RAMSF: A Novel Generic Framework for Optical Remote Sensing Multimodal Spatial–Spectral Fusion

Chuang Liu, Zhiqi Zhang, *Member, IEEE*, and Mi Wang, *Member, IEEE*

***Abstract*— Optical remote sensing (ORS) multimodal spatial–spectral fusion (MSF) aims to obtain high-resolution images containing fine-grained spatial details and high-fidelity spectral information, which are crucial for downstream tasks and real-world applications. Existing methods can yield promising outcomes in specific fusion scenarios. However, due to the coarse representation of spatial details and the imprecise alignment of spatial–spectral features, the majority of methods encounter difficulties in balancing spatial and spectral preservation. This imbalance tends to cause distortion in the fused image, rendering these task-specific methods less adaptable and more challenging to apply simultaneously to different ORS-MSF tasks. To address this gap, this article introduces a generic framework that focuses on generalization and practical applicability, rather than solely optimizing the performance of models in a specific fusion task. By conducting a comprehensive analysis of theoretical models and network architectures, we systematically decompose the fusion process into two distinct phases, namely, detail reconstruction and feature alignment. Consequently, the proposed framework consists of two fundamental components: low-frequency-driven high-frequency salient detail reconstruction (LHSDR) and coordinate-modal-guided spatial–spectral feature progressive alignment (CSFPA). In LHSDR, the joint spatial degradation process in various frequency directions from diverse modal data is estimated and salient details are derived in a hierarchical integration, with low frequencies driving high ones. These coupled high-frequency details could lay the foundation for subsequent implementation of high-fidelity fusion. Furthermore, CSFPA estimates the joint spectral degradation process by establishing coordinate-mode relations between coupled high-frequency details and corresponding spectral information in the continuous domain. As a result, high spatial–spectral fidelity fused images are obtained through fine detail reconstruction and accurate feature alignment. Ten datasets derived from three different ORS-MSF tasks are utilized for an experiment, comprising eight simulated and five real test sets. Our proposed methodology demonstrates robust fusion performance and generalization capability on data with different spectral bands at various resolutions. All implementations will be published on our website.**

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

I. Introduction

HIGH-RESOLUTION optical remote sensing (ORS) images are extensively used in downstream tasks such as land classification [1], environmental monitoring [2], and disaster warning [3]. Due to limitations in physical conditions and hardware budget, the spatial resolution and spectral resolution are mutually constrained. A single satellite sensor can only acquire high-spatial-resolution (HR) images with sparse spectral information or low-spatial-resolution (LR) images with abundant spectral information. These limitations hinder the real-world applications of ORS images. Fortunately, ORS multimodal spatial–spectral fusion (MSF) offers a potent remedy for these limitations [4]. The objective of ORS-MSF is to combine these two types of images into images with fine-grained spatial texture and plentiful spectral information. The typical ORS images include panchromatic image (PANI), multispectral image (MSI), and hyperspectral image (HSI), each of which exhibits distinct characteristics. Among them, PANIs possess HR but lack spectral information, MSIs possess several spectral bands with mid-spatial resolution, and HSIs possess dozens or hundreds of spectral bands with spatial resolution being further reduced. For the successful execution of downstream tasks, it is essential to investigate how to integrate the advantageous information from these diverse modal data. To achieve this objective, we concentrate on three distinct types of ORS-MSF tasks, namely, MSI pan sharpening (MSIP), HSI pan sharpening (HSIP), and MSI and HSI fusion (MHIF). As depicted in Fig. 1(a), MSIP and HSIP aim to combine the PANI with the LR MSI/HSI to obtain the HR MSI/HSI, whereas MHIF aims to combine the MSI with the LRHSI to obtain the HRHSI.

The key to ORS-MSF is to simultaneously preserve sufficient spatial information and corresponding spectral information. However, the significant disparities between diverse modal ORS images render it challenging to balance spatial and spectral preservation. As a consequence, this imbalance results in spatial or spectral distortions in the fused images, rendering

Received 24 December 2024; revised 12 February 2025 and 15 March 2025; accepted 17 March 2025. Date of publication 19 March 2025; date of current version 3 April 2025. This work was supported in part by the National Science Fund for Distinguished Young Scholars under Grant 62425102 and in part by the National Key Research and Development Program of China under Grant 2022YFB3902800. *(Corresponding author: Zhiqi Zhang.)*

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

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).

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

Digital Object Identifier 10.1109/TGRS.2025.3552937

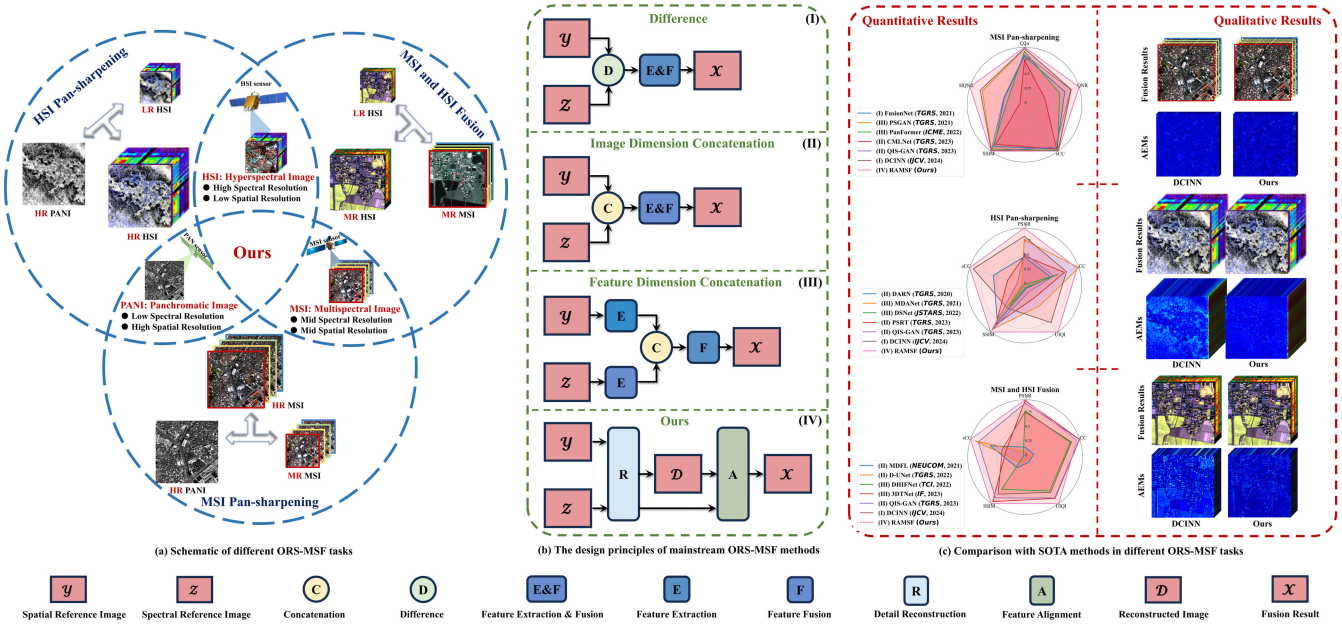

Fig. 1. Schematic illustrates the differences between the proposed RAMSF and the existing MSIP, HSIP, and MHIF algorithms. (a) Schematic of different ORS-MSF tasks. (b) Design principles of mainstream ORS-MSF methods. (c) Comparison with state of the art (SOTA) methods in different ORS-MSF tasks.

task-specific methods less adaptable and more challenging to apply to different ORS-MSF tasks. For most traditional ORS-MSF methods [5], [6], [7], [8], [9], [10], [11], [12], linear transformation is used to extract features, which possesses the advantages of high efficiency and low device dependence. However, they execute fusion in either the spatial or the spectral domains, ignoring information in the other, resulting in distortion in the fusion outcomes. Due to the rapid development of hardware devices, deep learning (DL) has been extensively studied in recent years. The existing DL-based ORS-MSF techniques can be broadly summed up into three distinct phases, namely, image preprocessing, feature extraction, and feature fusion [13], [15], [20]. As shown in Fig. 1(b), with respect to the spatial information representation employed in the image preprocessing phase, the DL-based ORS-MSF techniques encompass difference-based methods [13], [14], [15], image dimension concatenation-based methods [16], [17], [18], [19], and feature dimension concatenation-based methods [20], [21], [22], [23]. During image preprocessing, the majority of DL methods employ coarse spatial information representation to establish spatial information within diverse modal data. In particular, many DL-based fusion frameworks, such as detail-preserving conditional invertible network (DCINN) [15] and 3D-CNN and transformer prior network (3DT-Net) [24], focus on employing novel technologies to optimize modules designed in the extraction and fusion phases, thereby improving the preservation and integration capabilities of advantageous details. However, perfection remains elusive. Reconstructing HR ORS images in different ORS-MSF tasks remains a challenging endeavor, owing to two factors.

1) In terms of theoretical models, most DL methods [24], [25], [26] are based on the idea of degradation models, which iteratively optimize spatial and spectral degradation models by two parallel branches to seek exact solutions. However, ORS images of diverse modalities exhibit complementarity in both spatial and spectral dimensions. This implies that solving a degradation model constructed from a single known ORS image is not capable of retrieving the precise spatial or spectral characteristics of the desired image. Furthermore, existing theoretical models ignore the fact that spatial details and spectral distribution can be mutually reinforced during the fusion process. These render task-specific methods are struggling to be applied to MSIP, HSIP, and MHIF tasks simultaneously.
2) In terms of implementing the network architecture, it can be divided into three phases: image preprocessing, feature extraction, and feature fusion. During the image preprocessing phase, these methods typically use a single convolutional or linear layer to learn spatial and spectral degradation models. Due to the high complexity of real-world degradation, such simple simulations are inappropriate. In addition, these methods employ continuous pixelwise operations, such as difference and concatenation, to establish spatial information representations among diverse modal images. The pixelwise difference operation solely considers the unique information in one of the images, disregarding the complementary spatial information in the other image. The pixelwise concatenation operation directly concatenates diverse modal images in the channel dimension, which takes into account the complementarity between the diverse modal images but leads to a large amount of information redundancy. As a result, the fusion outcomes are highly dependent on the modules developed during the subsequent feature extraction phase, rendering it challenging to estimate the fine-grained detail representation. During the feature fusion stage, existing methods align spatial–spectral features of diverse modalities at various scales using explicit predefined functions

that represent pixels as discrete points. However, these discrete sampling techniques ignore the smooth characteristics of spectral degradation, leading to spatial or spectral distortion.

To attain high spatial–spectral fidelity fusion while being capable of adapting to diverse fusion tasks, it is imperative to contemplate how to exploit the complementary properties of different modal features in both spatial and spectral dimensions. Fine details can lead to a reasonable spectral distribution of the fused image, thereby enabling the acquisition of more appropriate spatial details. To this end, we systematically decompose the ORS-MSF task into two phases: detail reconstruction and feature alignment. Detail reconstruction aims to concentrate on estimating the joint spatial degradation process of diverse modal data in the spatial dimension, which in turn reconstructs complementary detail representations. In the meantime, it is necessary to synthesize unique and complementary detail information in the reconstruction process while avoiding a large amount of redundant information. Feature alignment, on the other hand, emphasizes the estimation of the joint spectral degradation process of diverse modal data, exploiting the complementarity in the spectral dimension, with the goal of maintaining smoothness in the spectral dimension. By enabling alignment and fusion of reconstructed details and the corresponding spectral information, high spatial–spectral fidelity fusion is achieved.

Inspired by the above theoretical analysis, we propose a novel generic framework, commonly referred to as RAMSF. Consequently, RAMSF consists of two main phases: low-frequency-driven high-frequency salient detail reconstruction (LHSDR) and coordinate-modal-guided spatial–spectral feature progressive alignment (CSFPA). A robust generalization capability that can be applied to different ORS-MSF tasks is displayed by the proposed RAMSF, which achieves a balance between spatial and spectral preservation during the fusion process. The following contribution can be drawn.

1) By analyzing the theoretical fusion models and the network architectures implemented by existing methods, we decompose for the first time the ORS-MSF task into two main phases: detail reconstruction and feature alignment. A generic fusion framework based on the above concept has been developed, known as RAMSF. As illustrated in Fig. 1(c), RAMSF achieves the SOTA performance by comparing it with the most advanced methods in different fusion tasks. In addition, the parameter settings of RAMSF are identical for different tasks, demonstrating its robust generalization capability and applicability.
2) To address the issue of coarse spatial information representation, we introduce LHSDR. It estimates the joint spatial degradation process in various frequency directions from diverse modal data and derives salient details in a hierarchical integration, with low frequencies driving high ones. Regardless of the ORS-MSF tasks, it is crucial to integrate detail representations from diverse modal data, and these reconstructed details can lay the foundation for the subsequent high-fidelity fusion.
3) To address the issue of imprecise alignment of spatial–spectral features, we propose CSFPA. With the establishment of coordinate-modal relationships among diverse modal features at various scales in the continuous domain, CSFPA is able to estimate the joint spectral degradation process. This rigorous alignment can maintain smoothness of the spatial–spectral features and thus is applicable to spectral images with different spectral bands at various resolutions.
4) Ten datasets captured by ten different sensors from three different ORS-MSF tasks are used to experiment, comprising eight simulated and five real test sets. Our proposed methodology surpasses that of representative advanced methods while maintaining a high level of efficiency.

II. Related Works

A. Traditional ORS-MSF Methods

The traditional ORS-MSF methods can mainly be classified into three distinct categories, including component replacement (CS), multiresolution analysis (MRA), and variational optimization (VO). In the CS-based method, the spatial component of the low-resolution spectral reference images (LRMSI/LRHSI) is substituted with the spatial reference image (PANI/MSI). Methods such as the principal component analysis (PCA) [5], [6], [7], intensity-hue-saturation (IHS) [8], [9], and adaptive Gram Schmidt (GSA) [10], [11] are included in this category. The sharpened outcomes of CS can enhance spatial fidelity; however, they are susceptible to spectral distortion in certain intricate scenes. The MRA-based method employs the multiresolution analysis framework to extract spatial information and inject it into spectral reference images. The MRA-based techniques possess the capability to preserve adequate spectral information; however, they are susceptible to spatial distortion owing to their inadequate extraction of spatial information. The VO-based methods primarily encompass model-based and sparse-based methods, among which the most representative methods are coupled nonnegative matrix factorization (CNMF) [27], Gaussian prior [28], and Bayesian naive Gaussian prior [29]. These methods possess limited representation capabilities and heavily rely on a priori assumptions.

B. DL-Based ORS-MSF Methods

Recently, DLs have garnered significant attention due to their potent nonlinear fitting capabilities. Inspired by the notion of the traditional detail injection pipelines, Deng et al. [13] proposed a fusion network (FusionNet) that directly executes a pixelwise difference operation on PANI and MSI to represent spatial detail information. Subsequently, they learn the corresponding detail representations by using a specially designed deep neural networks (DNN) and finally inject these detail representations into the original MSI. Prior to the pixelwise differencing, Chen et al. [14] utilized guided filtering to enhance the diverse modal images in order to obtain a finer detail representation. The fusion performance of these two

difference-based methods depends on the convolutional block or transformer block designed at the feature extraction stage, which can be prone to result in the loss of critical information. For this reason, Wang et al. [15] proposed DCINN, which utilizes a conditional invertible neural network to preserve the detail representation after differencing. DCINN can be applied to both MSIP and MHIF tasks. Diverse modal ORS images are imaged by different sensors and there are radiometric differences between them. This implies that an ORS image of any one modality contains information that is not present in the ORS images of the other modalities. As a result, the pixelwise difference operation solely considers the unique information in PANI or MSI, disregarding the complementary spatial information in LRMSI or LRHSI. To this end, many methods employ pixelwise concatenation operation to combine the two input images in the image dimension to represent the complementary spatial information that needs to be learned. Based on this idea, Wang et al. [17] extracted multiscale features of diverse modalities using cascadic multireceptive learning (CML) blocks after concatenating PANI and LRMSI in the channel dimension. Deng et al. [19] utilized transformer blocks to mine the global dependencies of the concatenated images. However, the significant disparities between diverse modal ORS images render image dimension-based concatenation methods challenging to mine fine spatial–spectral features. To this end, some methods use parallel branches to extract the diverse modal features, followed by the concatenated processing in the channel dimension, such as pansharpening generative adversarial network (PSGAN) [20], multistage dual-attention network (MDANet) [21], and 3DT-Net [24]. Using this strategy, Chen et al. [30] proposed a spectral-spatial transformer, which can achieve satisfactory sharpening performance. These methods are prone to producing a substantial quantity of redundant information, particularly for hyperspectral data comprising hundreds of bands. In addition, these methods only perform well on single datasets. It is imperative to enhance the generalization capabilities, as the lack of accuracy may have far-reaching effects on downstream tasks. Chen et al. [31], [32] proposed recurrent pansharpening network with arbitrary numbers of bands (ArbRPN), which is capable of reconstructing HRMSI for an arbitrary number of bands, thereby greatly improving the generalization capability of the model with respect to the number of bands. In addition, some methods focus specifically on the extraction of high-frequency details from PANI, such as high- and low-frequency fusion networks (HLFNet) [33] and Hyper-deep-shallow fusion network (DSNet) [34]. Seo et al. [35] proposed an unsupervised fusion framework that integrates registration and fusion during training and retains more detailed information through two designed loss functions.

