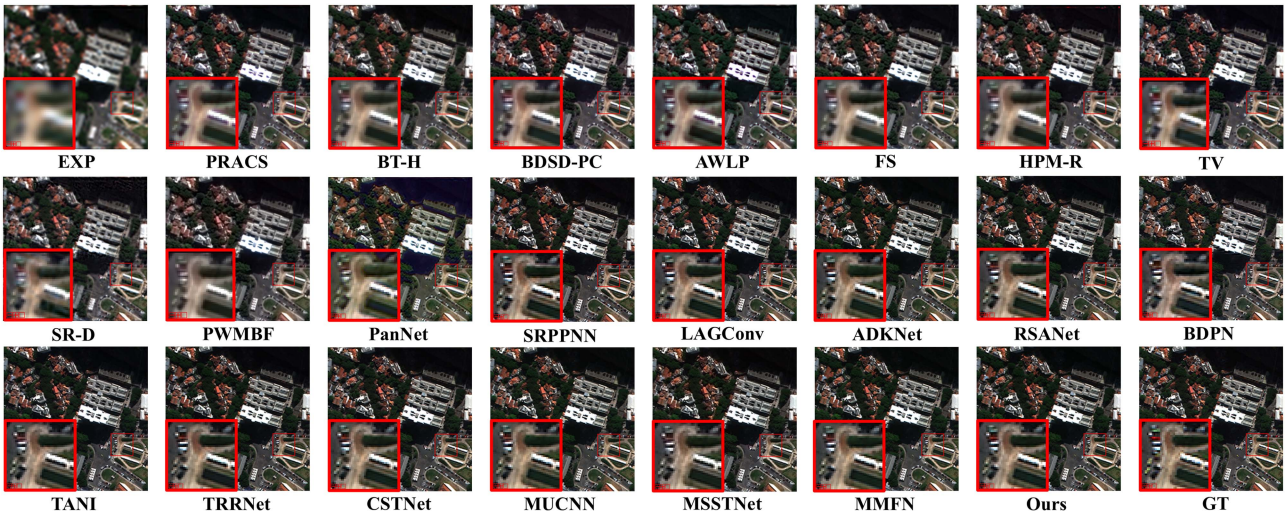

Fig. 10. Qualitative comparison under the RST on WV3 dataset (selected bands: red, green, and blue).

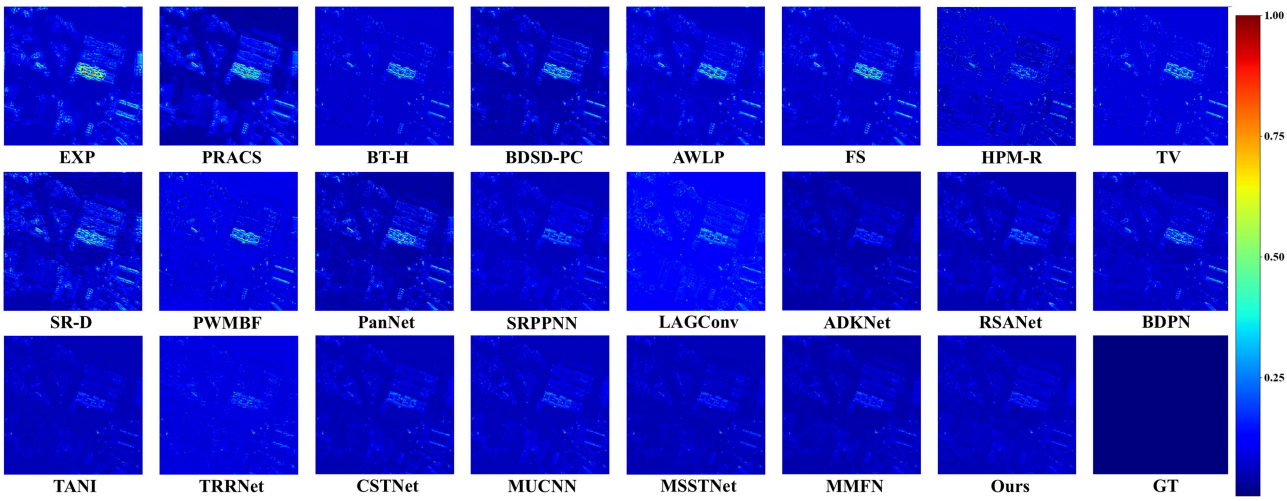

Fig. 11. AEMs that correspond to each method are depicted in Fig. 10.

not only achieves the highest values but also the smallest standard deviation values. This implies that our methodology achieves the highest degree of stability and optimal fusion performance. In FST, our method achieves the third-best value on the QNR and HQNR, after SR-D and MMFN. In contrast to the RST, where there is a reference image available, the five non-reference metrics in the FST use customized spectral and spatial references to calculate the degree of distortion in the sharpened image. For example, the spectral distortion metrics D_λ and D_λ^F use the interpolated MSI image as a spectral reference, i.e., EXP. However, the qualitative evaluations indicate that the spectral distributions of HRMSI and EXP are inconsistent, which makes the spectral distortion indices unavoidably introduce bias. On the other hand, some studies [4], [50] have demonstrated that spectral distortion metrics may treat the injected spatial details as spectral distortions. These factors partially explain the fact that the VO-based method SR-D exhibits poor spatial quality in qualitative evaluations, whereas it achieves the best scores in quantitative evaluations. To summarize, the combined performance in both qualitative and quantitative experiments demonstrates that our methodology is advantageous and competitive.

