Rethinking Pan-Sharpening via Spectral-Band Modulation

Xinyang Liu, Junming Hou, *Student Member, IEEE*, Xiaofeng Cong, Hao Shen, Zhuochen Lou, Liang-Jian Deng, *Senior Member, IEEE*, and Jian Wei You, *Senior Member, IEEE*

***Abstract*— Pan-sharpening aims to super-resolve the low-resolution (LR) multispectral (MS) image under the guidance of a high-resolution (HR) panchromatic (PAN) image. Existing deep learning (DL)-based pan-sharpening methods usually adhere to a common philosophy of learning complementary information between MS and PAN images. Despite remarkable advances, few studies consider the band-private characteristics which differ greatly from band to band. An ideal MS image, however, is jointly determined by its diverse spectral bands, thus the accurate restoration of every band will benefit the pan-sharpening performance. In this work, we propose a novel yet effective solution to reconstruct the HRMS image by explicitly modulating every spectral band under the conditions of the PAN image. As a result, we design a spatially-adaptive spectral modulation network, dubbed SSMNet, which consists of three core designs: source-aware spectral modulator (SSM), cross-band information aggregation (CBIA) module, and cross-stage feature integration (CSFI) module. The first predicts a series of spatially-adaptive kernels to capture the local information of every spectral band. Followed by, the second is responsible for facilitating the information communication among various bands to guarantee continuous spectral representations. Furthermore, the third attends to integrate the cross-stage output features to produce the pan-sharpened result. In addition, we also introduce the histogram loss to constrain the band-wise distribution of the final fused products. Extensive experiments demonstrate that our SSMNet achieves favorable performance against other state-of-the-art (SOTA) methods on multiple satellite datasets. The code is available at https://github.com/ez4lionky/SSMNet/.**

***Index Terms*— Band-private characteristics, deep learning (DL), histogram loss, pan-sharpening, spectral modulation.**

Manuscript received 1 September 2023; revised 30 October 2023; accepted 23 November 2023. Date of publication 7 December 2023; date of current version 29 December 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFB3813100, in part by the National Natural Science Foundation of China under Grant 62101124 and Grant 62288101, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20210209 and Grant BK20220808, in part by the Fundamental Research Funds for the Central Universities under Grant 2242023K5002, and in part by the Startup Research Fund of Southeast University under Grant RF1028623244. *(Xinyang Liu and Junming Hou contributed equally to this work.) (Corresponding authors: Liang-Jian Deng; Jian Wei You.)*

Xinyang Liu is with the State Key Laboratory of Millimeter Waves, School of Information Science and Engineering, Southeast University, Nanjing 210096, China, and also with the Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University, Hong Kong, SAR (e-mail: codex.lxy@gmail.com).

Junming Hou and Jian Wei You are with the State Key Laboratory of Millimeter Waves, School of Information Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: junming_hou@seu.edu.cn; jvyou@seu.edu.cn).

Xiaofeng Cong is with the School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: cxf_svip@163.com).

Hao Shen is with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China (e-mail: haoshenhs@gmail.com).

Zhuochen Lou is with the Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: zhuoclou@umich.edu).

Liang-Jian Deng is with the School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: liangjian.deng@uestc.edu.cn).

Digital Object Identifier 10.1109/TGRS.2023.3340193

I. Introduction

HIGH spatial resolution multispectral (HRMS) images are widely used in various fields [3], [4], such as land-cover classification [5], environmental protection [6] and change detection [7], [8], and so forth. However, existing multispectral (multispectral) sensors cannot directly obtain HRMS images due to their physical and technical limitations. Therefore, the common alternative is to equip remote sensing satellites with two types of imaging sensors, which can capture two modalities of images of the same scene, i.e., texture-rich panchromatic (PAN) images and high-spectral resolution MS images. To be specific, MS images usually hold desired spectral information but limited spatial resolution; while PAN images embrace rich spatial details but poor spectral resolution. Therefore, the pan-sharpening technique is developed to produce texture-rich MS images by integrating the complementary information of these two modalities of images. In other words, an ideally sharpened MS image should be spatially consistent with the PAN image while avoiding spectral distortions. In light of this, PAN image can be regarded as a high-resolution (HR) guidance to super-resolve low-resolution (LR) MS images in pan-sharpening learning.