C. Implicit Neural Representation

ORS images, such as MSI and HSI, are obtained through continuous imaging across a certain spectral range, whose spectral information is continuous. However, existing methods align spatial–spectral features of diverse modalities at various scales using explicit predefined functions that represent pixels as discrete points. These discrete sampling techniques ignore the smooth characteristics of spectral degradation and also disrupt the continuity of the spectral bands, resulting in spatial or spectral distortion. Therefore, implicit neural representation (INR) emerged, which is capable of parameterizing the signal into a continuous function representation through neural networks. With this property, INR has applications in several low-level vision tasks, such as hyperspectral reconstruction [36], [37] and 3-D reconstruction [38], [39], [40], [41]. Chen et al. [42] proposed spectral-wise attention based on INR, which is also capable of realizing accurate hyperspectral reconstruction. For the MHIF task, Zhu et al. [43] proposed generative adversarial network with quadtree implicit sampling (QIS-GAN), which utilizes the concept of quadtree to reformulate the INR. QIS-GAN achieves satisfactory fusion quality by concatenating the MSI and the LRHSI and then mining the potential spatial–spectral features using the INR. However, INR necessitates the computation of pixel values for all dimensions of diverse modal features, thereby undoubtedly elevating the computational complexity of the model.

III. Methodology

A. Problem Formulation and Overview

For ease of understanding, we use italicized Latin letters for scalars, e.g., H or b, handwritten letters for tensors, e.g., $\mathcal{Y}$ or $\mathcal{y}$, and bold letters for matrices, e.g., $\mathbf{Y}$. In particular, we denote the input spatial reference image as $\mathcal{Y} \in \mathbb{R}^{H\times W\times b}$, e.g., the PANI in two pan-sharpening tasks or MSI in the MHIF task; the input spectral reference image as $\mathcal{Z} \in \mathbb{R}^{h\times w\times B}$, e.g., the LRMSI or LRHSI; the upsampled spectral reference image as $\hat{\mathcal{Z}} \in \mathbb{R}^{H\times W\times B}$; and the output HRMSI or HRHSI as $\mathcal{X} \in \mathbb{R}^{H\times W\times B}$. $\mathcal{Y}$ possesses a high spatial resolution with height and width H and W, respectively, but has a low number of spectral bands b. $\mathcal{Z}$ possesses a high spectral resolution B, but its spatial resolution is low with height and width h and w, respectively. In ORS-MSF tasks, the ratio between the spatial resolution of $\mathcal{Y}$ and $\mathcal{Z}$ is constant r, i.e., $H = rh$ and $W = rw$. The goal of the ORS-MSF is to obtain the $\mathcal{X}$ with sufficient spatial information and corresponding spectral information by combining the advantageous information from $\mathcal{Y}$ and $\mathcal{Z}$.

It is evident that the information contained in the input image pairs ($\mathcal{Y}$ and $\mathcal{Z}$) is significantly smaller than the desired image ($\mathcal{X}$). For instance, if $\mathcal{Y}$ is a PANI with a spatial scale of 200 × 200 and $\mathcal{Z}$ is an eight-band LRMSI with eight bands with a size of 50 × 50 × 8, the total size of the input image pairs is 60 000 pixels, whereas the desired image (200 × 200 × 8) size is 320 000 pixels. Therefore, ORS-MSF can be considered as an ill-posed problem with an infinite number of solutions. For an accurate solution space, the spatial degradation process or spectral degradation process is estimated using a single input image and then solved for the inverse process to reconstruct the fusion result [16], [24], [25], [26]. Specifically, they construct a spectral degradation model between $\mathcal{Y}$ and $\mathcal{X}$ and spatial degradation model between $\mathcal{Z}$ and $\mathcal{X}$ based on the following physical observation models:

$$\mathbf{Y} = \mathbf{XR}, \quad \mathbf{Z} = \mathbf{PX} \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^{HW\times b}$, $\mathbf{Z} \in \mathbb{R}^{hw\times B}$, and $\mathbf{X} \in \mathbb{R}^{HW\times B}$ are the matrix forms after $\mathcal{Y}$, $\mathcal{Z}$, and $\mathcal{X}$ reshaping, respectively.

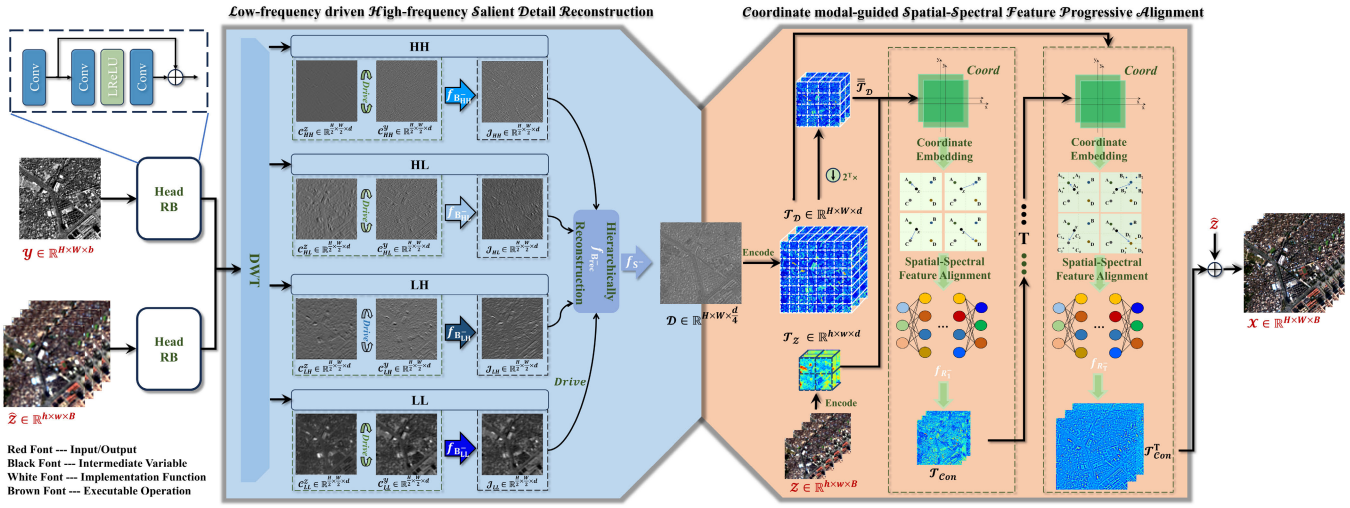

Fig. 2. Overall framework of the proposed RAMSF, which comprises two fundamental components: 1) LHSDR and 2) CSFPA.

Besides, $\mathbf{R} \in \mathbb{R}^{B\times b}$ and $\mathbf{P} \in \mathbb{R}^{hw\times HW}$ are the spectral response matrix modeling the spectral response function (SRF) and the spatial downsampling matrix modeling the point spread function (PSF), respectively. The above observation model suggests that $\mathbf{X}$ is typically obtained by resolving the following energy function:

$$\mathbf{X} = \arg\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{XR}\| + \|\mathbf{Z} - \mathbf{PX}\| + \Gamma(\mathbf{X}). \quad (2)$$

The initial two terms refer to the spectral and spatial fidelity terms, while the third term refers to the regularization term.

These studies seek an exact solution by iteratively optimizing through two parallel branches. The spatial reference image $\mathbf{Y}$ and the spectral reference image $\mathbf{Z}$ possess complementary information both spatially and spectrally. This means that solving a degradation model constructed from a single known reference image is not capable of retrieving the precise spatial or spectral characteristics of the desired image $\mathbf{X}$. Furthermore, spatial details and spectral distribution are mutually reinforcing in the fusion process. Fine details can lead to a reasonable spectral distribution of the fused image, thereby enabling the acquisition of more appropriate spatial details. In view of the above factors, it is imperative that the advantageous information of $\mathbf{Y}$ and $\mathbf{Z}$ should be fully considered when optimizing either spatial or spectral term. As a result, a new model is quantified as follows:

$$\begin{cases} \mathbf{Z} = \mathbf{P_1X}, & \mathbf{Y} = \mathbf{P_2X} \\ \mathbf{Y} = \mathbf{XR_1}, & \mathbf{Z} = \mathbf{XR_2}. \end{cases} \quad (3)$$

The spatial downsampling matrices $\mathbf{P_1} \in \mathbb{R}^{hw\times HW}$ and $\mathbf{P_2} \in \mathbb{R}^{HW\times HW}$ both consist of a spatial blurring matrix associated with the PSF and a downsampling matrix, wherein $\mathbf{P_1}$ exhibits the identical shape as $\mathbf{P}$ and $\mathbf{P_2}$'s downsampling matrix contains all elements of one. $\mathbf{R_1}$ and $\mathbf{R_2}$ are spectral response matrices associated with the SRF. Correspondingly, the energy function in (2) can be remodeled as follows:

$$\mathbf{X} = \arg\min_{\mathbf{X}} \frac{1}{2}(\|\mathbf{Z} - \mathbf{P_1X}\| + \|\mathbf{Y} - \mathbf{P_2X}\|) + \frac{1}{2}(\|\mathbf{Y} - \mathbf{XR_1}\| + \|\mathbf{Z} - \mathbf{XR_2}\|) + \Gamma(\mathbf{X}). \quad (4)$$

The first two items aim at obtaining complementary spatial details in diverse modal images, while the third and fourth items aim at obtaining spectral information corresponding to the spatial details. To this end, the above energy function can be further decomposed into two subproblems

$$\mathbf{D} = \arg\min_{\mathbf{D}} \frac{1}{2}(\|\mathbf{Z} - \mathbf{P_1D}\| + \|\mathbf{Y} - \mathbf{P_2D}\|) + \Gamma_1(\mathbf{D}) \quad (5)$$

$$\mathbf{X} = \arg\min_{\mathbf{X}} \frac{1}{2}(\|\mathbf{D} - \mathbf{XR_1}\| + \|\mathbf{Z} - \mathbf{XR_2}\|) + \Gamma_2(\mathbf{X}) \quad (6)$$

where $\mathbf{D} = \nabla\mathbf{X}$ represents the spatial details in $\mathbf{X}$. Based on the above theoretical analysis, we decompose ORS-MSF into two distinct phases, namely, detail reconstruction and feature alignment. This decomposition aims to solve (5) and (6).

Based on the above theoretical model, we propose a generic fusion framework for different ORS-MSF tasks, known as RAMSF. The proposed RAMSF consists of two fundamental components, namely, LHSDR and CSFPA, which correspond to specific implementations of detail reconstruction and feature alignment within the theoretical model, respectively. Initially, $\mathcal{Y}$ and $\hat{\mathcal{Z}}$ are fed into the head residual block (HRB). The structure of HRB is shown in the top-left corner of Fig. 2. It adopts a residual structure, wherein the initial convolution block is used to increase the dimensionality of the input image, whereas the subsequent residual blocks [44] are used to extract the shallow coded features. Subsequently, the coded features are fed into the LHSDR. The proposed LHSDR aims to model (5) to explore the spatial complementarity of diverse modal images. It estimates the joint spatial degradation process in various frequency directions from diverse modal data and derives salient details in a hierarchical integration, with LF driving HF, to obtain coupled high-frequency details. The coupled high-frequency details and encoded spectral information are then fed to the CSFPA to progressively optimize (6), which in turn effectively explores the spectral complementarity of the diverse modal data. The CSFPA estimates the joint spectral degradation process by establishing coordinate-mode relations between coupled high-frequency details and corresponding spectral information in the continuous domain. It possesses the capability to smoothly align the features at various scales

within the continuous domain. Finally, the high-fidelity fused image is obtained through fine detail reconstruction and accurate feature alignment.

B. Low-Frequency-Driven High-Frequency Salient Detail Reconstruction

The objective of (5) is to reconstruct complementary detail representations in diverse modal data. According to Moore's pseudoinverse in matrix theory, it is easy to check that $\mathbf{P_1^-Z}$ and $\mathbf{P_2^-Y}$ are invertible transformations of $\mathbf{Z}$ and $\mathbf{Y}$, respectively, and thus, $\mathbf{P_1^-Z}$ and $\mathbf{P_2^-Y}$ can theoretically retain all the information of $\mathbf{Z}$ and $\mathbf{Y}$, respectively. Since $\mathbf{D}$ and $\mathbf{X}$ are not available, the invertible $\mathbf{P_1^-Z}$ and $\mathbf{P_2^-Y}$ can be used to approximate the detail information $\mathbf{D}$ in $\mathbf{X}$. Specifically, we employ the 2-D Haar wavelet transform to obtain the frequency components from various directions in diverse modal images. The filters for various directions are quantified as follows:

$$f_{\mathrm{HH}} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad f_{\mathrm{HL}} = \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}$$
$$f_{\mathrm{LH}} = \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}, \quad f_{\mathrm{LL}} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \tag{7}$$

where HH, HL, LH, and LL represent the HF component, the HF component in the vertical direction, the HF component in the horizontal direction, and the LF component, respectively. The frequency information in various directions can be quantified as follows:

$$\begin{cases} \mathcal{C}^x_{\mathrm{LL}}(i,j) = \mathcal{F}_x(2i-1,2j-1) + \mathcal{F}_x(2i-1,2j) \\ \qquad + \mathcal{F}_x(2i,2j-1) + \mathcal{F}_x(2i,2j) \\ \mathcal{C}^x_{\mathrm{LH}}(i,j) = -\mathcal{F}_x(2i-1,2j-1) - \mathcal{F}_x(2i-1,2j) \\ \qquad + \mathcal{F}_x(2i,2j-1) + \mathcal{F}_x(2i,2j) \\ \mathcal{C}^x_{\mathrm{HL}}(i,j) = -\mathcal{F}_x(2i-1,2j-1) + \mathcal{F}_x(2i-1,2j) \\ \qquad - \mathcal{F}_x(2i,2j-1) + \mathcal{F}_x(2i,2j) \\ \mathcal{C}^x_{\mathrm{HH}}(i,j) = \mathcal{F}_x(2i-1,2j-1) - \mathcal{F}_x(2i-1,2j) \\ \qquad - \mathcal{F}_x(2i,2j-1) + \mathcal{F}_x(2i,2j) \end{cases} \tag{8}$$

where $x \in \{\mathcal{Y}, \mathcal{Z}\}$; $\mathcal{F}_x \in \mathbb{R}^{H\times W\times d}$ denotes the shallow features extracted by HRB; and d is the number of filters in the last convolution layer, whose value is determined to be 32 in practical implementation. Besides, (i, j) denotes the spatial location of the pixel values. The coefficient matrix $\mathbf{M}$ before and after the transformation in (8) can be quantified as

$$\mathbf{M} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 \end{bmatrix}. \tag{9}$$

It is apparent that $\mathbf{M}$ is an invertible matrix, which lays the foundation for the complete preservation of the input information for the diverse modalities.

There exists complementary spatial information in diverse modal images. In particular, $\mathcal{Y}$ contains HF textural details, whereas $\mathcal{Z}$ contains more LF global information. Therefore, complementary spatial information in diverse modal images can be obtained by driving HF texture details with LF global information. Utilizing the position information obtained from the discrete wavelet transform (DWT), the frequency components of the various directions in diverse modal images are integrated to encode a more robust feature representation by LF driving HF

$$\mathcal{D} = f_{\mathbf{P}^-_{1,2}}\left(\mathcal{C}^{\mathcal{Y}}, \mathcal{C}^{\mathcal{Z}}\right) \tag{10}$$

where $\mathcal{D} \in \mathbb{R}^{H\times W\times(d/4)}$ is the estimated detail representation. The joint implementation function corresponding to the inverse matrix of the spatial downsampling matrix $\mathbf{P_1}$ and $\mathbf{P_2}$ is denoted by $f_{\mathbf{P}^-_{1,2}}$. It comprises the implementation function of the inverse matrix of a spatial blur matrix associated with the PSF and the implementation function of the inverse matrix of a downsampling matrix (i.e., the upsampling matrix). In order to estimate the joint spatial degradation process of diverse modal data, we decompose $f_{\mathbf{P}^-_{1,2}}$ into three distinct components, namely, $f_{\mathbf{B}^-_{\boldsymbol{\gamma}}}$, $f_{\mathbf{B}^-_{\mathbf{rec}}}$, and $f_{\mathbf{S}^-}$. $f_{\mathbf{B}^-_{\boldsymbol{\gamma}}}$ and $f_{\mathbf{B}^-_{\mathbf{rec}}}$ are utilized to model the inverse matrix of spatial blur matrix hierarchically, whereas $f_{\mathbf{S}^-}$ is utilized to model the upsampling matrix. Specifically, $f_{\mathbf{B}^-_{\boldsymbol{\gamma}}}$ models the information in identical frequency directions in diverse modal images, leading to the joint representations in various frequency directions; $f_{\mathbf{B}^-_{\mathbf{rec}}}$ models the joint representations in various directions step-by-step to obtain complementary detail representations; and $f_{\mathbf{S}^-}$ models the upsampling matrix, leading to the coupled high-frequency details.