F. Experiments on WV3 Dataset

To further assess the generalizability, we conduct experiments on the eight-band WV3 dataset. Fig. 10 shows the RST results on the WV3 dataset. As the sharpened scene becomes more and more complex, i.e., there are more and more targets in the scene and they are more and more granular, it can be observed that the performance of the traditional methods demonstrates obvious limitations. All traditional methods exhibit varying degrees of spatial and spectral distortion, with SR-D exhibiting the most serious spatial distortion. Except for PanNet, BDPN, and TANI, the rest of the DL-based methods exhibit good visual results. The corresponding AEMs are provided for a more intuitive visualization, as illustrated in Fig. 11. It is evident that ADKNet

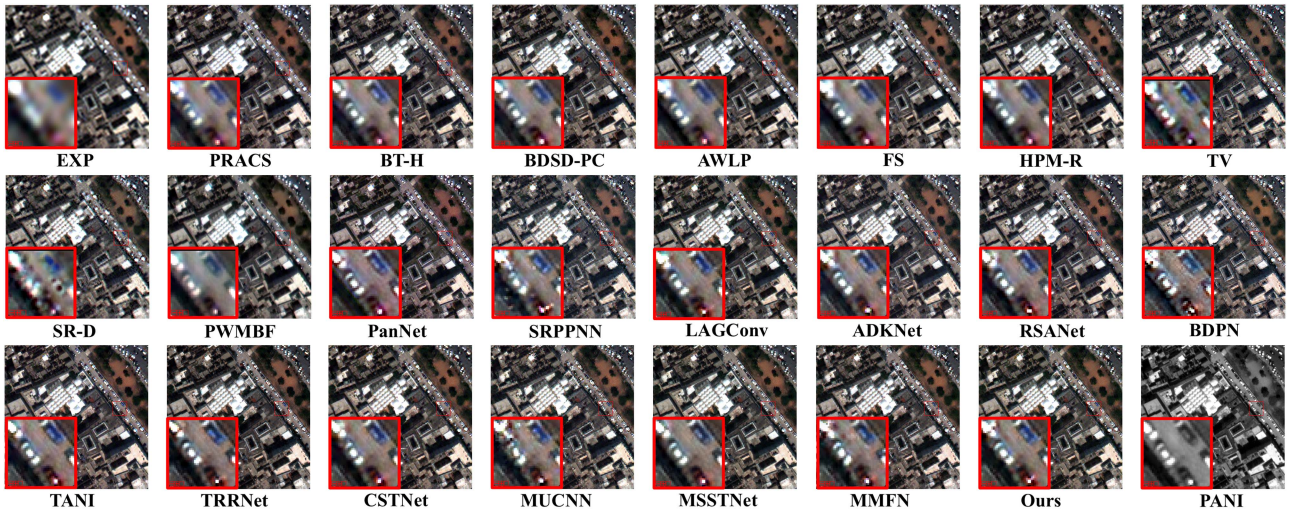

Fig. 12. Qualitative comparison under the FST on WV3 dataset (selected bands: red, green, and blue).

and MMFN correspond to AEMs with darker hues, whereas the proposed MMAPP exhibits the least residuals, especially in the flower bed area. The results of the FST are as shown in Fig. 12. It is noticed that the CS-based methods exhibit obvious spectral distortion, while the MRA-based methods all exhibit evident spatial distortion. Among the DL-based methods, TANI and MSSTNet perform best in terms of spatial quality, but they suffer from slight spectral distortion. SRPPNN, BDPN, and MUCNN show poor spatial quality. In contrast, the proposed MMAPP exhibits minor spatial distortion, specifically on the white sedan doors. As a whole, the MMAPP better balances spatial and spectral preservation.

Table IV shows the corresponding quantitative results. The highest values and the smallest standard deviations in RST show that our method is more capable of handling the more challenging scenario. In FST, the proposed MMAPP does not achieve the best scores on the two spectral distortion metrics (D_λ and D_λ^F). As previously mentioned, we have conducted an analysis of certain irrationalities associated with these two metrics in determining the spectral reference image. Our method achieves the highest score on the QNR, which implies that our method better balances spatial texture and spectral information. SRPPNN achieves the highest scores on HQNR, but its performance on qualitative results is very poor. The quality of sharpened images can be approximated by these metrics, and ought to evaluate them thoroughly by combining qualitative and quantitative evaluations. By combining qualitative and quantitative evaluations, it can be inferred that the proposed MMAPP achieves the optimal balance between preservation of spatial textures and spectral information.

G. Ablation Experiments

The proposed MMAPP primarily comprises of the multibranch pyramid structure, three AEIMs, and multiscale reconstruction constraints. These components are validated and analyzed accordingly in this section. For ease of comprehension, we refer to the scale variant in the multibranch pyramid structure as S-Variant; the variant in the ablation experiment for the three AEIMs as M-Variant; the variant related to each loss function in the multiscale reconstruction constraints as L-Variant; and the variant related to the coefficients of each loss function as C-Variant.

1) Multibranch Pyramid Structure: This part is used to verify the impact of the number of preset scales within the pyramid structure on the overall performance. We set up four scale variants by controlling the number of samples, including one scale, two scales, three scales, and four scales. The qualitative experimental results for each variant are shown in Fig. 13(a). S-Variant I and S-Variant II exhibit obvious spectral distortions as a result of not full exploiting the multiscale information, resulting in a more rigid fusion process. However, the fusion performance suffers degradation as the number of scales becomes four. The scale ratio between PANI and MSI is four. The scale of the input features of the MS branch and the fusion branch is half of the MSI scale, as we additionally execute a downsampling operation on the MSI, resulting in some information lost during the sampling process. It can be observed from the corresponding AEMs that, although the performance is degraded compared to S-Variant III, S-Variant IV still outperforms S-Variant II in preserving the local texture, proving the necessity of utilizing multiscale information. Table V illustrates the quantitative results consistent with the qualitative analysis. When the ratio between PANI and MSI is four, the three scales are capable of establishing the optimal connection between the two images, thereby achieving high fidelity fusion outcomes.

2) AEIMs: The purpose of this section is to validate the efficacy of each AEIM across three distinct branches, namely TSAM, SICM, and AFM. Four variants are devised, comprising a baseline version devoid of any AEIMs, followed by the gradual incorporation of TSAM, SICM, and AFM. The qualitative experimental results for each variant are shown in Fig. 13(b), where M-Variant I is baseline, and M-Variant II, M-Variant III, and M-Variant IV are the results after sequential addition of TSAM,