Instead of solely focusing on the spatial information in spatial reference image ($\mathcal{Y}$), we consider the various frequency components in diverse modal images. An average fusion scheme is employed to integrate the frequencies of diverse modal images in the same direction. In this manner, we jointly model $\mathbf{P_1^-Z}$ and $\mathbf{P_2^-Y}$ in terms of frequency directions by driving the HF with LF. The process can be quantified as follows:

$$\mathcal{J}_\gamma = f_{\mathbf{B}^-_{\boldsymbol{\gamma}}}\left(\mathcal{C}^{\mathcal{Y}}_\gamma \oplus \mathcal{C}^{\mathcal{Z}}_\gamma\right), \quad \text{s.t. } \gamma \in \{\mathrm{LL}, \mathrm{LH}, \mathrm{HL}, \mathrm{HH}\} \tag{11}$$

where $f_{\mathbf{B}^-_{\boldsymbol{\gamma}}}$ denotes the function corresponding to $\mathbf{B}^-_{\boldsymbol{\gamma}}$. $\mathbf{B}^-_{\boldsymbol{\gamma}}$ denotes the invertible version of the spatial blur matrix associated with the PSF, which is used to model the complementary information of frequency components in the same direction in diverse modal data. Besides, $\mathcal{C}^{\mathcal{Y}}_\gamma$ and $\mathcal{C}^{\mathcal{Z}}_\gamma$ stand for diverse modal frequency features obtained by (8). Joint representations in the same frequency direction are integrated in an LF-driven HF manner, taking advantage of the complementary strengths of the diverse modal data. Furthermore, we continue to model the joint representations in various directions hierarchically by driving HF with LF (e.g., $\mathcal{J}_{\mathrm{LL}}$ driving $\mathcal{J}_{\mathrm{LH}}$). Finally, the coupled high-frequency details $\mathcal{D}$ are obtained. The process can be quantified as follows:

$$\mathcal{D} = f_{\mathbf{S}^-}\left(f_{\mathbf{B}^-_{\mathbf{rec}}}\left(\mathcal{J}_\gamma \oplus \mathcal{J}_\beta\right)\right), \quad \text{s.t. } \beta \neq \gamma \tag{12}$$

where $f_{\mathbf{B}^-_{\mathbf{rec}}}$ bears resemblance to $f_{\mathbf{B}^-_{\boldsymbol{\gamma}}}$ and is a PSF-related operation that can be executed by HRB. The convolution kernel size, step size, and padding of convolutional layers in

HRB are 3, 1, and 1 respectively. $f_{\mathbf{S}^-}$ denotes the upsampling function, which can be executed with inverse DWT. Since the coefficient matrix $\mathbf{M}$ in (9) is invertible, the input information from diverse modalities can be retained intact during the execution of upsampling via inverse DWT. By jointly modeling $\mathbf{P}_1^- \mathbf{Z}$ and $\mathbf{P}_2^- \mathbf{Y}$ and decomposing them into three operations hierarchically, fine-grained detail representations in diverse modal images can be obtained. Fig. 2 illustrates how the frequency components of the same direction in different modal features are integrated to reveal finer edge details. By observing $\mathcal{Y}$, $\mathcal{Z}$, and $\mathcal{X}$ and the coupled high-frequency detail $\mathcal{D}$, it can be determined that the proposed LHSDR effectively exploits the detail representation in diverse modal data and proficiently models the fine-grained detail information.

It is worth noting that when modeling the frequency components in various directions, we ignore the LF components in the spectral reference features, i.e., $\mathcal{C}_{\mathrm{LL}}^{\mathcal{Z}}$. Two primary motives drive this approach. First, there exists a substantial quantity of redundant LF information in $\mathcal{C}_{\mathrm{LL}}^{\mathcal{Z}}$, which is opposed to the objective of reconstructing fine HF details in $\mathcal{X}$. Second, DNNs are more accustomed to anticipating HF information during the training process, and this redundant LF information may potentially hinder the DNNs from learning the HF details to a certain extent. On the contrary, we consider the LF components in the spatial reference features, i.e., $\mathcal{C}_{\mathrm{LL}}^{\mathcal{Y}}$. A modest amount of low frequencies drives the high ones, which is capable of achieving a more salient detail representation while avoiding a substantial amount of redundant information. We will discuss this section further in Section IV-E.

C. Coordinate-Modal-Guided Spatial–Spectral Feature Progressive Alignment

The objective of (6) is to further explore the complementarity of the reconstructed coupled detail representation ($\mathcal{D}$) and the spectral reference image ($\mathcal{Z}$) in the spectral dimension, thereby obtaining a high spatial–spectral fidelity image ($\mathcal{X}$). From the perspective of the observation model, the values on the different channels in $\mathcal{D}$ can be expressed as the accumulation of different responses in the spectral range. However, the spectral range is sampled as a wavelength interval in practical applications. They are generally quantified in the following discrete form:

$$\mathcal{D}_c(p) = \sum_{k=1}^{B} \mathcal{X}(i, j, \lambda_k) f_{\boldsymbol{R}_c}(\lambda_k) \tag{13}$$

where $c \in \{1, 2, 3, \ldots, (d/4)\}$ is the channel index of $\mathcal{D}$. The spectral range is denoted by the wavelengths $\lambda_1, \ldots, \lambda_B$. $\mathcal{X}(i, j, \lambda)$ denotes the spectral radiance value at spatial position (i, j) on λ. $f_{\mathbf{R_c}}$ denotes the SRF corresponding to index c. Continuing from Section III-B, we model the spectral complementarity of $\mathcal{D}$ and $\mathcal{Z}$ by utilizing coordinate-modal guidance. Similar to Section III-B, we model $f_{\boldsymbol{R}^-}$ by exploring the spectral complementarity of $\mathcal{D}$ and $\mathcal{Z}$. The modeling procedure consists of three fundamental steps: 1) establishing the coordinate encoding mode between pixels; 2) establishing the coordinate-modal relationship between $\mathcal{D}$ and $\mathcal{Z}$ through continuous coordinate mapping; and 3) approximating $f_{\boldsymbol{R}^-}$ by utilizing the implicit encoding function and aligning diverse modal features at various scales within the continuous domain.

To ensure continuity across input pixels, the coordinate map is scaled to a square grid of $[-1, 1] \times [-1, 1]$, which enables features of different scales to share a 2-D coordinate for subsequent alignment and fusion. For convenience, we represent pixels with the center position of the pixel. The normalized 2-D coordinate $\mathcal{C}$ can be quantified as

$$\mathcal{C}(i, j) = \left[-1 + \frac{2i+1}{H}, -1 + \frac{2j+1}{W}\right] \tag{14}$$

where $i \in [0, H-1]$ and $j \in [0, W-1]$. Prior to performing position coding, we employ the HRB to extract the shallow features of $\mathcal{D}$ and $\mathcal{Z}$ and map them to the same channel dimension. The shallow features of $\mathcal{D}$ and $\mathcal{Z}$ are denoted by $\mathcal{T}_{\mathcal{D}}$ and $\mathcal{T}_{\mathcal{Z}}$, respectively. Since $\mathcal{D}$ takes full advantage of the position information of DWT during the detail reconstruction phase, it is used as a benchmark to establish the coordinate-modal relationship between $\mathcal{T}_{\mathcal{D}}$ and $\mathcal{T}_{\mathcal{Z}}$. This alleviates the computational burden caused by calculating the coordinates of different modal data at the same time. The abovementioned coordinates allow us to identify the position of $\mathcal{T}_{\mathcal{D}}$ and $\mathcal{T}_{\mathcal{Z}}$ and establish the coordinate-modal relationship between $\mathcal{T}_{\mathcal{D}}$ and $\mathcal{T}_{\mathcal{Z}}$. In this way, each pixel feature within diverse modal features is represented in a continuous manner. Subsequently, the positional encoded $\mathcal{T}_{\mathcal{D}}$ and $\mathcal{T}_{\mathcal{Z}}$ are fed into an implicit decoding function. The implicit decoding function takes into account diverse modal features and their relative coordinate positions, with the capability of aligning features at various scales in the same coordinate dimension. As depicted in Fig. 2, the spectral feature value at each query coordinate $\mathcal{C}_q$ on the continuous feature map $\mathcal{T}_{\mathcal{C}\mathrm{on}}$ can be quantified as

$$\mathcal{T}_{\mathcal{C}\mathrm{on}}(\mathcal{C}_q) = f_{\boldsymbol{R}^-}\left(\mathcal{T}_{\mathcal{Z}}(\mathcal{C}_t), \mathcal{C}_q - \mathcal{C}_t, \bar{\bar{\mathcal{T}}}_{\mathcal{D}}(\mathcal{C}_q)\right). \tag{15}$$

We employ a multilayer perceptron as the implicit decoding function to approximate $f_{\boldsymbol{R}^-}$, which is capable of aligning features at various scales on the same coordinate scale via implicit neural functionalities. $f_{\boldsymbol{R}^-}$ consists of one input layer, three hidden layers, and one output layer, with the number of nodes in the three hidden layers being 128, 256, and 128, respectively. Besides, $\mathcal{T}_{\mathcal{Z}}(\mathcal{C}_t)$ and $\bar{\bar{\mathcal{T}}}_{\mathcal{D}}(\mathcal{C}_q)$ represent the spectral feature values at the query coordinates $\mathcal{C}_t$ and the detail feature values at the query coordinates $\mathcal{C}_q$, respectively. $\bar{\bar{\mathcal{T}}}_{\mathcal{D}}$ is the result of downsampling of $\mathcal{T}_{\mathcal{D}}$, which is performed by an adaptive maximum pooling layer. It is worth mentioning that the number of multi-layer perceptrons (MLPs) and the downsampling times are determined by the number of times of progressive alignment.

In order to guarantee the continuity of the spectral features, we utilize four spectral vectors closest to the current position to generate an estimation of the spectral feature values. Since the spectral vectors in various directions have different distances from the estimated values, they exhibit dissimilar importance to the estimated values. We employ a normalized confidence voting mechanism, which calculates the weights of the individual pixel values based on the proportion of the subspace

corresponding to the four spectral vectors in the vicinity of the overall space. Therefore, the weights corresponding to the four spectral feature values in the vicinity of $\mathcal{T}_{\mathcal{Z}}$ can be quantified as

$$w_{q,t} = \frac{\mathcal{S}_{q,t}}{\sum_{t \in N_q} \mathcal{S}_{q,t}} \tag{16}$$

where N_q represents the set of neighborhood pixels surrounding the spectral feature value $\mathcal{T}_{\mathcal{Z}}$ and $\mathcal{S}$ represents the subspace corresponding to the spectral feature values in a certain neighborhood. Furthermore, the spectral feature value is weighted according to four spectral feature vectors in the neighborhood. The process in (15) is remodeled as follows:

$$\mathcal{T}_{\mathcal{C}\text{on}}\left(\mathcal{C}_q\right) = \sum_{i \in N_q} w_{q,t} f_{\boldsymbol{R}^-}\left(\mathcal{T}_{\mathcal{Z}}\left(\mathcal{C}_t\right), \mathcal{C}_q - \mathcal{C}_t, \bar{\bar{T}}_{\mathcal{D}}\left(\mathcal{C}_q\right)\right). \tag{17}$$

Moreover, in order to guarantee continuity among diverse modal features, we employ a progressive alignment to sample diverse modal features to the same scale

$$\mathcal{T}^{T}_{\mathcal{C}\text{on}}\left(\mathcal{C}_k\right) = \sum_{k \in N_k} w_{k,q} f_{\mathbf{R}_T^-}\left(\mathcal{T}^{T-1}_{\mathcal{C}\text{on}}\left(\mathcal{C}_q\right), \mathcal{C}_k - \mathcal{C}_q, \mathcal{T}_{\mathcal{D}}\left(\mathcal{C}_k\right)\right) \tag{18}$$

where $\mathcal{T}^{T}_{\mathcal{C}\text{on}}$ denotes the coupled high-frequency details and spectral information following the progressive alignment and $T \leq \lfloor \sqrt{r} \rfloor$ denotes the number of times of progressive alignment. In general, the higher the times of alignment, the finer the spatial–spectral features obtained. Nonetheless, taking into consideration the efficiency of fusion, the value of T will be determined experimentally, which will be further discussed in Section IV-E. Fig. 2 presents the results of visualizing the spatial–spectral features before and after the alignment of CSFPA. It is obvious that after the progressive alignment of spatial–spectral features at different scales, targets at different scales are accompanied by specific spectral information. These reasonable spectral distributions can promote the formation of finer spatial–spectral details. Evidently, CSFPA is capable of accurately aligning diverse modal data while ensuring the continuity of spectral features.

D. Training and Implementation Details

With LHSDR and CSFPA, coupled high-frequency details and corresponding spectral information are accurately reconstructed and aligned. Additionally, a high-fidelity fused image is obtained as follows:

$$\mathcal{X} = \mathcal{T}^{T}_{\mathcal{C}\text{on}} \oplus \hat{\mathcal{Z}}. \tag{19}$$

The residual learning enables RAMSF to be more inclined to learn the HF details and corresponding spectral information required in LR MSI/HSI to HR MSI/HSI. This further aligns with the learning requirements of both detail reconstruction and feature alignment. Since the main innovation of this work lies in the design principles of RAMSF, including both LHSDR and CSFPA, we directly adopt the commonly used mean absolute error (MAE) as the loss function for training. MAE is able to focus more on learning edge detail information, which further aligns with our objective

$$\text{Loss} = \frac{\sum_{n=1}^{H}\sum_{m=1}^{W}\sum_{l=1}^{B} \|\mathcal{X}(n,m,l) - G(n,m,l)\|_1}{\text{HWB}} \tag{20}$$

where G denotes the ground truth (GT).

We validate the proposed RAMSF on three different ORS-MSF tasks, including MSIP, HSIP, and MHIF. It is worth noteworthy to mention that we utilize identical hyperparameter settings for all three tasks, thereby obviating the need to adjust the hyperparameter setting according to the different task. This indicates its excellent applicability and generalization capability. In particular, the initial learning rate, epochs, and batch size are set to $1e^{-4}$, 400, and 16, respectively. Besides, we choose Adam as the optimizer and decay the learning rate by half every 100 epochs.

IV. Experiments

A. Datasets and Settings

1) Dataset Details: We conduct extensive experiments on three different ORS-MSF tasks, namely, MSIP, HSIP, and MHIF. Specifically, we conduct experiments on ten datasets, encompassing 13 test sets containing data primarily from GaoFen-2 (GF2), QuickBird (QB), WorldView-3 (WV3), WorldView-2 (WV2), Pavia Center (PC), Botswana, FR1, Pavia University (PU), Chikusei, and ZiYuan-1E (ZY1E) sensors. The GF2, QB, WV3, and WV2 datasets[1] [45] are used to validate model performance on the MSIP task; the PC, Botswana, and FR1 datasets[2] [34] are used to validate model performance on the HSIP task; and the PU[3] [46], Chikusei[4] [47], and ZY1E[5] [48] datasets are used to validate the model performance on the MHIF task. Table I contains the basic information about ten datasets. It is noteworthy to mention that the partitioning of the training, validation, and test sets within each dataset adheres to the original partitioning or the partitioning described in the published papers, wherein the ratio between the number of samples in the training set and the number of samples in the validation set is 9:1. Each of the four MSIP datasets contains one simulation test set and one real test set, each of which contains 20 test samples. The spatial scale of the PANIs in the simulation test set is 256 × 256, while the spatial scale of the PANIs in the real test set is 512 × 512. For the HSIP dataset, the test sets corresponding to PC, Botswana, and FR1 contain two PANIs with spatial scales of 400 × 400, four PANIs with spatial scales of 128 × 128, and one PANI with a spatial scale of 456 × 76, respectively. For the MHIF dataset, the corresponding test sets of Chikusei, PU, and ZY1E contain 11 MSIs with a spatial scale of 512 × 512, four MSIs with a spatial scale of 256 × 256, and 90 MSIs with a spatial scale of 270 × 270, respectively. All datasets are

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

[2] https://github.com/liangjiandeng/HyperPanCollection

[3] https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes

[4] https://www.sal.t.u-tokyo.ac.jp/hyperdata

[5] https://github.com/meiruni/MIMFormer

TABLE I
BASIC INFORMATION OF TEN DATASETS

Tasks	MSIP				HSIP			MHIF		
Sources	GF2	QB	WV3	WV2	PC	Botswana	FR1	Chikusei	PU	ZY1E
Bands	4	4	8	8	102	145	69	128	93	76
Range (μm)	0.45-0.9	0.45-0.9	0.45-0.9	0.45-0.9	0.4-0.9	0.4-2.5	0.4-2.5	0.36-1.02	0.43-0.86	0.4-2.5
Resolution (m)	0.8	0.61	0.31	0.5	1.3	30	30	2.5	1.3	30
Ratio	4	4	4	4	4	4	6	4	4	3
Size	6907×7300	4096×4096	2811×3408	485×2020	1096×715	1496×256	2400×2400	2517×2335	1096×1096	4986×4581
Scenes	Buildings, Coasts, Cars, Water, Trees, Seasonal Swamps, Roofs, Streets, Urban									

TABLE II
QUANTITATIVE EVALUATION RESULTS ON MSIP SIMULATED TEST SETS (BOLD: BEST AND UNDERLINE: SECOND BEST)

Method	GF2			QB			WV3		
	SAM (↓)	ERGAS (↓)	$Q4$ (↑)	SAM (↓)	ERGAS (↓)	$Q4$ (↑)	SAM (↓)	ERGAS (↓)	$Q8$ (↑)
EXP (*TGRS*, 2002)	1.8531	2.4094	0.7971	8.5575	12.0189	0.5782	5.8351	7.1354	0.6027
PRACS (*TGRS*, 2011) [CS]	1.7135	1.6482	0.8975	8.2863	8.4567	0.7861	5.6081	5.2059	0.7721
TV (*GRSL*, 2014) [VO]	1.9190	1.7767	0.9051	7.5653	7.7524	0.8204	6.3392	4.9836	0.7855
MF (*TIP*, 2016) [MRA]	1.6842	1.7763	0.8784	8.0350	8.9013	0.8120	5.3162	4.9191	0.7957
FusionNet (*TGRS*, 2021) [I]	1.0135	1.0438	0.9603	4.9853	4.2581	0.9224	3.3821	2.4778	0.8939
PSGAN (*TGRS*, 2021) [III]	0.8800	0.7887	0.9755	4.6754	3.8613	0.9304	3.5527	2.6521	0.8963
PanFormer (*ICME*, 2022) [III]	0.8972	0.8436	0.9727	4.7662	4.1235	0.9273	3.2780	2.4681	0.9012
CMLNet (*TGRS*, 2023) [II]	0.8780	0.8134	0.9737	4.8879	4.6149	0.9176	3.0436	2.2344	0.9068
QIS-GAN (*TGRS*, 2023) [II]	0.8246	0.7679	0.9759	4.6785	3.9063	0.9241	3.1704	2.3472	0.9027
DCINN (*IJCV*, 2024) [I]	0.8259	0.7530	**0.9768**	4.6210	3.9333	0.9304	2.9979	2.2383	0.9113
Ours	**0.8082**	**0.7441**	**0.9768**	**4.6070**	**3.8127**	**0.9317**	**2.9863**	**2.2265**	**0.9162**

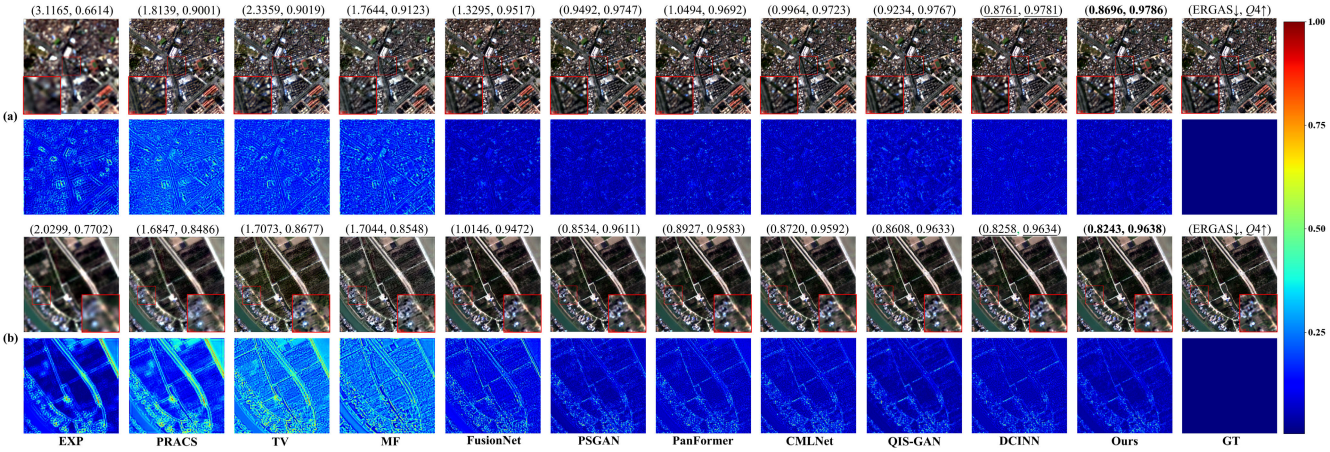

Fig. 3. Qualitative evaluation results on GF2 simulated test set. The fusion results are presented in odd rows, while the corresponding AEMs are listed in even rows. (a) First set of results on the GF2 simulated test set. (b) Second set of results on the GF2 simulated test set.

publicly available, which can be accessed in published articles and corresponding websites.

2) Evaluation Metrics: There are eight simulated test sets and five real test sets. For simulated data, we employ four reduced-scale test (RST) metrics to evaluate the fusion performance, namely, peak signal-to-noise ratio (PSNR), spectral angle mapper (SAM) [49], Erreur Relative Globale Adimensionnelle de Synthèse (ERGAS) [50], and $Q2n$ ($Q4$ for four bands and $Q8$ for eight bands) [51]. For real data, we employ four full-scale test (FST) metrics to evaluate the fusion performance, namely, D_λ, D_s, quality with no reference (QNR) [52], and hybrid QNR (HQNR) [53].

3) Comparative Methods: We selected different types of methods previously categorized, which have been validated in each fusion task. For the MSIP task, four traditional methods and six DL methods are selected. The traditional methods include 23-tap polynomial interpolation (EXP) [54], adaptive component-substitution by using partial replacement (PRACS) [55], total variation (TV) [56], and morphological filters (MF) [57]; and the DL methods include FusionNet [13], PSGAN, PanFormer, CML network (CMLNet) [17], QIS-GAN [43], and DCINN [15]. For the HSIP task, we also select four traditional and six DL methods. The traditional methods include EXP, GSA [11], CNMF [27], and modulation transfer function-generalized Laplacian pyramid (MTF-GLP) [50], while the DL methods include deep prior and dual attention residual network (DHP-DARN) [16], MDANet [21], Hyper-DSNet (DSNet) [34], pyramid shuffle-and-reshuffle transformer (PSRT) [19], QIS-GAN, and DCINN. Similarly, we select four traditional methods, including EXP, smoothing filter-based intensity modulation (SFIM) [59], CNMF [27], and guided filtering principal component analysis (GFPCA) [58], and six DL methods, including multi-domain feature learning (MDFL) [60], dual U-shape network (D-UNet) [18], deep hyperspectral image fusion network (DHIFNet) [25], 3DT-Net [24], QIS-GAN, and DCINN for the MHIF task. EXP is

TABLE III
QUANTITATIVE EVALUATION RESULTS ON MSIP REAL TEST SETS (BOLD: BEST AND UNDERLINE: SECOND BEST)

Method	GF2			QB			WV3		
	D_s (↓)	QNR (↑)	HQNR (↑)	D_s (↓)	QNR (↑)	HQNR (↑)	D_s (↓)	QNR (↑)	HQNR (↑)
EXP (*TGRS*, 2002)	0.0263	0.9737	0.9601	0.0529	0.9470	0.9142	0.0340	0.9660	0.9437
PRACS (*TGRS*, 2011) (CS)	0.0469	0.9413	0.8943	0.1399	0.8375	0.7556	0.0757	0.9120	0.8906
TV (*GRSL*, 2014) (VO)	0.0478	0.9317	0.8967	0.0939	0.8788	0.8551	0.0686	0.9155	0.9097
MF (*TIP*, 2016) (MRA)	0.0593	0.9104	0.8858	0.1105	0.8589	0.8356	0.0853	0.8909	0.8898
FusionNet (*TGRS*, 2021) (I)	0.0483	0.9369	0.9130	0.0284	0.9315	0.8998	0.0548	0.9236	0.9208
PSGAN (*TGRS*, 2021) (III)	0.0342	**0.9593**	0.9301	0.0306	0.9412	0.9324	0.0752	0.9055	0.8941
PanFormer (*ICME*, 2022) (III)	0.0419	0.9369	0.9192	0.0473	0.9236	0.8945	0.0699	0.9107	0.9063
CMLNet (*TGRS*, 2023) (II)	0.0384	0.9524	0.9362	0.0634	0.9009	0.8767	0.0436	0.9226	0.9363
QIS-GAN (*TGRS*, 2023) (II)	0.0344	0.9501	0.9424	0.0193	0.9407	0.9007	0.0625	0.9195	0.8988
DCINN (*IJCV*, 2024) (I)	0.0357	0.9536	0.9383	0.0267	0.9414	0.9355	0.0565	**0.9313**	0.9285
Ours	**0.0333**	0.9585	**0.9471**	**0.0154**	**0.9533**	**0.9506**	**0.0433**	0.9219	**0.9386**

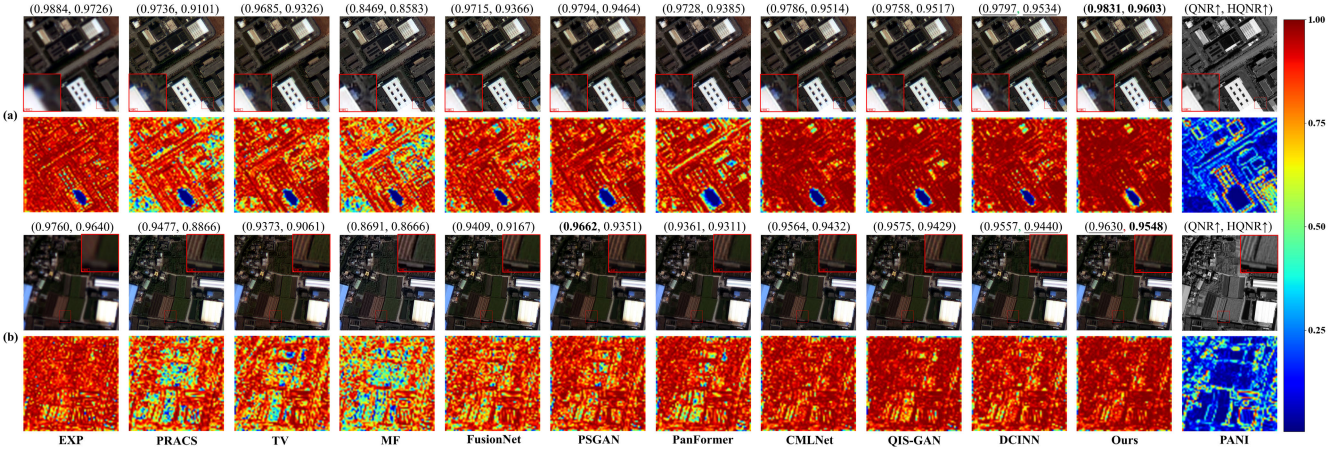

Fig. 4. Qualitative evaluation results on GF2 real test set. The fusion results are presented in odd rows, while the corresponding HMs are listed in even rows. (a) First set of results on the GF2 real test set. (b) Second set of results on the GF2 real test set.

an upsampling technique and is not counted in the quantitative evaluation ranking. In particular, the two-branch INR-related method QIS-GAN and the generalized multimodal fusion method DCINN based on detail injection are used as baseline methods for comparison experiments in all three fusion tasks.

In the comparison experiments, we strictly adhere to the guidelines provided in the official code repositories and papers. All codes are executed on the same computer with an i5-11600 CPU and two GTX-3060 GPUs. In addition, all implementations of this project will be released to ensure reproducibility.

B. Experiments for MSIP

Three MSI datasets are employed to validate the performance of each method in the MSIP task, including the GF2 dataset, the QB dataset, and the WV3 dataset, each of which contains a simulated test set and a real test set.

1) GF2 (4-Band): Fig. 3 depicts two sets of RST results and their corresponding absolute error maps (AEMs) on the GF2 simulated test set, encompassing typical scenarios such as buildings, cropland, and rivers. In addition, representative quantitative metric scores are presented above the subjective fusion results to provide a comprehensive assessment of the fusion performance. The AEMs corresponding to the fusion outcomes of the traditional methods exhibit more residuals and more colorful pixels, indicating severe spatial and spectral distortions. As shown in Fig. 3(a), the DL methods demonstrate superior fusion performance in the large buildings scenario, as evidenced by the fact that their corresponding AEMs have fewer residuals. In Fig. 3(b), FusionNet exhibits apparent spatial and spectral distortions when the ground features are more demanding in terms of fine granularity, and PSGAN, PanFormer, CMLNet, and QIS-GAN exhibit more residuals in the road-edge and small building regions. DCINN and the proposed RAMSF exhibit considerable performance. The proposed RAMSF exhibits a more consistent spectral distribution with GT in large-scale cropland areas, while the corresponding AEMs are darker in hue and contain fewer residuals. This implies that RAMSF possesses better spatial and spectral preservation capabilities, owing to its sufficient exploration of complementary spatial details and the accurate representation of the spatial–spectral features. Table II presents the quantitative results along with the mean. The best results are in boldface, while the second-best results are underlined. The publication abbreviation, year of publication, and corresponding category of each method are labeled after method names in the first column. It can be observed that DL-based methods exhibit superior performance compared to traditional methods on all metrics. The proposed RAMSF achieves the best scores on all the metrics with significant advantages, which further validates the effectiveness of our methodology.

Fig. 4 depicts two sets of FST results and their corresponding HQNR maps (HMs) on the GF2 real test set, encompassing typical scenarios such as large buildings, small buildings, cropland, and roads. The fusion results of EXP can be used as a spectral reference despite its poor results in terms of spatial quality. Traditional methods still demonstrate signifi-

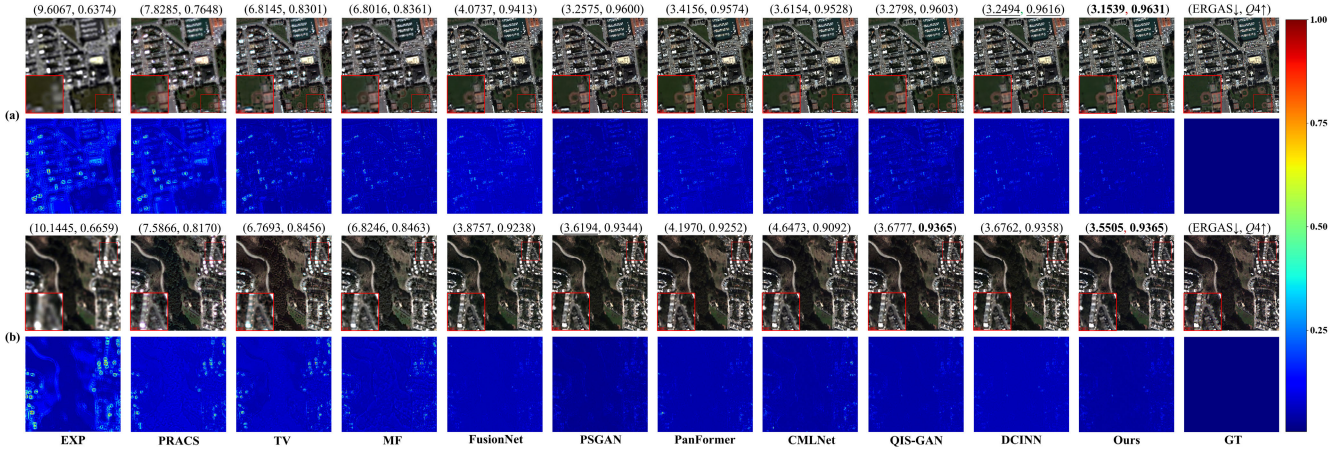

Fig. 5. Qualitative evaluation results on QB simulated test set. The fusion results are presented in odd rows, while the corresponding AEMs are listed in even rows. (a) First set of results on the QB simulated test set. (b) Second set of results on the QB simulated test set.

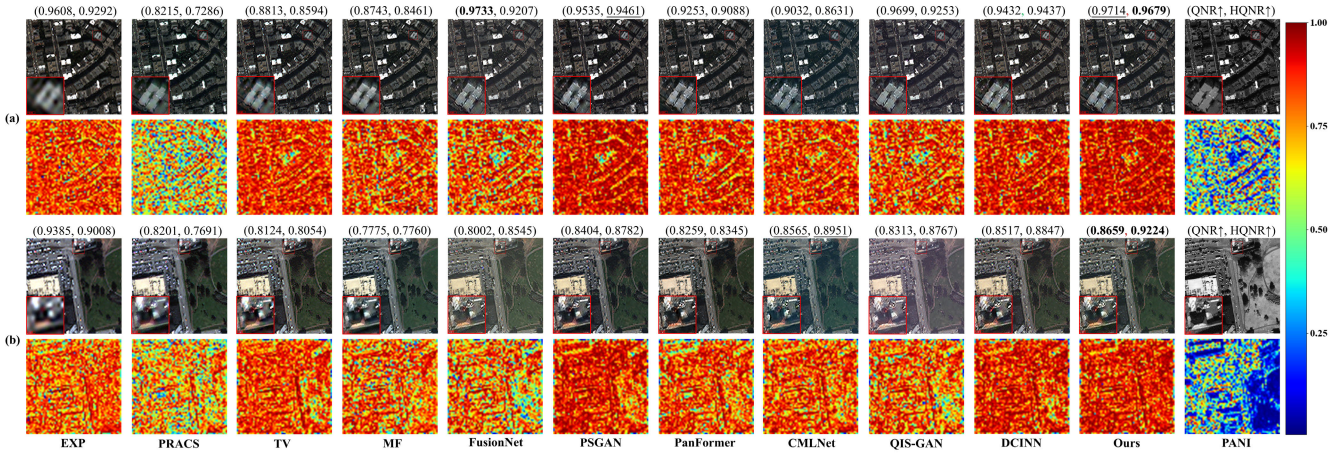

Fig. 6. Qualitative evaluation results on QB real test set. The fusion results are presented in odd rows, while the corresponding HMs are listed in even rows. (a) First set of results on the QB real test set. (b) Second set of results on the QB real test set.

cant spatial and spectral distortions. As shown in Fig. 4(a), most DL methods exhibit satisfactory fusion performance when the spatial resolution is 0.8 m and the fusion scene is relatively monolithic. However, as shown in Fig. 4(b), these DL methods exhibit excessive smoothing in the cropland areas where fine grainedness is in demand. In contrast, the proposed RAMSF exhibits a finer texture. Furthermore, the corresponding HMs further demonstrate that the RAMSF achieves more fine-grained spatial details and more reasonable spectral distributions. The quantitative evaluation results in Table III further validate the effectiveness of the proposed LHSDR and CSFPA.

2) QB (4-Band): Fig. 5 depicts two sets of RST results obtained from QB simulated test set, including typical scenarios such as cities and forests. The fusion results of MF show significant spatial distortion, while PRACS demonstrates evident spectral distortion. These results demonstrate the inherent flaw of both MRA-based and CS-based methods. Similarly, the VO-based method TV exhibits apparent spatial and spectral distortions that could be attributed to irrational parameter settings. At reduced scales, the DL methods demonstrate strong fitting capabilities. In contrast, PSGAN and the proposed RAMSF show comparable fusion performance, as evidenced by the overall darker hue of PSGAN's AEMs and fewer residuals in RAMSF's AEMs. The corresponding quantitative evaluation outcomes are shown in Table II, which shows that the proposed RAMSF achieves the best scores on all metrics. This further validates that the RAMSF is capable of reconstructing finer details and more reasonable spectral distributions, thanks to fine detail reconstruction and accurate feature alignment.