TABLE IV
QUANTITATIVE COMPARISON ON WV3 DATASET

Comp.	Quantitative Evaluation in RST					Quantitative Evaluation in FST				
	ERGAS↓(±std)	SAM↓(±std)	sCC↑(±std)	$Q2n$↑(±std)	RASE↓(±std)	D_s↓(±std)	D_λ↓(±std)	D_λ^F↓(±std)	QNR↑(±std)	HQNR↑(±std)
EXP	7.1354±1.5641	5.8351±1.6720	0.9226±0.0290	0.6027±0.0889	22.2931±6.3631	0.0340±0.0133	0.0000±0.0000	0.0232±0.0064	0.9660±0.0133	0.9437±0.0166
CS-based methods										
PRACS	5.2059±1.4694	5.6081±1.6734	0.9495±0.0149	0.7721±0.1122	14.9953±4.9185	0.0757±0.0264	0.0136±0.0087	0.0368±0.0134	0.9120±0.0329	0.8906±0.0351
BT-H	4.5107±1.2973	4.8984±1.2695	0.9586±0.0123	0.8182±0.0993	13.9308±4.7883	0.1001±0.0375	0.0268±0.0198	0.0574±0.0226	0.8764±0.0516	0.8489±0.0511
BDSD-PC	4.6499±1.4270	5.4643±1.6708	0.9552±0.0117	0.8117±0.1036	13.8559±4.7681	0.0912±0.0362	**0.0129±0.0096**	0.0624±0.0228	0.8973±0.0432	0.8528±0.0509
MRA-based methods										
AWLP	4.6971±1.3285	5.2762±1.3649	0.9508±0.0135	0.8066±0.1012	14.1126±4.7256	0.0761±0.0312	0.0219±0.0155	0.0163±0.0078	0.9040±0.0425	0.9090±0.0366
FS	4.6450±1.4062	5.3228±1.6112	0.9560±0.0114	0.8177±0.0989	13.7713±4.8366	0.0851±0.0307	0.0197±0.0168	0.0197±0.0076	0.8973±0.0432	0.8971±0.0352
HPM-R	5.6107±3.1016	5.3694±1.6011	0.9431±0.0254	0.8142±0.1000	13.4847±4.7394	0.0855±0.0308	0.0205±0.0172	0.0205±0.0080	0.8963±0.0436	0.8960±0.0355
VO-based methods										
TV	4.9836±1.3213	6.3392±1.5459	0.9396±0.0235	0.7855±0.1207	15.2599±4.2853	0.0686±0.0256	0.0173±0.0115	0.0234±0.0059	0.9155±0.0336	0.9097±0.0285
SR-D	5.3739±1.5781	6.3255±1.5472	0.9456±0.0210	0.7650±0.1316	16.5443±4.6934	0.0472±0.0235	0.0220±0.0165	0.0302±0.0140	0.9318±0.0287	0.9240±0.0241
PWMBF	5.3958±1.5718	6.6000±1.6982	0.9275±0.0162	0.7863±0.1088	15.8269±4.8340	0.0898±0.0267	0.0230±0.0152	0.0590±0.0180	0.8896±.0377	0.8569±0.0387
Single-branch structures										
PanNet	5.2335±1.5127	6.9395±1.3863	0.9321±0.0355	0.7491±0.1464	15.5878±4.1562	0.0933±0.0348	0.0251±0.0192	0.0627±0.0237	0.8845±0.0485	0.8503±0.0482
SRPPNN	2.3587±0.5462	3.1867±0.5588	0.9859±0.0056	0.8927±0.0889	7.3885±2.0307	**0.0355±0.0145**	0.0338±0.0188	**0.0164±0.0072**	0.9320±0.0246	**0.9487±0.0165**
LAGConv	2.2435±0.5154	3.1007±0.5211	0.9870±0.0053	0.9039±0.0864	7.0068±1.8672	0.0572±0.0192	0.0178±0.0126	0.0258±0.0108	0.9261±0.0256	0.9185±0.0230
ADKNet	2.2905±0.5034	3.1382±0.5618	0.9860±0.0053	0.9051±0.0834	7.1265±1.8758	0.0484±0.0190	0.0214±0.0164	0.0187±0.0073	0.9312±0.0236	0.9338±0.0181
RSANet	2.2904±0.4945	3.1252±0.5458	0.9864±0.0056	0.8995±0.0882	7.1681±1.8556	0.0498±0.0232	0.0203±0.0148	0.0232±0.0096	0.9309±0.0270	0.9281±0.0244
Dual-branch structures										
BDPN	3.3709±0.7864	4.4479±0.9630	0.9623±0.0160	0.8438±0.1075	9.9593±2.6070	0.0532±0.0167	0.0237±0.0160	0.0241±0.0091	0.9245±0.0281	0.9241±0.0213
TANI	2.6869±0.5628	3.6054±0.6695	0.9781±0.0108	0.8912±0.0879	8.0661±2.3056	0.0849±0.0319	0.0157±0.0151	0.0353±0.0144	0.9012±0.0433	0.8832±0.0415
TRRNet	2.3638±0.5133	3.2627±0.5607	0.9847±0.0058	0.9045±0.0858	7.3628±1.9141	0.0735±0.0236	0.0171±0.0150	0.0364±0.0139	0.9110±0.0353	0.8931±0.0326
CSTNet	2.3211±0.4828	3.2271±0.6020	0.9862±0.0055	0.9014±0.0853	7.1889±1.8010	0.0680±0.0221	0.0179±0.0122	0.0206±0.0072	0.9155±0.0307	0.9130±0.0261
UNet-based multiscale structures										
MUCNN	2.3646±0.5424	3.1591±0.6193	0.9874±0.0052	0.9091±0.0828	7.3216±2.0291	0.0406±0.0214	0.0264±0.0210	0.0225±0.0099	0.9339±0.0267	0.9377±0.0210
MSSTNet	2.7775±0.6243	3.6973±0.7252	0.9734±0.0126	0.8965±0.0825	8.3879±2.4221	0.0860±0.0318	0.0200±0.0193	0.0372±0.0166	0.8963±0.0473	0.8804±0.0431
MMFN	2.2488±0.4954	2.9772±0.5516	0.9883±0.0052	0.9138±0.0813	7.0415±1.8871	0.0446±0.0207	0.0224±0.0141	0.0200±0.0077	0.9340±0.0242	0.9363±0.0221
The proposed multibranch pyramid structure										
Ours	**2.1664±0.4594**	**2.9499±0.5142**	**0.9895±0.0048**	**0.9141±0.0824**	**6.8507±1.7779**	0.0387±0.0167	0.0201±0.0104	0.0216±0.0092	**0.9420±0.0207**	0.9406±0.0188

The red, green, and blue highlights indicate the best, the second-best, and the third-best values, respectively.