Fig. 6 depicts two sets of FST results on QB real test set. Obviously, each method shows varying degrees of performance degradation upon increasing the resolution to 0.6 m. FusionNet, CMLNet, and QIS-GAN exhibit obvious spatial and spectral distortions. This is attributed to the fact that the theoretical models or network structures adopted by these methods ignore the complementary properties of diverse modal data and are unable to reconstruct fine-grained detail representations, resulting in the loss of spatial details, which in turn causes an irrational overall spectral distribution. PSGAN and DCINN exhibit comparable performance benefiting from their rational exploitation of advanced generative adversarial networks and invertible neural network models. Compared with these DL methods, RAMSF's results are more closely aligned with those of EXP in terms of the overall spectral distribution. Furthermore, the proposed RAMSF exhibits much finer spatial details, as evidenced by the white vehicle in the zoomed-in area. The quantitative outcomes are presented in Table III. The proposed RAMSF shows notable superiority in

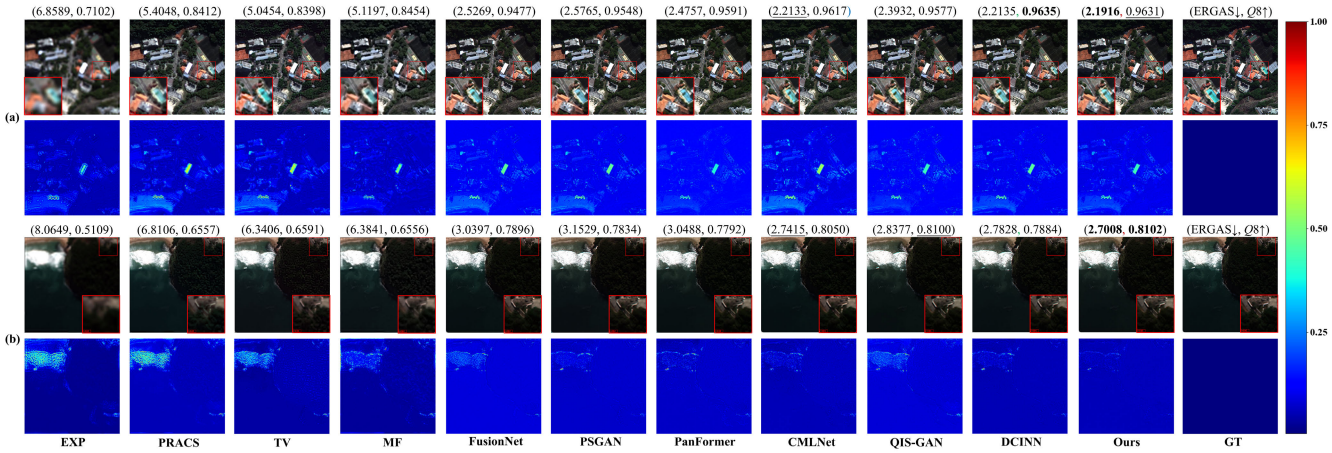

Fig. 7. Qualitative evaluation results on WV3 simulated test set. The fusion results are presented in odd rows, while the corresponding AEMs are listed in even rows. (a) First set of results on the WV3 simulated test set. (b) Second set of results on the WV3 simulated test set.

two comprehensive evaluation metrics, QNR and HQNR. This indicates that RAMSF effectively balances spatial and spectral information in diverse modal images.

3) WV3 (8-Band): Fig. 7 illustrates the RST results on the WV3 simulated test set, including typical scenarios such as street buildings, coast, and jungle. Due to the limited nonlinear fitting capability, traditional methods exhibit severe spatial blurring and spectral distortion in intricate scenes, as exemplified by blurred spatial textures and a substantial number of colored pixels in the AEMs. As shown in the edge region of the pool within the red box, FusionNet and DCINN clearly suffer from a partial loss of detail. These two methods construct representations between diverse modal data by means of differencing (detail injection), disregarding complementary details in the spectral reference image, resulting in spectral distortion. The two methods PSGAN and CMLNet, which belong to the concatenation, also show slight spectral distortions, as evidenced by the presence of a large number of green pixels on the white roofs in the AEMs. The concatenation approach builds up the detail representation between the diverse modal data rather roughly, making the fusion results highly dependent on the extraction capability of the designed modules. QIS-GAN and the proposed RAMSF demonstrate comparable performance. Compared with QIS-GAN, RAMSF preserves more details, as evidenced by the overall darker tone and fewer residuals in the corresponding AEMs. QIS-GAN employs implicit neural sampling resembling a quadtree, which is capable of better preserving the complementary information between diverse modalities. However, QIS-GAN uses stitching to rudely establish the detail representation, which makes the model lose a large amount of detail information in the process of learning.

Fig. 8 depicts two sets of FST results and their corresponding HMs on the WV3 real test set, encompassing typical scenarios such as small buildings, shrubbery, and cars. The fusion scenario becomes more challenging when the resolution reaches 0.3 m. As evident from the HMs, PanFormer, CMLNet, DCINN, and the proposed RAMSF demonstrate comparable fusion quality. In contrast, PanFormer shows slight spectral distortion as seen from the zoomed-in region. In addition, CMLNet and DCINN exhibit severe spatial distortions, as evidenced by the severe deformation appearing in the vehicle, which is detrimental to the subsequent detection and segmentation tasks. The proposed RAMSF still exhibits clear edges and consistent spectral distributions in highly fine-grained scenes, thanks to the model's emphasis on complementary details of diverse modalities.

The quantitative results of RST and FST on WV3 data are shown in Tables II and III, respectively. It can be seen that the best scores are achieved by the proposed RAMSF, except for QNR. This phenomenon may be attributable to the inherent drawback of QNR, which is that it may treat the increased detail information as spectral distortions, as mentioned in [61] and [62].

C. Experiments for HSIP

Three HSIP datasets are used to further validate the generalization capability of the proposed RAMSF on different ORS-MSF tasks, including the simulated PC dataset with 102 bands, the simulated Botswana dataset with 145 bands, and the real FR1 dataset with 69 bands.

1) PC (102-Band): The qualitative results are shown in Fig. 9. Since EXP is an upsampling method that does not use spatially referenced images, its fusion results demonstrate significant blurring. Among the traditional methods, CNMF, GFPCA, and MTF-GLP demonstrate obvious spatial distortion and spectral distortion. DL-based fusion methods demonstrate similar fusion performance, so we further analyze the subjective performance through AEMs. In contrast, the AEM corresponding to RAMSF has the darkest colors and the least residuals. Specifically, in the building edge region, the fusion results of RAMSF exhibit fewer residuals, which means that our method is capable of preserving more spatial details. In addition, we show the spectral vectors for two different spatial locations in the test sample shown in Fig. 10(a). It can be observed that the spectral vectors of RAMSF are closest to GT, which proves that our method has the best spectral preservation ability. Table IV shows the corresponding quantitative results. RAMSF achieves the best scores on the PSNR and ERGAS, which further proves the effectiveness of RAMSF.

2) Botswana (145-Band): The HSIs in the Botswana test set comprise 145 spectral bands, rendering the sharpening

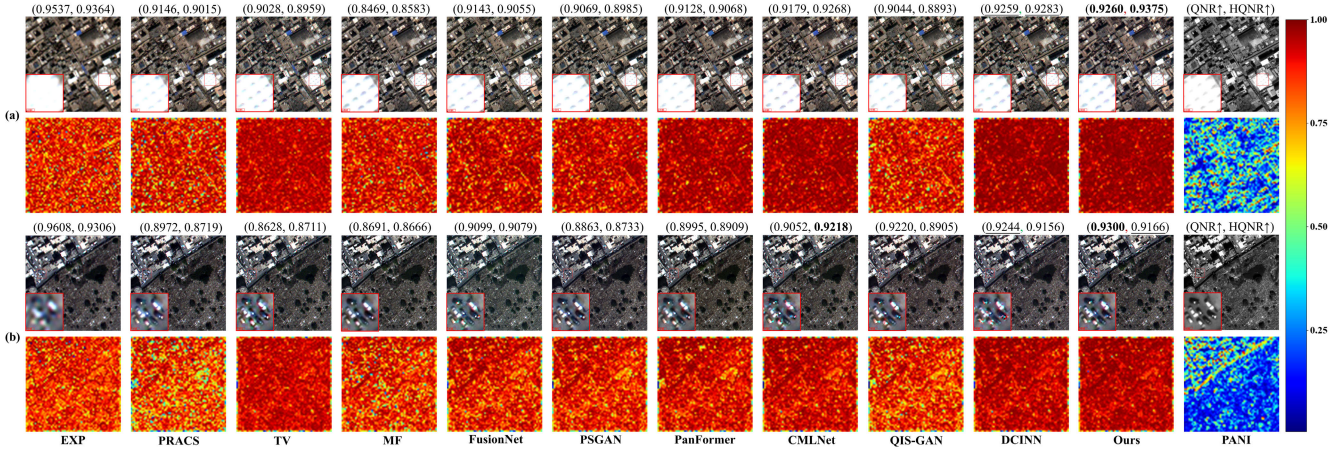

Fig. 8. Qualitative evaluation results on WV3 real test set. The fusion results are presented in odd rows, while the corresponding HMs are listed in even rows. (a) First set of results on the WV3 real test set. (b) Second set of results on the WV3 real test set.

TABLE IV

QUANTITATIVE EVALUATION RESULTS ON HSIP TEST SETS (BOLD: BEST AND UNDERLINE: SECOND BEST)

Method	PC (Simulation data)			Botswana (Simulation data)			FR1 (Real data)		
	PSNR (↑)	SAM (↓)	ERGAS (↓)	PSNR (↑)	SAM (↓)	ERGAS (↓)	D_λ (↓)	D_s (↓)	QNR (↑)
EXP (*TGRS*, 2002)	25.9337	7.1899	9.1698	31.3785	2.0029	3.0295	0.0000	0.0094	0.9906
GSA (*TGRS*, 2007) (CS)	29.2380	7.6170	6.8566	34.8872	1.8206	2.1891	0.0220	0.0688	0.9106
CNMF (*TGRS*, 2012) (VO)	29.4520	6.6936	6.3537	31.9540	2.0324	2.8813	0.0305	0.0576	0.9137
MTF-GLP (*TGRS*, 2017) (MRA)	26.2928	9.9142	8.8554	33.1565	1.9279	2.5634	0.0448	0.0593	0.8986
DARN (*TGRS*, 2020) (II)	33.9589	4.8643	3.9922	37.5380	1.5765	1.8209	0.0287	0.0615	0.9116
MDANet (*TGRS*, 2021) (III)	35.3967	**4.0844**	3.3743	37.8565	1.5151	1.9192	0.0287	0.0475	0.9252
DSNet (*JSTARS*, 2022) (III)	34.4274	4.5454	3.7359	37.0110	1.6464	1.8512	0.0249	0.0476	0.9287
PSRT (*TGRS*, 2023) (II)	35.1552	4.3610	3.4371	37.5537	1.5498	1.8140	0.0479	0.0691	0.8863
QIS-GAN (*TGRS*, 2023) (II)	33.2985	5.1233	4.3069	37.6065	1.5215	1.8727	**0.0219**	0.0625	0.9170
DCINN (*IJCV*, 2024) (I)	34.8466	4.2728	3.4479	37.9243	1.5912	1.7959	0.0222	0.0467	0.9322
Ours	**35.4784**	4.1896	**3.2933**	**38.0786**	**1.5027**	**1.7137**	0.0224	**0.0421**	**0.9364**

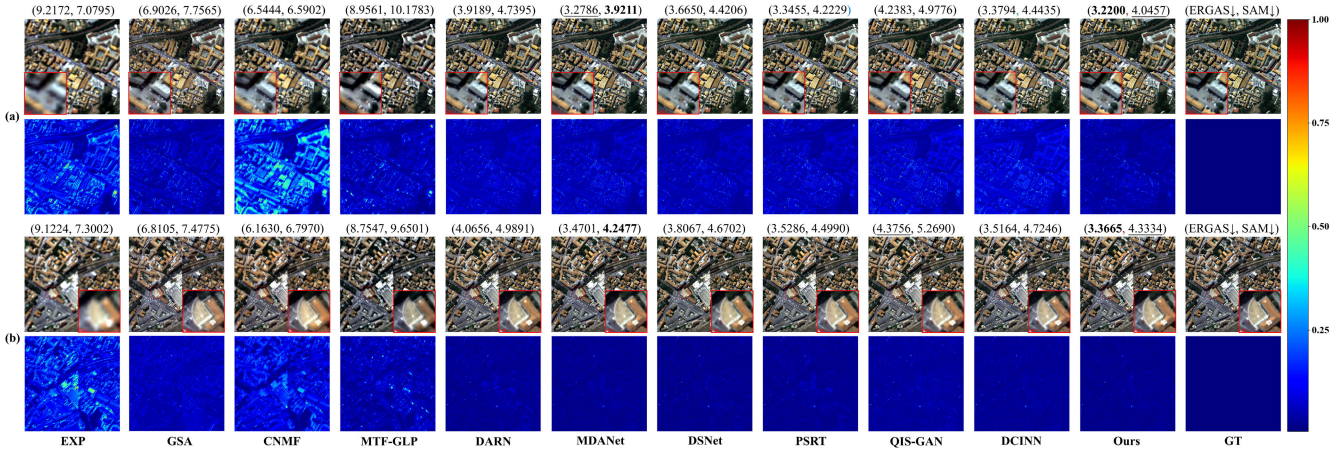

Fig. 9. Qualitative evaluation results on PC test set. The fusion results are presented in odd rows, while the corresponding AEMs are listed in even rows. (a) First set of results on the PC test set. (b) Second set of results on the PC test set.

process more challenging. Fig. 11 shows two sets of results and their corresponding AEMs on the Botswana data, encompassing typical scenarios such as mountains and rocks. Most of the methods exhibit severe performance degradation. Among them, CNMF demonstrates the worst fusion performance. GFPCA and MTF-GLP demonstrate significant spatial and spectral distortions. From the enlarged area depicted in the red box, the proposed RAMSF demonstrates the closest spatial details and spectral distribution to GT. The corresponding AEMs in Fig. 11(a) reveal that only PSRT and the proposed RAMSF demonstrate comparable detail preservation capabilities. PSRT is an image dimension concatenation method that employs transformer as a feature extractor. The benefit of transformer's global modeling capability allows PSRT to achieve decent performance even after rough preprocessing. However, as shown in Fig. 11(b), PSRT loses a lot of edge

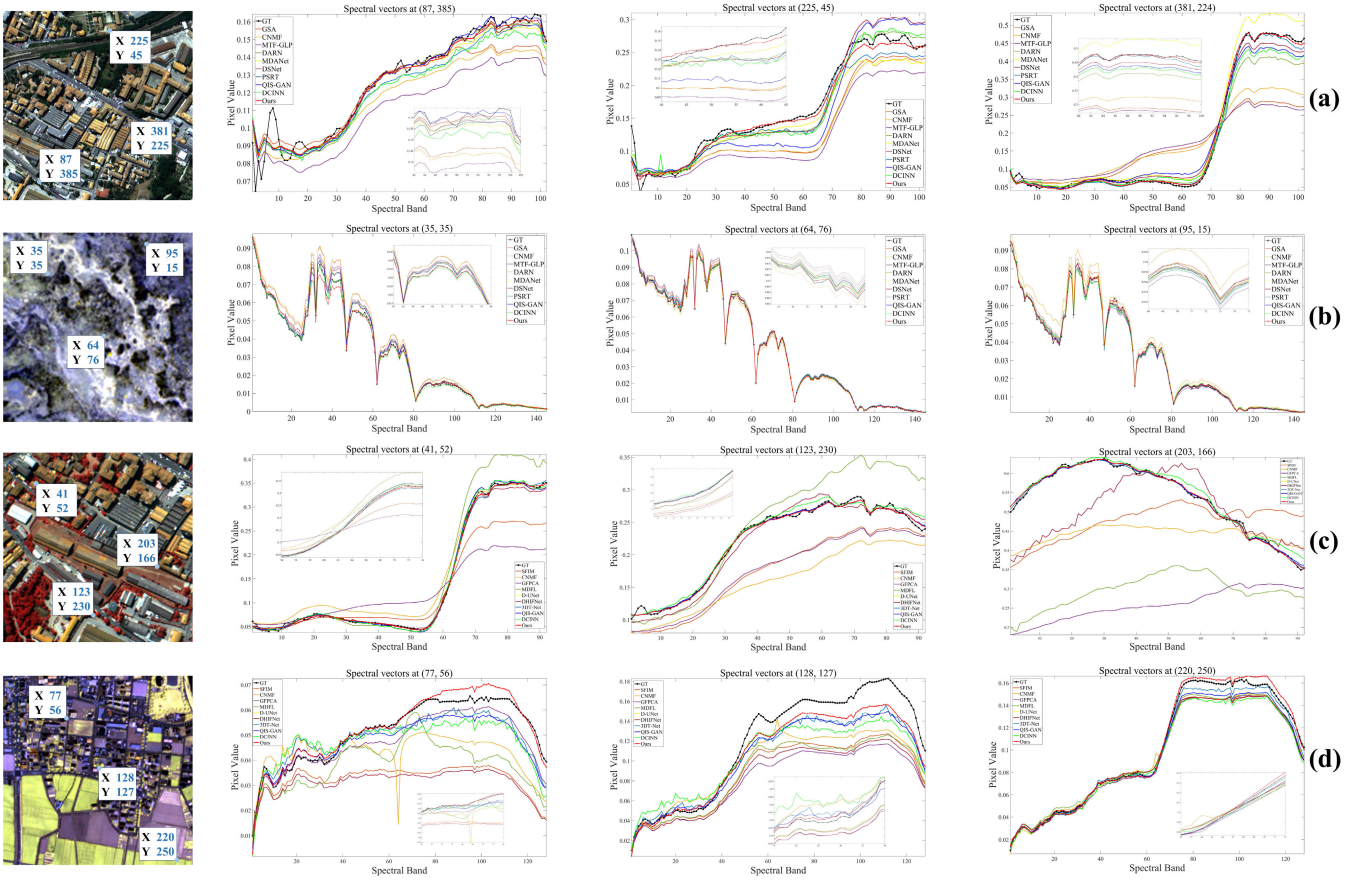

Fig. 10. Comparisons of spectral vectors from different spatial locations on different ORS-MSF data. (a) Spectral vectors from different spatial locations on PC test data. (b) Spectral vectors from different spatial locations on Botswana test data. (c) Spectral vectors from different spatial locations on PU test data. (d) Spectral vectors from different spatial locations on Chikusei test data.