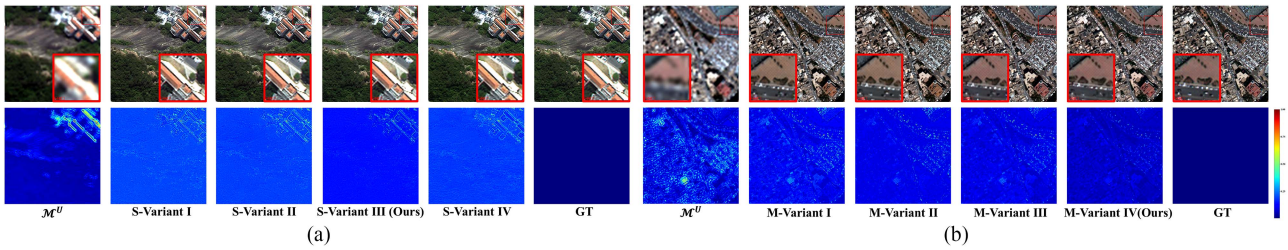

Fig. 13. Ablation experimental results on the number of scales in multibranch pyramid structure and AEIM in each branch. (a) Ablation experimental results on the number of scales in multibranch pyramid structure. (b) Ablation experimental results on three AEIMs (selected bands: red, green, and blue).

SICM, and AFM respectively. Even without using any AEIMs, M-Variant I (baseline) still displays the good visual effect. AEMs are able to visualize the differences between variants more intuitively. It can be clearly seen that the residuals in the AEM corresponding to M-Variant II become less after the addition of TSAM. After the addition of SICM, the spectral distribution of M-Variant III tends to agree with GT, while the local hue of the AEM becomes darker. In addition, the residuals of M-Variant III are further reduced compared to M-Variant II due to the additional consideration of complementary spatial details in diverse source images by SICM. Finally, the spatial texture and spectral distribution of the fused images are further enhanced by the addition of AFM, which is attributed to the ability of AFM to focus on both the spatial and spectral dimensions. Table V illustrates the quantitative results consistent with the qualitative analysis. The best score for the spectral distortion index (D_λ) of

TABLE V
QUANTITATIVE RESULTS OF ABLATION EXPERIMENTS ON WV3 DATASET

Variants					Quantitative evaluation in RST					Quantitative evaluation in FST				
					ERGAS↓	SAM↓	sCC↑	$Q2n$↑	RASE↓	D_s↓	D_λ↓	D_λ^F↓	QNR↑	HQNR↑
S-Variant	One	Two	Three	Four	Experiments on the number of scales in multibranch pyramid structure									
I	✓	×	×	×	2.5127	3.3466	0.9830	0.9011	8.0390	0.0685	0.0184	0.0343	0.9147	0.8997
II	×	✓	×	×	2.2622	3.0830	0.9871	0.9056	7.2033	0.0557	0.0140	0.0298	0.9313	0.9164
III(ours)	×	×	✓	×	**2.1664**	**2.9499**	**0.9895**	**0.9141**	**6.8507**	**0.0387**	0.0201	**0.0216**	**0.9420**	**0.9406**
IV	×	×	×	✓	2.3286	3.0501	0.9883	0.9090	7.3498	0.0533	**0.0115**	0.0357	0.9359	0.9131
M-Variant	Baseline	TSAM	SICM	AFM	Experiments on three AEIMs									
I	✓	×	×	×	2.7898	3.7846	0.9680	0.8951	8.3938	0.0745	0.0296	0.0774	0.8983	0.8541
II	✓	✓	×	×	2.4203	3.2856	0.9847	0.9029	7.6857	0.0634	0.0206	0.0323	0.9175	0.9064
III	✓	✓	✓	×	2.3585	3.1475	0.9851	0.9073	7.4217	0.0518	**0.0156**	0.0249	0.9335	0.9246
IV(ours)	✓	✓	✓	✓	**2.1664**	**2.9499**	**0.9895**	**0.9141**	**6.8507**	**0.0387**	0.0201	**0.0216**	**0.9420**	**0.9406**
L-Variant	$\mathcal{L}_{ori}$	$\mathcal{L}_{d2}$	$\mathcal{L}_{d4}$	$\mathcal{L}_{spe}$	Experiments on the loss function									
I	✓	×	×	×	2.4952	3.5773	0.9842	0.9023	7.5385	0.0459	0.0136	0.0289	0.9412	0.9265
II	✓	✓	×	×	2.2352	3.0515	0.9870	0.9099	7.0362	0.0486	**0.0118**	0.0224	0.9403	0.9283
III	✓	✓	✓	×	2.1992	2.9620	0.9882	0.9126	6.9542	0.0422	0.0139	0.0276	**0.9446**	0.9314
IV(ours)	✓	✓	✓	✓	**2.1664**	**2.9499**	**0.9895**	**0.9141**	**6.8507**	**0.0387**	0.0201	**0.0216**	0.9420	**0.9406**
C-Variant	λ_1	λ_2	λ_3	λ_4	Experiments with coefficients in the loss function									
I	1	1	1	0.01	2.2859	3.0796	0.9881	0.9045	7.2691	0.0508	0.0113	0.0245	0.9386	0.9261
II	0.1	0.2	0.7	0.01	2.2580	3.0449	0.9883	0.9089	7.1549	0.0475	**0.0098**	0.0221	0.9433	0.9314
III	0.1	0.3	0.6	0.01	2.2060	3.0012	0.9888	0.9116	7.0442	0.0401	0.0135	**0.0214**	**0.9470**	0.9394
IV(ours)	0.2	0.3	0.5	0.01	**2.1664**	**2.9499**	**0.9895**	**0.9141**	**6.8507**	**0.0387**	0.0201	0.0216	0.9420	**0.9406**
V	0.2	0.3	0.5	0.05	2.2180	3.0082	0.9888	0.9054	7.2341	0.0542	0.0135	0.0230	0.9332	0.9241
VI	0.2	0.3	0.5	0.10	2.2555	3.0380	0.9884	0.9090	7.1341	0.0536	0.0132	0.0252	0.9341	0.9226
VII	0.2	0.3	0.5	0.15	2.2637	3.0935	0.9882	0.8992	7.2764	0.0540	0.0117	0.0236	0.9351	0.9238

The red highlights indicate the best values.