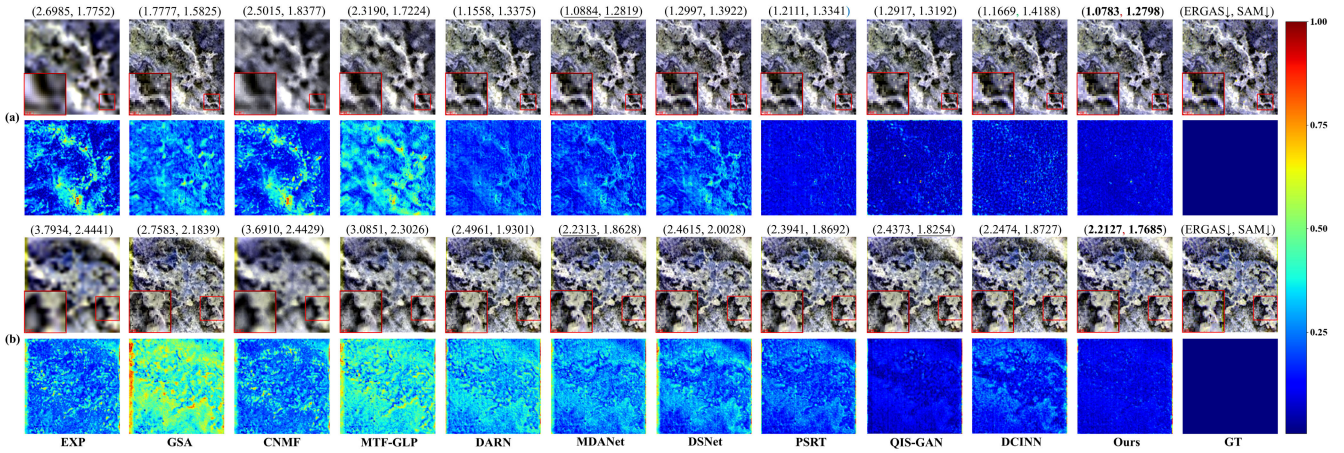

Fig. 11. Qualitative evaluation results on Botswana test set. The fusion results are presented in odd rows, while the corresponding AEMs are listed in even rows. (a) First set of results on the Botswana test set. (b) Second set of results on the Botswana test set.

details when the texture of the mountains and rivers is more abundant, which verifies the importance of establishing a fine detail representation between diverse modal data. In addition, in the second set of examples, except for QIS-GAN and the proposed RAMSF, all the other methods show significant spectral distortion, as evidenced by the colored pixels in the edge regions of the AEMs. This confirms the necessity of preserving the continuity of the spatial–spectral features. The quantitative test results on Botswana data in Table IV validate the effectiveness of decoupling the fusion problem into LHSDR and CSFPA. Similarly, we also show the spectral vectors in the test sample, as shown in Fig. 10(b). It can be seen that the spectral vectors of RAMSF are closest to those of GT, which further demonstrates that our methodology also has the best spectral preservation on the more challenging Botswana dataset.

3) FR1 (69-Band): Fig. 12 depicts the qualitative results on the real FR1 test set. Since HQNR is only applicable to assess

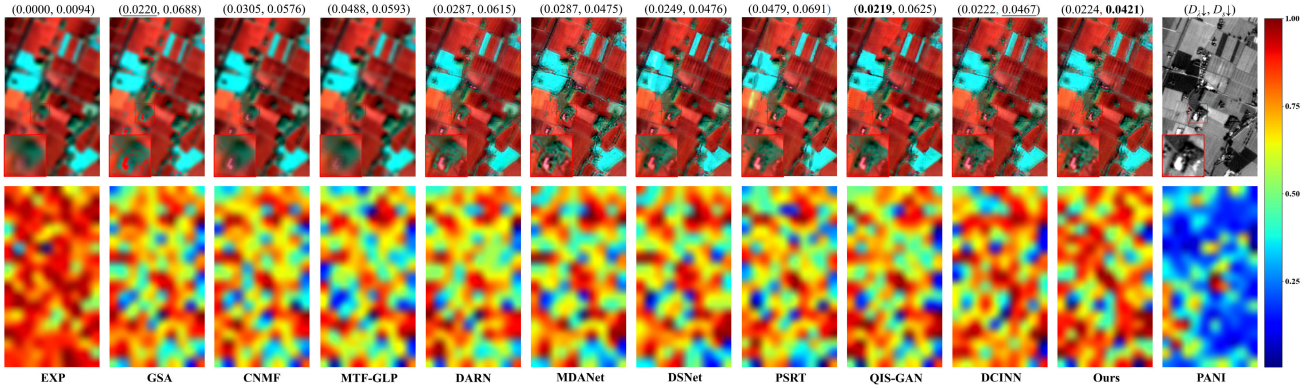

Fig. 12. Qualitative evaluation results on FR1 test set. The fusion results are presented in odd rows, while the corresponding QMs are listed in even rows.

TABLE V

QUANTITATIVE EVALUATION RESULTS ON MHIF TEST SETS (BOLD: BEST AND UNDERLINE: SECOND BEST)

Method	PU (Simulation data)			Chikusei (Simulation data)			ZY1E (Real data)		
	PSNR (↑)	SAM (↓)	ERGAS (↓)	PSNR (↑)	SAM (↓)	ERGAS (↓)	D_λ (↓)	D_s (↓)	QNR (↑)
EXP (*TGRS*, 2002)	25.0169	6.4494	9.6371	37.4279	2.9778	6.6727	0.0000	0.0072	0.9928
SFIM (*IJRS*, 2000) (MRA)	26.2683	7.0703	8.4280	37.6643	2.8862	5.9979	0.0129	0.0319	0.9557
CNMF (*TGRS*, 2012) (VO)	30.6587	5.1807	5.4630	35.6715	4.5624	10.3532	0.0115	0.0221	0.9667
GFPCA (*JSTARS*, 2015) (CS)	25.7571	7.6052	8.8297	37.2628	3.4593	5.8204	0.0083	0.0146	0.9773
MDFL (*NEUCOM*, 2021) (II)	28.0597	6.2106	6.8369	34.9832	3.1631	6.1240	0.0075	0.0097	0.9828
D-UNet (*TGRS*, 2022) (II)	38.6230	3.3862	2.6473	34.5305	2.9753	6.6755	0.0051	0.0127	0.9823
DHIFNet (*TCI*, 2022) (III)	47.4610	1.5594	**0.8632**	38.5241	2.6178	3.8121	0.0051	0.0129	0.9821
3DT-Net (*IF*, 2023) (III)	47.2101	1.5549	0.8958	39.5564	2.2712	3.5121	0.0047	0.0153	0.9800
QIS-GAN (*TGRS*, 2023) (II)	47.3713	1.5631	0.8719	39.3013	2.1978	3.6856	**0.0018**	0.0136	0.9846
DCINN (*IJCV*, 2024) (I)	44.6116	1.9501	1.1001	38.4410	2.5224	3.9038	0.0109	0.0126	0.9766
Ours	**47.5913**	**1.5489**	0.8665	**40.1836**	**2.1476**	**3.3505**	0.0025	**0.0077**	**0.9898**

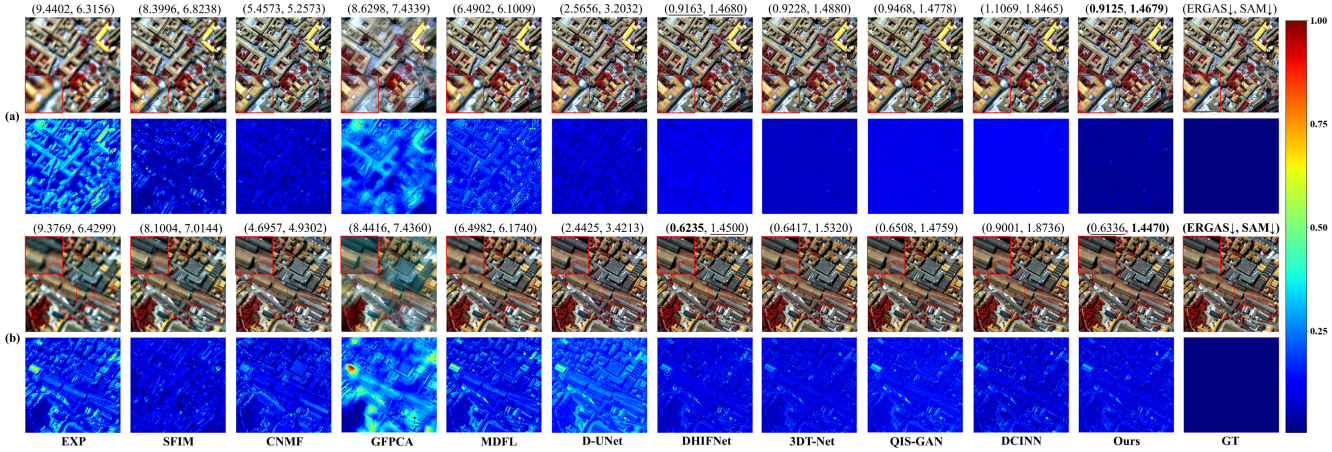

Fig. 13. Qualitative evaluation results on PU test set. The fusion results are presented in odd rows, while the corresponding AEMs are listed in even rows. (a) First set of results on the PU test set. (b) Second set of results on the PU test set.

the comprehensive quality of 4-band or 8-band ORS images, we visualize another comprehensive quality evaluation metric, QNR maps (QMs), to show the scores of each local region. As shown in the zoomed-in region, GSA and CNMF exhibit unrealistic spectral distributions, while MTF-GLP exhibits inferior spatial quality. The fusion outcomes obtained from the image dimension concatenation methods DARN, PSRT, and QIS-GAN exhibit excessive smoothness and lack critical details. This is attributable to the that direct stitching disregards the complementarity of diverse modal data, rendering it challenging to reconstruct a precise representation. The feature dimension concatenation methods MDANet and DSNet demonstrate slight spectral distortion. This is mainly due to the fact that such preprocessing tends to cause a large amount of redundant information, which makes it difficult to mine fine detail representations. In addition, the discrete sampling method used in the feature extraction stage makes it difficult to align multimodal features at various scales, which in turn leads to spectral distortion. DCINN and the proposed RAMSF show comparable performance. In contrast, RAMSF exhibits more consistent spatial details with PANI. The quantitative evaluation results of each method are shown

in Table IV. The proposed RAMSF achieves the best scores in D_s and QNR, which implies that the RAMSF preserves more spatial details and possesses better comprehensive quality. It is worth mentioning that the RAMSF slightly outperforms the QIS-GAN in terms of the spectral distortion index D_λ. In fact, the qualitative results demonstrate that the proposed RAMSF preserves more spatial details. As previously mentioned, since D_λ is using the upsampled HSI as the spectral reference image, it may consider the enhanced spatial details as spectral distortion. Overall, the proposed RAMSF demonstrates the most impressive fusion performance.

D. Experiments for MHIF

For MHIF, we also use three datasets for validation, including the simulated PU dataset with 92 bands, the simulated Chikusei dataset with 128 bands, and the real ZY1E dataset with 76 bands.

1) PU (92-Band): Fig. 13 shows two sets of results and their corresponding AEMs on the PU data, encompassing typical scenarios such as buildings and roads. Compared with the other two classical methods, CNMF exhibits decent fusion performance. For DL methods, the fusion results of MDFL and D-UNet exhibit poor spatial quality. This is due to the fact that the direct concatenation ignores the complementarity of diverse modal data, resulting in the loss of critical details. As shown in the AEMs, the feature dimension concatenation methods DHIFNet and 3DT-Net exhibit high reconstruction errors in the center region of the building. This feature dimension concatenation method tends to cause a large amount of redundant information, which makes it difficult to mine fine detail representations and thus leads to the loss of some key details. QIS-GAN, DCINN, and the proposed RAMSF exhibit comparable fusion quality. In Fig. 13(a), the AEM corresponding to the RAMSF exhibits a darker hue, which implies a less global error. In Fig. 13(b), the AEM corresponding to DCINN exhibits darker hues in some localized regions, while the proposed RAMSF exhibits fewer residuals. The spectral curves distributed at different positions in Fig. 10(c) further validate that the proposed RAMSF can achieve a more consistent spectral distribution with GT. Table V shows the corresponding quantitative results. The proposed RAMSF achieves the highest score on PSNR, which means that its fusion results are clearer and more informative in detail. Meanwhile, the RAMSF achieves the best score on SAM, which means that our method demonstrates a more consistent spectral distribution with GT.

2) Chikusei (128-Band): Fig. 14 shows two sets of results and their corresponding AEMs, encompassing typical scenarios such as buildings, cropland, and rivers. Both from the fusion results and the corresponding AEMs, the traditional methods demonstrate severe spatial blurring and spectral distortion. For DL methods, MDFL and D-UNet also exhibit obvious spatial blurring, as shown in the blurred yellow building edges in Fig. 14(a) and the smooth cropland area in Fig. 14(b). DHIFNet, 3DT-Net, and QIS-GAN show slight distortion, specifically in the center area of the building. This direct concatenation cannot accurately represent the complementary characteristics of the diverse modal data, which results in the loss of some key details. DCINN and the proposed RAMSF exhibit slight distortions, specifically in the blue building area located in the bottom-left corner. In the cropland area, the RAMSF exhibits a finer texture. Furthermore, it can be seen from the corresponding AEMs of each method that the proposed RAMSF possesses a less reconstruction error. The quantitative results are shown in Table V. The proposed RAMSF achieves the best scores on SAM and ERGAS, indicating that our method can preserve more spectral information and spatial details.

3) ZY1E (76-Band): Fig. 15 shows the fusion results and their corresponding QMs on real ZY1E data, encompassing typical scenarios such as shrubbery, coastline, and rivers. In vegetation and cropland areas, traditional methods exhibit inferior fine grainedness, even though their corresponding QMs are more reddish in hue. The fusion results of MDFL exhibit severe distortions. D-UNet, DHIFNet, and 3DT-Net exhibit comparable fusion quality. The fusion outcomes of DCINN exhibit excessive smoothing, resulting in the loss of detail information. DCINN establishes the representation between diverse modal data by means of differencing, ignoring the complementary details in the spectral reference image. In contrast, the proposed RAMSF exhibits the best spatial and spectral preservation capabilities, as evidenced by the rich vegetation texture, clear coastline, and well-defined road boundaries. The test results on the three nonreference metrics are shown in Table V, where the proposed RAMSF achieves the best scores on D_s and QNR. This indicates that the RAMSF is able to preserve more spatial details while attaining a higher overall quality. In addition, the proposed RAMSF performs second only to the QIS-GAN on D_λ. Combining the qualitative evaluation results and the inherent defects of D_λ, we conclude that the proposed RAMSF is still the most promising.

E. Ablation Experiments

Two components of the proposed RAMSF, LHSDR and CSFPA, are the subject of ablation experiments. We conduct corresponding experiments on the MSIP dataset QB, the HSIP dataset PC, and the MHIF dataset Chikusei.

1) Low-Frequency-Driven High-Frequency Salient Detail Reconstruction: In the process of reconstructing the HF details, we take into account the LF information in the spatial reference image ($\mathcal{C}_{\mathrm{LL}}^{Y}$) while disregarding the LF information in the spectral reference image ($\mathcal{C}_{\mathrm{LL}}^{Z}$). We conduct the corresponding experiments to validate the efficacy of LHSDR. Moreover, we compare the proposed LHSDR with the common pixelwise operation in three different ORS-MSF tasks. Fig. 16(a) shows the visualization results. It can be seen that subtraction and concatenation show significant spectral distortion. Since continuous pixelwise operations cannot adequately represent the detail information in the original image pairs, lost detail information changes the overall spectral distribution. From the corresponding AEMs, it can be observed that w $\mathcal{C}_{\mathrm{LL}}^{Z}$ exhibits more residuals, indicating that a portion of the detail information is lost subsequent to the addition of $\mathcal{C}_{\mathrm{LL}}^{Z}$, This implies that the LF information presented LRMSI may inadvertently hinder the process of detail reconstruction. In addition, the fusion results show a slight spectral distortion after the addition of $\mathcal{C}_{\mathrm{LL}}^{Y}$, proving that the fine details can

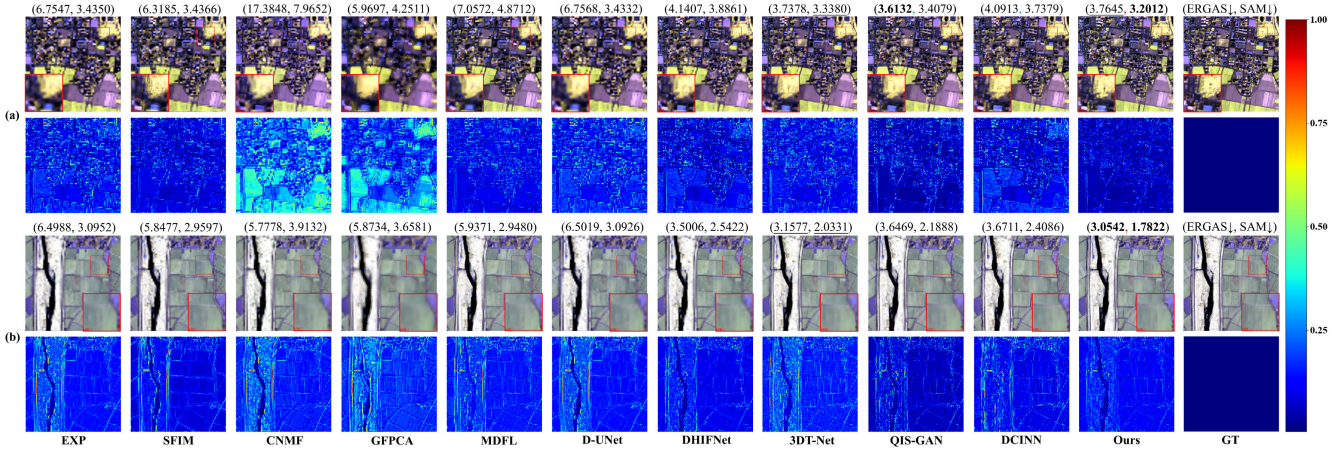

Fig. 14. Qualitative evaluation results on Chikusei test set. The fusion results are presented in odd rows, while the corresponding AEMs are listed in even rows. (a) First set of results on the Chikusei test set. (b) Second set of results on the Chikusei test set.

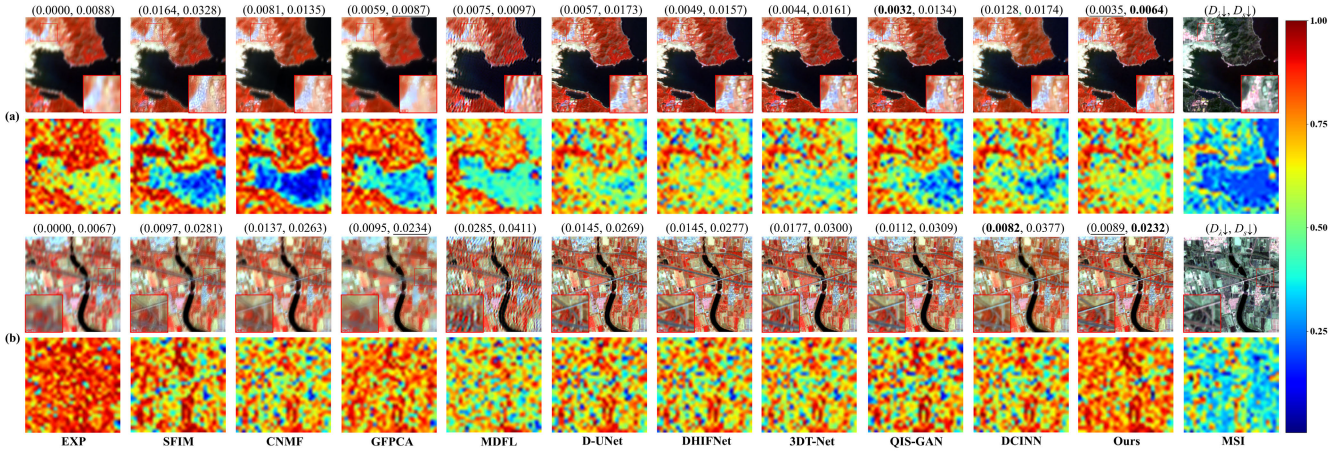

Fig. 15. Qualitative evaluation results on ZY1E test set. The fusion results are presented in odd rows, while the corresponding QMs are listed in even rows. (a) First set of results on the ZY1E test set. (b) Second set of results on the ZY1E test set.

TABLE VI

QUANTITATIVE ABLATION EXPERIMENTAL RESULTS FOR LHSDR ON DIFFERENT ORS-MSF TASKS (BOLD: BEST)

Method	MSIP (4-band)			HSIP (102-band)			MHIF (128-band)		
	SAM (↓)	ERGAS (↓)	Q2n (↑)	PSNR (↑)	SAM (↓)	ERGAS (↓)	PSNR (↑)	SAM (↓)	ERGAS (↓)
Subtraction	5.5314	5.2065	0.9016	33.6155	4.9505	4.1696	38.1526	2.4552	3.9926
Concatenation	5.6499	5.3944	0.8982	33.4821	5.0258	4.2041	39.3518	2.2780	3.6000
w $\mathcal{C}_{LL}^{\mathcal{Z}}$	4.7548	4.0069	0.9258	35.2663	4.2768	3.3911	39.6225	2.1731	3.3984
w/o $\mathcal{C}_{LL}^{\mathcal{Y}}$	4.7625	3.9436	0.9275	35.3282	4.3370	3.3458	40.0326	2.1765	3.3842
LHSDR	**4.6070** (-3.1/3.2/16.7/18.4%)	**3.8127** (-3.3/4.8/26.8/29.3%)	**0.9317** (+0.5/0.6/3.3/3.7%)	**35.4784** (+0.4/0.6/5.5/6.0%)	**4.1896** (-2.0/3.4/15.4/16.6%)	**3.2933** (-1.6/2.9/21.0/21.7%)	**40.1836** (+0.4/1.4/2.1/5.3%)	**2.1476** (-1.2/1.3/5.7/12.5%)	**3.3505** (-1.0/1.4/6.9/16.1%)

TABLE VII

QUANTITATIVE ABLATION EXPERIMENTAL RESULTS FOR CSFPA ON DIFFERENT ORS-MSF TASKS (BOLD: BEST)

Method	MSIP (4-band)			HSIP (102-band)			MHIF (128-band)		
	SAM (↓)	ERGAS (↓)	Q2n (↑)	PSNR (↑)	SAM (↓)	ERGAS (↓)	PSNR (↑)	SAM (↓)	ERGAS (↓)
Bilinear	4.8938	4.1192	0.9231	34.1061	4.6837	3.7112	39.6350	2.2580	3.4500
Bicubic	4.8995	4.1124	0.9258	34.0696	4.5797	3.6934	39.7274	2.2134	3.4080
Deconvolution	4.8046	4.0146	0.9263	35.2664	4.4161	3.3696	39.8737	2.2199	3.4737
PixelShuffle	4.8705	4.0430	0.9253	35.2151	4.4622	3.3753	39.7575	2.2062	3.4139
CSFPA	**4.6070** (-4.1/5.4/5.8/6.0%)	**3.8127** (-5.0/5.7/7.3/7.4%)	**0.9317** (+0.6/0.6/0.7/0.9%)	**35.4784** (+0.6/0.7/4.0/4.1%)	**4.1896** (-5.1/6.2/8.5/10.5%)	**3.2933** (-2.3/2.4/10.8/11.3%)	**40.1836** (+0.8/1.1/1.1/1.4%)	**2.1476** (-2.7/3.0/3.3/4.9%)	**3.3505** (-1.7/1.9/2.8/3.4%)

facilitate the formation of a more reasonable spectral distribution. Furthermore, the outcomes from the HSIP and MHIF tasks confirm this finding. The fusion results of subtraction and concatenation exhibit varying degrees of blurring, as shown by the building edge regions in the enlarged area of the red box. From the corresponding AEMs, the proposed LHSDR shows

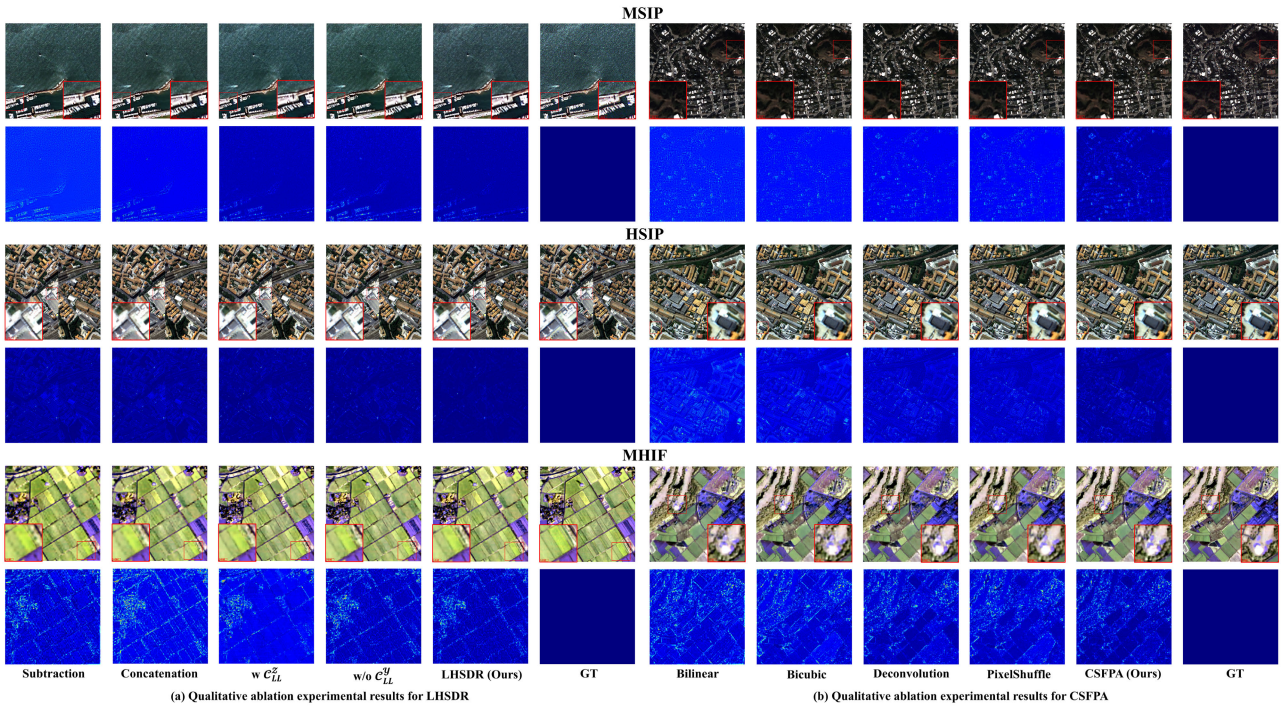

Fig. 16. Ablation experimental results. Qualitative ablation experimental results for (a) LHSDR and (b) CSFPA.

TABLE VIII
QUANTITATIVE ABLATION EXPERIMENTAL RESULTS FOR THE NUMBER OF TIMES OF PROGRESSIVE ALIGNMENT IN CSFPA (BOLD: BEST)

Method	MSIP (4-band)			HSIP (102-band)			MHIF (128-band)		
	SAM (↓)	ERGAS (↓)	Q2n (↑)	PSNR (↑)	SAM (↓)	ERGAS (↓)	PSNR (↑)	SAM (↓)	ERGAS (↓)
T = 0	4.8751	4.0945	0.9266	34.1771	5.1233	4.3069	38.5690	2.6761	3.8773
T = 1	4.7106	3.9103	0.9296	34.6265	5.0186	3.9111	39.4661	2.2301	3.6385
T = 2	**4.6078**	3.8127	0.9317	**35.4784**	4.1896	**3.2933**	40.1836	2.1476	**3.3505**
T = 3	4.6111	**3.7848**	0.9329	35.4452	**4.0539**	3.7680	**40.1943**	2.1320	3.3549
T = 4	4.6362	3.7972	**0.9333**	34.7030	4.4653	3.8955	39.8230	**2.1225**	3.3542

the least residuals. This further demonstrates that $\mathcal{C}_{\mathrm{LL}}^{y}$ facilitates the reconstruction of more fine-grained HF information. However, excessively redundant LF information, such as $\mathcal{C}_{\mathrm{LL}}^{z}$, hinders the reconstruction of details. Table VI presents the corresponding quantitative results, and it is evident that RAMSF exhibits superior performance across all metrics. For the QB dataset, our methodology improves at least 3.1%, 3.3%, and 0.5% on SAM, ERGAS, and $Q2n$ metrics, respectively. For the PC dataset, our method achieves at least 0.4%, 2.0%, and 1.6% improvement on the PSNR, SAM, and ERGAS metrics, respectively. Likewise, for the Chikusei dataset, our RAMSF achieves at least 0.4%, 1.2%, and 1.0% improvement on the PSNR, SAM, and ERGAS metrics, respectively. Therefore, we can conclude that a moderate amount of LF information can drive the formation of more salient HF information.

2) Coordinate-Modal-Guided Spatial–Spectral Feature Progressive Alignment: We perform ablation experiments on CSFPA. The sampling process in CSFPA is replaced with common discrete-based sampling techniques such as bilinear, bicubic, deconvolution, and pixel shuffle. Fig. 16(b) illustrates the qualitative results on three different ORS-MSF tasks. For MSIP, the enlarged area shows that the proposed CSFPA is capable of generating spatial information and spectral distributions closest to GT. Furthermore, the corresponding AEMs further demonstrate that the CSFPA preserves more spatial detail information. The proposed CSFPA demonstrates the closest spatial and spectral distribution of the GT, as shown in the building area in the red box zoomed-in region. Table VII presents the corresponding quantitative results. For the QB dataset, RAMSF improves at least 4.1%, 5.0%, and 0.6%, on the SAM, ERGAS, and $Q2n$ metrics, respectively. For the PC dataset, our methodology improves at least 0.6%, 5.1%, and 2.3% on the PSNR, SAM, and ERGAS metrics, respectively. Similarly, for the PC dataset, our methodology improves by at least 0.8%, 2.7%, and 1.7% on the PSNR, SAM, and ERGAS metrics, respectively. In summary, CSFPA is capable of accurately aligning diverse modal data while ensuring the continuity of spectral features. These reasonable spectral distributions can promote the formation of finer details.

We have performed an ablation analysis of the number of times of progressive alignment (T) in CSFPA. As depicted in Table VIII, the fusion performance in different ORS-MSF tasks shows a consistent upward trend as T increases. However, when $T > 2$, the upward trend of fusion performance slows down. Taking into consideration the efficiency burden associated with the increase of T, we choose $T = 2$ as the number of times of progressive alignment.

TABLE IX

QUANTITATIVE ABLATION EXPERIMENTAL RESULTS FOR DIFFERENT DETAIL REPRESENTATION AND FUSION STRATEGY COMBINATIONS (BOLD: BEST)

Combinations	MSIP (4-band)			HSIP (102-band)			MHIF (128-band)		
	SAM (↓)	ERGAS (↓)	Q2n (↑)	PSNR (↑)	SAM (↓)	ERGAS (↓)	PSNR (↑)	SAM (↓)	ERGAS (↓)
QIS-GAN	4.6785	3.9063	0.9241	33.2985	5.1233	4.3069	39.3013	2.1978	3.6856
RAMSF with QIS	4.6625	3.8669	0.9288	34.5625	4.5165	3.3264	39.8377	2.1633	3.5098
QIS-GAN with CSFPA	4.6311	3.8748	0.9309	34.6405	4.7124	3.3203	40.0195	2.1656	3.4900
RAMSF	**4.6070**	**3.8127**	**0.9317**	**35.4784**	**4.1896**	**3.2933**	**40.1836**	**2.1476**	**3.3505**

TABLE X

EFFICIENCY EVALUATION ON DIFFERENT ORS-MSF TASKS (BOLD: BEST)

Method	Epochs	Training time (h)	Testing time (s)	Memory (MB)	FLOPs (G)	MACs (G)	Params (M)
MSIP							
FusionNet (*TGRS*, 2021) (I)	600	**5.20**	**0.005**	**34.78**	2.45	1.23	0.31
PSGAN (*TGRS*, 2021) (III)	**50**	9.03	0.038	46.90	4.07	2.04	2.38
PanFormer (*ICME*, 2022) (III)	400	79.18	0.039	101.59	**1.47**	**0.74**	1.52
CMLNet (*TGRS*, 2023) (II)	1000	10.60	0.026	294.15	2.68	1.34	0.33
QIS-GAN (*TGRS*, 2023) (II)	1500	56.59	0.042	60.82	3.35	1.67	0.52
DCINN (*IJCV*, 2024) (I)	350	77.53	0.061	1.65×10^4	1.58	0.79	0.46
Ours	400	7.20	0.017	43.39	2.61	1.31	**0.24**
HSIP							
DARN (*TGRS*, 2020) (II)	1500	2.122	**0.015**	32.72	3.79	1.90	0.47
MDANet (*TGRS*, 2021) (III)	**300**	3.913	0.032	688.68	88.74	44.37	11.43
DSNet (*JSTARS*, 2022) (III)	2000	3.301	**0.015**	**20.00**	**1.95**	**0.98**	**0.25**
PSRT (*TGRS*, 2023) (II)	2000	15.13	0.037	386.47	2.63	1.32	0.34
QIS-GAN (*TGRS*, 2023) (II)	1000	5.96	0.039	22.41	14.64	7.32	0.73
DCINN (*IJCV*, 2024) (I)	350	44.52	0.221	2.63×10^5	21.53	10.76	1.89
Ours	400	**1.358**	0.025	46.63	5.86	2.93	0.84
MHIF							
MDFL (*NEUCOM*, 2021) (II)	200	10.64	0.196	**7.42**	429.12	214.56	1.93
D-UNet (*TGRS*, 2022) (II)	200	**1.36**	**0.011**	162.43	12.28	6.14	7.27
DHIFNet (*TCI*, 2022) (III)	500	12.7	0.018	2.82×10^{12}	530.11	265.06	35.55
3DT-Net (*IF*, 2023) (III)	**150**	11.24	0.234	9.05×10^6	148.19	74.09	4.25
QIS-GAN (*TGRS*, 2023) (II)	1000	8.08	0.035	91.90	12.78	6.39	2.51
DCINN (*IJCV*, 2024) (I)	350	27.72	0.143	3.33×10^4	21.08	10.54	8.52
Ours	400	2.55	0.016	45.47	**5.38**	**2.69**	**0.78**

TABLE XI

EVALUATION OF GENERALIZABILITY ON THE WV2 DATASET (BOLD: BEST AND UNDERLINE: SECOND BEST)

Method	SAM↓	ERGAS↓	$Q8$↑
EXP (*TGRS*, 2002)	6.5870	6.8123	0.6401
PRACS (*TGRS*, 2011) (CS)	6.3483	5.3019	0.7668
TV (*GRSL*, 2014) (VO)	6.8319	5.0652	0.7984
MF (*TIP*, 2016) (MRA)	6.1838	<u>4.6625</u>	0.8169
FusionNet (*TGRS*, 2021) (I)	6.3901	4.8320	0.8136
PSGAN (*TGRS*, 2021) (III)	6.4454	5.3880	0.8175
PanFormer (*ICME*, 2022) (III)	7.1750	5.5227	0.8170
CMLNet (*TGRS*, 2023) (II)	6.3543	4.9603	0.8074
QIS-GAN (*TGRS*, 2023) (II)	6.3997	5.9108	0.7824
DCINN (*IJCV*, 2024) (I)	<u>5.6653</u>	4.7975	<u>0.8214</u>
Ours	**5.5011**	**4.4668**	**0.8345**

Furthermore, the necessity of the two phases in the proposed framework is further verified. We compare the proposed method with another INR-based image dimension concatenation method, QIS-GAN, both in terms of module and network architecture. Four different variants are obtained by combining different detail representations and fusion strategies. Table IX shows that the proposed RAMSF achieves the best fusion performance with different combinations of modules and network architectures, which further proves the necessity and superiority of the proposed LHSDR and CSFPA.