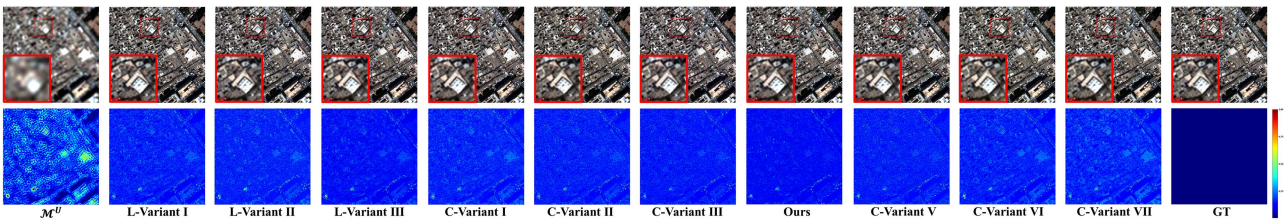

Fig. 14. Ablation experimental results on the proposed multiscale reconstruction constraint (selected bands: red, green, and blue).

M-Variant III was achieved after the addition of SICM, which demonstrates the ability of SICM to focus on improving the spectral distribution of the fused image. Moreover, D_λ instead decreases after the addition of AFM, which is due to the fact that AFM aims to achieve a balance between spatial and spectral preservation in order to improve the overall quality of the image.

3) Multiscale Reconstruction Constraint: The purpose of this section is to validate the efficacy of each component in the proposed loss function. Similarly, four variants are designed by gradually introducing constraints at different scales. The qualitative experimental results for each variant are shown in Fig. 14. It can be clearly seen from the AEMs that the residuals of the fused images are gradually reduced with the gradual introduction of the multiscale constraints. In addition, the fidelity of the fused image is further improved after the introduction of $\mathcal{L}_{\text{spe}}$. Table V illustrates the quantitative results. QNR is obtained by weighting D_s and D_λ. The proposed MMAPP fails to achieve the best scores on D_λ and QNR. This is due to the fact that the introduction of multiscale constraints can make full use of the information available at different scales, which in turn improves the spatial quality of the image. However, it has been demonstrated that QNR could potentially regard the increased detail as the spectral distortion [50], which in turn results in the decrease D_λ and QNR.

The coefficients corresponding to the components of multiscale constraints are subject to ablation experiments. There are two distinct parts to these experiments: sensitivity experiments on the coefficients that correspond to the three preset scales (λ_1, λ_2, and λ_3), and sensitivity experiments on the coefficients that correspond to $\mathcal{L}_{\text{spe}}$ (λ_4). The experimental results are shown in the right part of Fig. 14 and in the bottom part of Table V. C-Variant I–IV are coefficient variants corresponding to different scale constraints. It is evident that, when the same weights are assigned to different scale constraints, the fusion results of C-Variant I demonstrate the loss of a large amount of critical information, such as building edge regions. It is well known that the purpose of pansharpening is to exploit the spatial information from PANI to sharpen the MSI to the original scale of PANI. Furthermore, the GT at the original scale retains the most complete information, whereas the downsampling operation results in some information being lost. Therefore, the original scale should be given a greater weighting coefficient. The fusion results of C-Variant IV exhibit the sharpest building edges and the least residuals when λ_4 is adjusted to 0.5. It is worth noting that, in FST, C-Variant II achieves the best score on D_λ, making its QNR score slightly higher than that of C-Variant IV. C-Variant III achieves the best scores on D_λ^F and QNR. This is due to the higher weighting coefficient assigned to the original scales allow

TABLE VI
QUANTITATIVE COMPARISON OF EFFICIENCY ON WV3 DATASET

Effi.	PanNet	SRPPNN	LAGConv	ADKNet	RSANet	BDPN	TANI	TRRNet	CSTNet	MUCNN	MSSTNet	MMFN	Ours
NoPs (M)	0.15	1.72	0.30	0.06	**0.039**	2.96	0.042	5.23	0.12	2.32	25.35	2.37	0.41
FLOPs (G)	2.50	10.59	**0.26**	0.36	0.47	30.28	0.34	3.75	0.96	3.49	24.94	10.85	0.74
Time (s)	0.029	0.034	0.032	0.036	0.052	0.033	**0.002**	0.085	0.033	0.028	0.079	0.009	0.096

The red highlights indicate the best values.

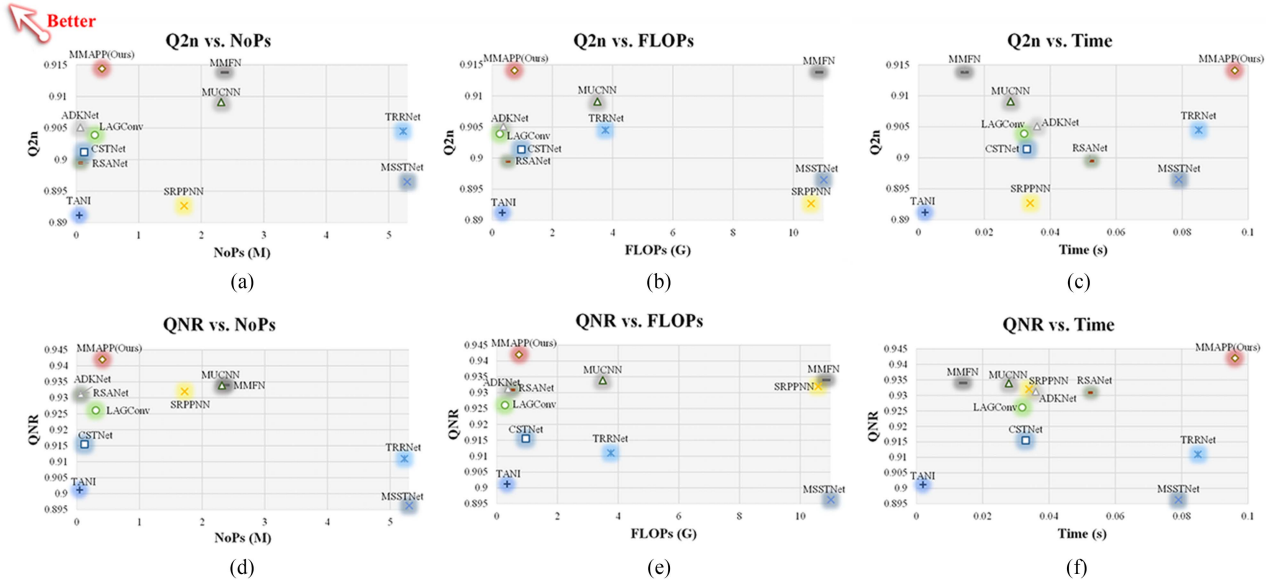

Fig. 15. Comparison of the efficiency of existing state-of-the-art DL-based methods. The first line shows the relationship between reduced-scale performance and efficiency, and the second line shows the relationship between full-scale performance and efficiency. The first column shows the relationship between performance and NoPs, the second column shows the relationship between performance and the computational complexity, and the second column shows the relationship between performance and the time complexity.

the fused images to maintain a more consistent spectral distribution over the GT. However, some of the multiscale information is lost, resulting in a poor performance on D_s. C-Variant IV–VII are coefficient variants of the corresponding coefficients of $\mathcal{L}_{\text{spe}}$. As the value of λ_4 increases, the fusion results show a spectral distribution that is more consistent with $\mathcal{M}^U$, as evidenced by an overall decreasing trend in D_s. However, the spatial information in $\mathcal{M}^U$ is limited, and assigning larger weighting coefficients to $\mathcal{L}_{\text{spe}}$ causes the fusion results to perform poorly in terms of spatial information preservation. As a result, we set λ_1, λ_2, λ_3, and λ_4 to 0.2, 0.3, 0.5, and 0.01, respectively.

H. Efficiency Analysis

We validate the efficiency of each method in three aspects: the number of parameters (NoPs), computational complexity (FLOPs), and time complexity (testing time). Table VI shows the efficiency test results. Specifically, our method is second to PanNet, LAGConv, ADKNet, RSANet, TANI, and CSTNet in terms of NoPs, and second only to LAGConv, ADKNet, RSANet, and TANI in terms of the amount of computation. Furthermore, taking into account the significance of sharpening performance, we affix model performance with three efficiency aspects to provide a comprehensive assessment. Our method exhibits the optimal performance in both R-Test and F-Test, as shown in Fig. 15. In terms of performance vs. NoPs, the proposed MMAPP achieves the best balance. In terms of performance vs. computational complexity, our method still achieves the optimal balance. In terms of performance vs. time, MMFN achieves the best balance. In general, the proposed MMAPP achieves the optimal balance between fusion performance, NoPs, and computational volume, but performs slightly less well in terms of test time. In practical applications, lower parameter counts and computations can reduce the algorithm's requirements on hardware specifications, thereby achieving high-fidelity fusion within a limited hardware environment. Furthermore, in practical scenarios, the image can be segmented and subsequently, a pipeline for fusion processing can be constructed to minimize the processing duration [51], [52]. As a result, we believe that our methodology has potential in real-world fusion scenarios.

V. DISCUSSION

The design principles followed by existing DL methods include single-branch structure, dual-branch structure, and multiscale structure. These structures fail to adequately exploit the scale disparities and heterogeneity between the diverse source data, resulting in spatial or spectral distortions in the fused images. To alleviate this effect, we propose an efficient multibranch pyramid structure for pansharpening, comprising PAN, MS, and fusion branches. In addition, each branch contains a pyramid structure that efficiently and seamlessly integrates data flows

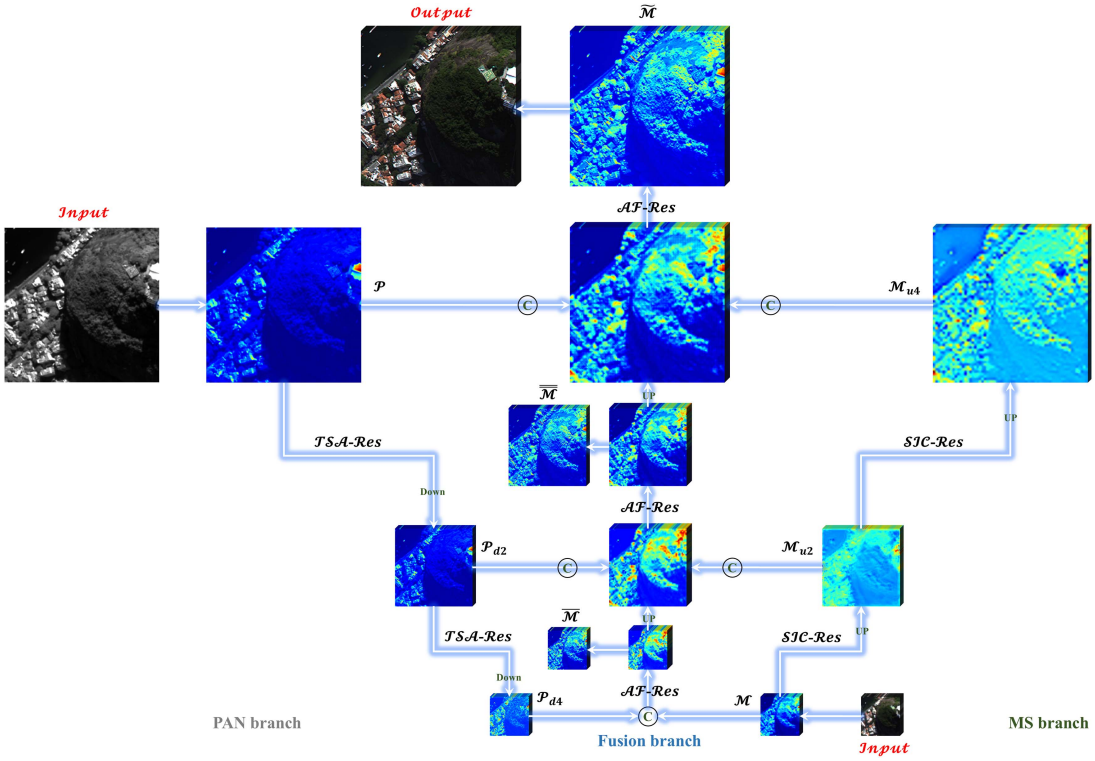

Fig. 16. Visualization of intermediate features in the proposed MMAPP.