F. Efficiency Analysis

Table X presents the results of efficiency testing in different ORS-MSF tasks in terms of five aspects, including the number of training epochs (epochs), time complexity (training time and testing time), runtime memory occupation (memory), computational complexity [floating-point operations (FLOPs) and multiply-accumulate operations (MACs)], and model complexity (Params). In addition, Fig. 17 presents a visual comparison of the relationship between the fusion performance and fusion efficiency. As shown in Fig. 17(a), we demonstrate the relationship between performance, training time, and computational complexity (FLOPs) in three different tasks. It is evident that RAMSF achieves the optimal balance between fusion performance, training time, and computational complexity in all three different tasks. Fig. 17(b) illustrates the relationship between performance, testing time, and model complexity (Params). In the MSIP task, the proposed RAMSF ranks second only to FusionNet in terms of testing time while exhibiting a notable advantage in fusion performance on real data. Moreover, RAMSF achieves the optimal balance in the HSIP and MHIF tasks. To summarize, the proposed RAMSF exhibits great potential and promise in three different ORS-MSF tasks.

G. Generalizability Analysis

It has been demonstrated that the proposed RAMSF has superior fusion performance and efficiency in different ORS-MSF tasks. In this section, we focus on the generalization ability of the model, i.e., the ability to fuse data acquired from different sensors through a single model. To this end,

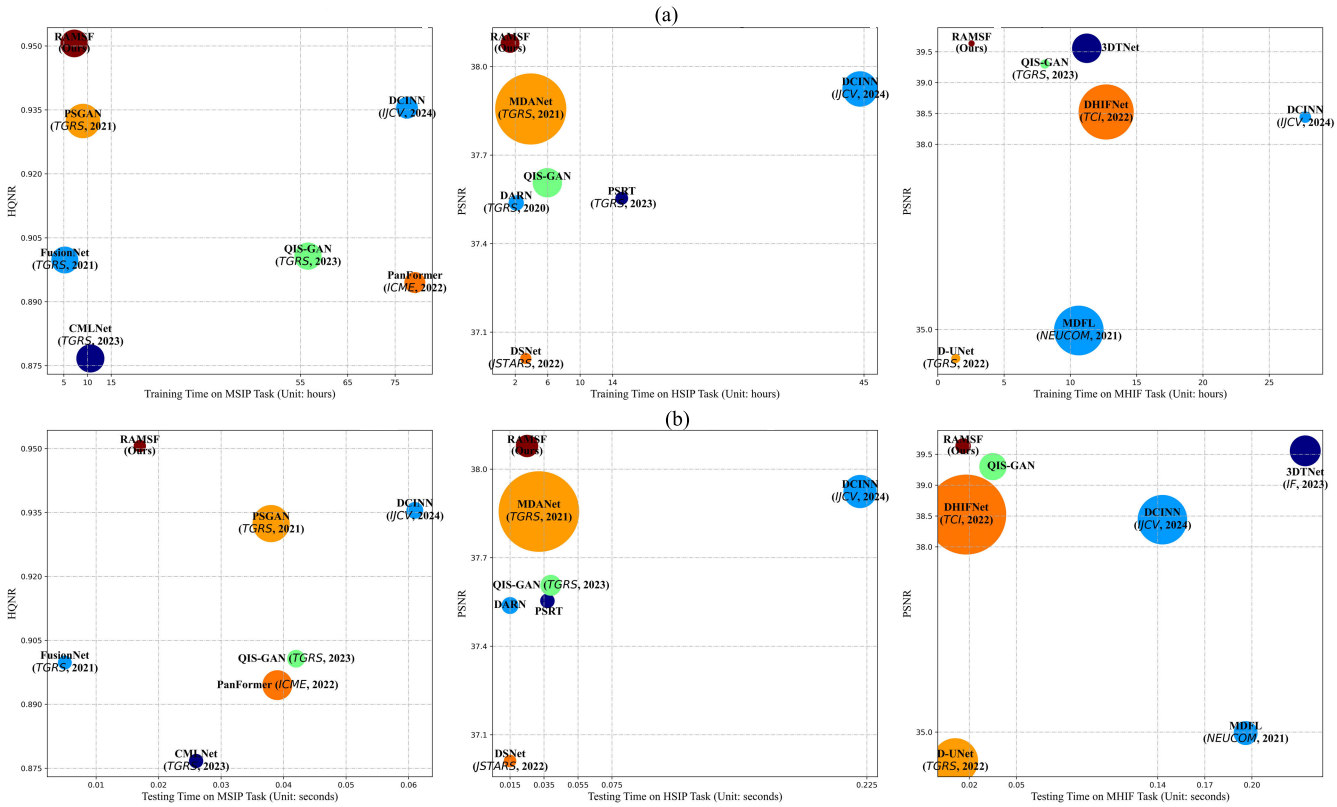

Fig. 17. Comparison of the efficiency of existing SOTA DL-based methods in diverse ORS-MSF tasks. (a) Comparison of fusion performance, training time, and computational complexity. The area of the nodes denotes the computational complexity (FLOPs). (b) Comparison of fusion performance, testing time, and model complexity. The area of the nodes denotes the model complexity (Params).

we evaluate the models trained on WV3 data with each comparison method on WV2 test data. The results of the quantitative experiments are shown in Table XI. It is evident that the methods show obvious performance degradation, particularly CMLNet and QIS-GAN. In contrast, the proposed RAMSF emerges as the most superior method, surpassing the benchmarks in three metrics, specifically SAM, ERGAS, and $Q8$, indicating that its fusion outcomes possess fewer reconstruction errors and superior sharpening quality in both spatial and spectral dimensions.

V. Conclusion

In this article, we provide a comprehensive analysis of the theoretical models and network architectures implemented in existing methods and decompose ORS-MSF into two main phases: detail reconstruction and feature alignment. Inspired by these analyses, we propose a generic framework, called RAMSF. The proposed RAMSF comprises two fundamental components, namely, the LHSDR and the CSFPA. LHSDR estimates the joint spatial degradation process in various frequency directions from diverse modal data and derives salient details in a hierarchical integration, with HF driving LF. The reconstructed salient details can lay the foundation for the subsequent high-fidelity fusion. CSFPA, on the other hand, estimates the joint spectral degradation process by establishing coordinate-mode relations between coupled high-frequency details and corresponding spectral information in the continuous domain. CSFPA is capable of accurately aligning diverse modal data while ensuring the continuity of spectral features. We conduct extensive ablation experiments and comparison experiments in three different ORS-MSF tasks, and the qualitative and quantitative results show that the proposed RAMSF can achieve superior fusion results. Furthermore, efficacy evaluations further demonstrate that the proposed RAMSF has achieved the optimal balance between fusion performance and efficacy. However, in terms of testing time, the proposed method still exhibits room for improvement. In future work, we will further improve the fusion performance and efficacy. Furthermore, we will expand our methodology to other modal remote sensing data, such as synthetic aperture radar, light laser detection and ranging (LiDAR), and thermal infrared data, to further integrate the advantageous information of diverse modal data and establish a solid foundation for downstream tasks and real-world applications.

References

[1] M. A. Moharram and D. M. Sundaram, "Land use and land cover classification with hyperspectral data: A comprehensive review of methods, challenges and future directions," *Neurocomputing*, vol. 536, pp. 90–113, Jun. 2023.

[2] Y. Himeur, B. Rimal, A. Tiwary, and A. Amira, "Using artificial intelligence and data fusion for environmental monitoring: A review and future perspectives," *Inf. Fusion*, vols. 86–87, pp. 44–75, Oct. 2022.

[3] S. Salcedo-Sanz et al., "Machine learning information fusion in Earth observation: A comprehensive review of methods, applications and data sources," *Inf. Fusion*, vol. 63, pp. 256–272, Nov. 2020.

[4] X. Meng et al., "A large-scale benchmark data set for evaluating pan-sharpening performance: Overview and implementation," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 1, pp. 18–52, Mar. 2021.

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

[6] V. K. Shettigara, "A generalized component substitution technique for spatial enhancement of multispectral images using a higher resolution data set," *Photogramm. Eng. Remote Sen.*, vol. 58, no. 5, pp. 561–567, 1992.

[7] V. P. Shah, N. H. Younan, and R. L. King, "An efficient pan-sharpening method via a combined adaptive PCA approach and contourlets," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1323–1335, May 2008.

[8] W. J. Carper, "The use of intensity-hue-saturation transformations for merging SPOT panchromatic and multispectral image data," *Photogramm. Eng. Remote Sens.*, vol. 56, no. 4, pp. 457–467, 1990.

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

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

[11] 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.

[12] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989.

[13] L.-J. Deng, G. Vivone, C. Jin, and J. Chanussot, "Detail injection-based deep convolutional neural networks for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6995–7010, Aug. 2021.

[14] 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, pp. 12631–12648, Jun. 2024.

[15] W. Wang, L.-J. Deng, R. Ran, and G. Vivone, "A general paradigm with detail-preserving conditional invertible network for image fusion," *Int. J. Comput. Vis.*, vol. 132, no. 4, pp. 1029–1054, Apr. 2024, doi: 10.1007/s11263-023-01924-5.

[16] Y. Zheng, J. Li, Y. Li, J. Guo, X. Wu, and J. Chanussot, "Hyperspectral pansharpening using deep prior and dual attention residual network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 8059–8076, Nov. 2020.

[17] J.-D. Wang, L.-J. Deng, C.-Y. Zhao, X. Wu, H.-M. Chen, and G. Vivone, "Cascadic multireceptive learning for multispectral pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5408416.

[18] J. Xiao, J. Li, Q. Yuan, and L. Zhang, "A dual-UNet with multistage details injection for hyperspectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5515313.

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

[20] Q. Liu, H. Zhou, Q. Xu, X. Liu, and Y. Wang, "PSGAN: A generative adversarial network for remote sensing image pan-sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10227–10242, Dec. 2021.

[21] P. Guan and E. Y. Lam, "Multistage dual-attention guided fusion network for hyperspectral pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5515214.

[22] H. Zhou, Q. Liu, D. Weng, and Y. Wang, "Unsupervised cycle-consistent generative adversarial networks for pan sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5408814.

[23] H. Zhou, Q. Liu, and Y. Wang, "PanFormer: A transformer based model for pan-sharpening," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2022, pp. 1–6, doi: 10.1109/ICME52920.2022.9859770.

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

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

[26] W. Wang, W. Zeng, Y. Huang, X. Ding, and J. Paisley, "Deep blind hyperspectral image fusion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4149–4158.

[27] 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.

[28] Q. Wei, J. Bioucas-Dias, N. Dobigeon, and J.-Y. Tourneret, "Hyperspectral and multispectral image fusion based on a sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3658–3668, Jul. 2015.

[29] 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.

[30] L. Chen, G. Vivone, J. Qin, J. Chanussot, and X. Yang, "Spectral–spatial transformer for hyperspectral image sharpening," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 11, pp. 16733–16747, Aug. 2023.

[31] L. Chen, Z. Lai, G. Vivone, G. Jeon, J. Chanussot, and X. Yang, "ArbRPN: A bidirectional recurrent pansharpening network for multispectral images with arbitrary numbers of bands," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5406418.

[32] L. Chen, G. Vivone, Z. Nie, J. Chanussot, and X. Yang, "Spatial data augmentation: Improving the generalization of neural networks for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5401711.

[33] W. Diao, F. Zhang, H. Wang, W. Wan, J. Sun, and K. Zhang, "HLF-Net: Pansharpening based on high- and low-frequency fusion networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[34] Y.-W. Zhuo, T.-J. Zhang, J.-F. Hu, H.-X. Dou, T.-Z. Huang, and L.-J. Deng, "A deep-shallow fusion network with multidetail extractor and spectral attention for hyperspectral pansharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 7539–7555, 2022.

[35] S. Seo et al., "UPSNet: Unsupervised pan-sharpening network with registration learning between panchromatic and multi-spectral images," *IEEE Access*, vol. 8, pp. 201199–201217, 2020.

[36] L. Liu, Z. Zou, and Z. Shi, "Hyperspectral remote sensing image synthesis based on implicit neural spectral mixing models," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5500514.

[37] Y. Chen, S. Liu, and X. Wang, "Learning continuous image representation with local implicit image function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8624–8634.

[38] L. Shen, J. Pauly, and L. Xing, "NeRP: Implicit neural representation learning with prior embedding for sparsely sampled image reconstruction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 1, pp. 770–782, Jan. 2024.

[39] C. Jiang, A. Sud, A. Makadia, J. Huang, M. Nießner, and T. Funkhouser, "Local implicit grid representations for 3D scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6000–6009.

[40] Z. Chen et al., "VideoINR: Learning video implicit neural representation for continuous space-time super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2037–2047.

[41] A. Gropp, L. Yariv, N. Haim, M. Atzmon, and Y. Lipman, "Implicit geometric regularization for learning shapes," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 3789–3799.

[42] H. Chen et al., "Spectral-wise implicit neural representation for hyperspectral image reconstruction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 5, pp. 3714–3727, May 2024.

[43] C. Zhu, S. Deng, Y. Zhou, L.-J. Deng, and Q. Wu, "QIS-GAN: A lightweight adversarial network with quadtree implicit sampling for multispectral and hyperspectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5531115.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[45] 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.

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

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

[48] R. Li, L. Zhang, Z. Wang, and X. Li, "MIMFormer: Multiscale inception mixer transformer for hyperspectral and multispectral image fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 15122–15135, 2024.

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

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

[51] 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.

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

[53] B. Aiazzi, L. Alparone, S. Baronti, R. Carlà, A. Garzelli, and L. Santurri, "Full scale assessment of pansharpening methods and data products," *Remote Sens.*, vol. 9244, Oct. 2014, Art. no. 924402.

[54] 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.

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

[56] 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.

[57] R. Restaino, G. Vivone, M. D. Mura, and J. Chanussot, "Fusion of multispectral and panchromatic images based on morphological operators," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2882–2895, Jun. 2016.

[58] W. Liao et al., "Processing of multiresolution thermal hyperspectral and digital color data: Outcome of the 2014 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2984–2996, Jun. 2015.

[59] J. G. Liu, "Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details," *Int. J. Remote Sens.*, vol. 21, no. 18, pp. 3461–3472, Jan. 2000.

[60] Q. Li, Y. Yuan, and Q. Wang, "Hyperspectral image super-resolution via multi-domain feature learning," *Neurocomputing*, vol. 472, pp. 85–94, Feb. 2022.

[61] Z. Zhang, C. Liu, L. Wei, and S. Xiang, "MMAPP: Multibranch and multiscale adaptive progressive pyramid network for multispectral image pansharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 20129–20149, 2024.

[62] H. Zhang, H. Wang, X. Tian, and J. Ma, "P2Sharpen: A progressive pansharpening network with deep spectral transformation," *Inf. Fusion*, vol. 91, pp. 103–122, Mar. 2023.

Chuang Liu is currently pursuing 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.

Zhiqi Zhang (Member, IEEE) received the B.Sc. degree in geographic information system from 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 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. His research interests include system architecture, algorithm optimization, AI, and high-performance processing of remote sensing.

Mi Wang (Member, IEEE) received the B.Eng., M.Sc., and Ph.D. degrees in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1997, 1999, and 2001, respectively.

Since 2008, he has been a Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University. His research interests include measurable seamless stereo orthoimage databases, geographic information systems (GIS), and high-precision remote sensing image processing.