at various scales across distinct branches. In the meantime, we have developed three specialized AEIMs for each branch. These AEIMs are specifically designed to cater to diverse sources and distinct stages of the pansharpening process, enabling the extraction and fusion of more advantageous information. Finally, multiscale constraints are developed to obtain high-fidelity fused images. The results of Section IV-D–IV-F demonstrate that the proposed MMAPP is capable of achieving more realistic fusion outcomes with fine-grained spatial texture and high-fidelity spectral distribution. Section IV-G validates the efficacy of the proposed methodology. Furthermore, the efficiency analysis section shows that our method effectively restricts the model parameters and computational complexity, which is attributed to the highly cohesive design of the proposed fusion structure. In addition, we visualize the intermediate features of the model. An instance with images from the WV3 dataset is delivered, as shown in Fig. 16. It is evident that the proposed MMAPP is capable of efficiently and seamlessly integrating data flows at various scales in distinct branches, while preserving the advantageous information in the heterogeneous data. This allows for high-fidelity fusion.

VI. Conclusion

In this article, we provide a detailed analysis of the scale disparities and heterogeneity of the diverse source data. We propose a multibranch pyramid structure, which can build bridges between diverse source images at various scales. It contains three distinct branches, including the PAN, MS, and fusion branches, which efficiently and seamlessly integrates the data flow in distinct branches by means of the pyramid structure. Furthermore, in order to preserve more advantageous information, we have developed three AEIMs for each branch, namely, the TSAM for the PAN branch, the SICM for the MS branch, and the AFM for the fusion branch. These AEIMs are specifically designed to cater to diverse sources and distinct stages of the pansharpening process. The adaptive weights they generate can be used to extract and fuse more advantageous information. Ultimately, high-fidelity sharpening outcomes are obtained by minimizing the reconstruction errors at various scales in distinct branches. We conduct extensive ablation experiments and comparison experiments on three public datasets, and the qualitative and quantitative results show that the proposed MMAPP can achieve more realistic results. Furthermore, efficacy evaluations further demonstrate that MMAPP has achieved the optimal balance between fusion performance, model complexity, and computational complexity. However, in terms of testing time, the proposed method still exhibits significant room for improvement. In future article, we will further improve the fusion performance and efficacy. Furthermore, we intend to enhance the parallelization of our model, for instance, by parallelizing the data flows at various scales in distinct branches, in order to more efficiently reflect actual geographical characteristics and establish a solid foundation for downstream tasks and real-world applications.

Data Availability

Data link has been shared in the article. All codes will be available at https://github.com/JUSTM0VE0N/MMAPP.

REFERENCES

[1] M. Shimoni, R. Haelterman, and C. Perneel, "Hypersectral imaging for military and security applications: Combining myriad processing and sensing techniques," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 101–117, Jun. 2019.

[2] C. Kyrkou and T. Theocharides, "EmergencyNet: Efficient aerial image classification for drone-based emergency monitoring using atrous convolutional feature fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1687–1699, Jan. 2020.

[3] C. Ye et al., "Landslide detection of hyperspectral remote sensing data based on deep learning with constrains," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 12, pp. 5047–5060, Dec. 2019.

[4] G. Vivone et al., "A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 1, pp. 53–81, Mar. 2021.

[5] S. Lolli, L. Alparone, A. Garzelli, and G. Vivone, "Haze correction for contrast-based multispectral pansharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2255–2259, Dec. 2017.

[6] 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.

[7] J. Choi, K. Yu, and Y. Kim, "A new adaptive component substitutionbased satellite image fusion by using partial replacement," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 1, pp. 295–309, Jan. 2011.

[8] T. Zhao et al., "Artificial intelligence for geoscience: Progress, challenges and perspectives," *Innovation*, vol. 5, no. 5, 2024, Art. no. 100691.

[9] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens.*, vol. 8, no. 7, 2016, Art. no. 594.

[10] J. Cai and B. Huang, "Super-resolution-guided progressive pansharpening based on a deep convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5206–5220, Jun. 2021.

[11] Z.-R. Jin, T.-J. Zhang, T.-X. Jiang, G. Vivone, and L.-J. Deng, "LAGConv: Local-context adaptive convolution kernels with global harmonic bias for pansharpening," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 1113–1121.

[12] J. Shi et al., "Source-adaptive discriminative kernels based network for remote sensing pansharpening," in *Proc. Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 1283–1289.

[13] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "PanNet: A deep network architecture for pan-sharpening," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1753–1761.

[14] Y. Zhang, L. Chi, M. Sun, and Y. Ou, "Pan-sharpening using an efficient bidirectional pyramid network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5549–5563, Aug. 2019.

[15] Y. Wang, L. J. Deng, T. J. Zhang, and X. Wu, "SSconv: Explicit spectral-to-spatial convolution for pansharpening," in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 4472–4480.

[16] L. Jian, S. Wu, L. Chen, G. Vivone, R. Rayhana, and D. Zhang, "Multi-scale and multi-stream fusion network for pansharpening," *Remote Sens.*, vol. 15, no. 6, 2023, Art. no. 1666.

[17] K. Malik, C. Robertson, D. C. Braun, and C. Greig, "U-net convolutional neural network models for detecting and quantifying placer mining disturbances at watershed scales," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 104, 2021, Art. no. 102510.

[18] X. Otazu, M. Gonzalez-Audicana, O. Fors, and J. Nunez, "Introduction of sensor spectral response into image fusion methods. Application to wavelet-based methods," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 10, pp. 2376–2385, Oct. 2005.

[19] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "MTF-tailored multiscale fusion of high-resolution MS and PAN imagery," *Photogrammetric Eng. Remote Sens.*, vol. 72, no. 5, pp. 591–596, 2015.

[20] M. Ghahremani, Y. Liu, P. Yuen, and A. Behera, "Remote sensing image fusion via compressive sensing," *ISPRS J. Photogrammetry Remote Sens.*, vol. 152, pp. 34–48, 2019.

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

[22] 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.

[23] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "A new pansharpening algorithm based on total variation," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 318–322, Jan. 2014.

[24] 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.

[25] M. R. Vicinanza, R. Restaino, G. Vivone, M. Dalla 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.

[26] P. Wang, Z. He, B. Huang, M. Dalla Mura, H. Leung, and J. Chanussot, "VOGTNet: Variational optimization-guided two-stage network for multispectral and panchromatic image fusion," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2024.3409563.

[27] R. Wen, L.-J. Deng, Z.-C. Wu, X. Wu, and G. Vivone, "A novel spatial fidelity with learnable nonlinear mapping for panchromatic sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jan. 2023, Art. no. 5401915.

[28] P. Wang et al., "Low rank tensor completion pansharpening based on haze correction," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Jan. 2024, Art. no. 5405720.

[29] Y. Chen, H. Liu, and F. Fang, "A novel pansharpening method based on cross stage partial network and transformer," *Sci. Rep.*, vol. 14, no. 1, 2024, Art. no. 12631.

[30] W. Diao, F. Zhang, H. Wang, J. Sun, and K. Zhang, "Pansharpening via triplet attention network with information interaction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3576–3588, Jan. 2022.

[31] K. Zhang, Z. Li, F. Zhang, W. Wan, and J. Sun, "Pan-sharpening based on transformer with redundancy reduction," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jan. 2022, Art. no. 5513205.

[32] C. Liu, W. Lü, Z. Zhang, X. Feng, and X. Shao, "Recursive self-attention modules based network for panchromatic and multispectral image fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 10067–10083, Oct. 2023.

[33] X. Jia, B. De Brabandere, T. Tuytelaars, and L. Van Gool, "Dynamic filter networks," in *Proc. Neural Inf. Process. Syst.*, 2016, pp. 667–675.

[34] J. Wu, D. Li, Y. Yang, C. Bajaj, and X. Ji, "Dynamic filtering with large sampling field for convnets," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 185–200.

[35] X. Chen, H. Wang, and B. Ni, "X-volution: On the unification of convolution and self-attention," 2021, *arXiv:2016.02253*.

[36] X. Pan et al., "On the integration of self-attention and convolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 815–825.

[37] J. Zhou, V. Jampani, Z. Pi, Q. Liu, and M.-H. Yang, "Decoupled dynamic filter networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6647–6656.

[38] L.-J. Deng et al., "Machine learning in pansharpening: A benchmark, from shallow to deep networks," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 3, pp. 279–315, Sep. 2022.

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

[40] R. H. Yuhas, A. F. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm," in *Proc. Summaries 3rd Annu. JPL Airborne Geosci. Workshop*, 1992, pp. 147–149.

[41] J. Zhou, D. L. Civco, and J. A. Silander, "A wavelet transform method to merge Landsat TM and SPOT panchromatic data," *Int. J. Remote Sens.*, vol. 19, no. 4, pp. 743–757, 1998.

[42] A. Garzelli and F. Nencini, "Hypercomplex quality assessment of multi/hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 662–665, Oct. 2009.

[43] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva, "Multispectral and panchromatic data fusion assessment without reference," *Photogrammetric Eng. Remote Sens.*, vol. 74, no. 2, pp. 193–200, 2008.

[44] B. Aiazzi, L. Alparone, S. Baronti, R. Carlà, A. Garzelli, and L. Santurri, "Full scale assessment of pansharpening methods and data products," in *Proc. SPIE*, Oct. 2014, Art. no. 924402.

[45] B. Aiazzi, L. Alparone, S. Baronti, and A. Garzelli, "Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 10, pp. 2300–2312, Oct. 2002.

[46] G. Vivone, "Robust band-dependent spatial-detail approaches for panchromatic sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6421–6433, Sep. 2019.

[47] G. Vivone, R. Restaino, and J. Chanussot, "A regression-based high-pass modulation pansharpening approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 984–996, Feb. 2018.

[48] G. Vivone, R. Restaino, and J. Chanussot, "Full scale regression based injection coefficients for panchromatic sharpening," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3418–3431, Jul. 2018.
[49] 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.
[50] A. Arienzo, G. Vivone, A. Garzelli, L. Alparone, and J. Chanussot, "Full resolution quality assessment of pansharpening: Theoretical and hands-on approaches," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 3, pp. 168–201, Sep. 2022.
[51] Z. Zhang, Z. Qu, S. Liu, D. Li, J. Cao, and G. Xie, "Expandable on-board real-time edge computing architecture for luojia3 intelligent remote sensing satellite," *Remote Sens.*, vol. 14, no. 15, 2022, Art. no. 3596.
[52] Z. Zhang, W. Lü, X. Shao, G. Xie, C. Liu, and M. Xu, "Task-driven on-board real-time panchromatic multispectral fusion processing approach for high-resolution optical remote sensing satellite," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 7636–7661, Aug. 2023.

Zhiqi Zhang received the B.Sc. degree in geographic information system from the Huazhong Agricultural University, Wuhan, China, in 2006, the B.Eng. degree in computer science and technology from the Huazhong University of Science and Technology, Wuhan, in 2006, and the M.Eng. degree in computer technology and the Ph.D. degree in photogrammetry and remote sensing from the Wuhan University, Wuhan, in 2015 and 2018, respectively.

He is currently an Associate Professor with the School of Computer Science, Hubei University of Technology, Wuhan, China. His research interests include system architecture, algorithm optimization, AI, and high-performance processing of remote sensing.

Chuang Liu is currently working toward the M.Eng. degree in computer technology with the School of Computer Science, Hubei University of Technology, Wuhan, China.

His research interests include remote sensing image processing, multimodal image fusion, low-level vision, machine learning, and deep learning.

Lu Wei received the B.Eng. degree in software engineering from the Wuhan University of Technology, Wuhan, China, in 2006, and the M.Eng. degree in computer technology from the Wuhan University, Wuhan, in 2019.

She was a Senior Engineer with the Huawei Technologies Corporation and is currently an Associate Professor with the School of Information Science and Engineering, Wuchang Shouyi University, Wuhan. Her research interests include image radiance correction, multisensor image fusion, image quality assessment, and intelligent image processing.

Shao Xiang received the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Surveying, Mapping and Remote Sensing Information Engineering, Wuhan University, Wuhan, China, in 2024.

His research interests include artificial intelligence, pattern recognition, and remote sensing image processing.