Treated as a PAN-guided MS super-resolution task, pan-sharpening is an ill-posed and challenging problem [9], [10]. The up-to-date solutions to a pan-sharpening problem can be roughly divided into two large categories: traditional methods and DL-based methods [3], [4], [11]. Traditional approaches are usually based on specific assumptions, for example, most of them regard the PAN image as a linear combination of all spectral bands of HRMS image [3], [12], [13]. In other words, the model performance is highly dependent on the exact assumptions. However, prior traditional methods, such as component substitution (CS)-based methods [14], [15] and multiresolution analysis (MRA)-based methods [16], [17], fail to establish the accurate relationship between HRMS and PAN images, thus suffering from significant spectral or spatial distortions. Unlike CS and MRA-based methods, variational optimization (VO)-based techniques [18], [19], taking the relationships among LRMS, PAN, and HRMS images into

account, have elegant mathematical expression, and achieve a better tradeoff between spectral-spatial preservation. Nevertheless, the high computational burden hinders their practical applications. More recently, DL-based pan-sharpening methods that are mainly based on convolutional neural networks (CNNs) have shown great superiority to their conventional counterparts due to the accessibility of large-scale remote sensing images. Masi et al. [20] is the first to employ CNN to implement pan-sharpening learning. Though the network architecture is very simple, it still achieves favorable results against traditional methods. Yang et al. [21] developed the first deep pan-sharpening network based on the residual learning [22]. Since then, more complicated and deeper CNN architectures [23], [24], [25], [26] have been exploited to enhance the pan-sharpening performance. Despite the dramatic progress made by DL-based methods, most of them follow a common fusion paradigm that mainly focuses on the entire image, simply ignoring the band-private local characteristics. Compared with commonly used RGB images, nonetheless, the main reason MS images embrace more abundant and diversified information is attributed to their multiple spectral bands which reflect various contents of the global surface [27], [28], [29], [30]. In addition, the cross-band information is complementary and indivisible for remote sensing image interpretation. For example, some bands, e.g., R, G, and B, usually contain similar color and texture features, rendering it difficult to discriminate. Therefore, we can employ other bands with significant spectral heterogeneity, like the NIR band, to understand the imagery scenes. In other words, an informative MS image is jointly determined by its various spectral bands. Fig. 1 presents the pixel statistics of every sharpened band from different DL techniques and the corresponding reference band. From Fig. 1, it is clearly observed that there are obvious differences in the distribution patterns of different bands, demonstrating the significant band-private characteristics in MS image. In view of this, it is necessary to accurately estimate every spectral band, thus obtaining an ideal MS image with abundant spatial textures and spectral information.

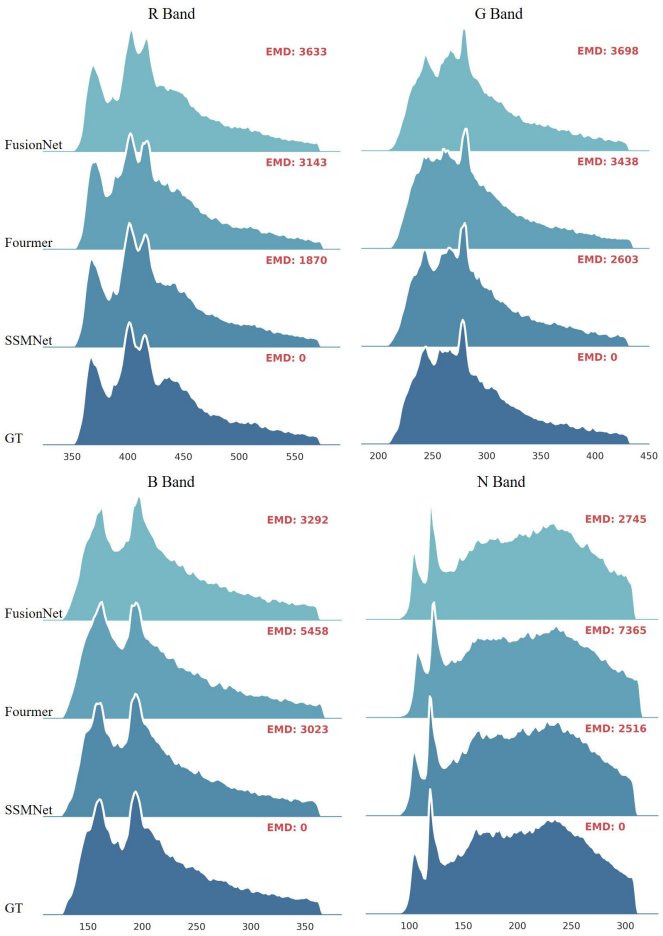

Fig. 1. **Our motivation.** The illustration of band-wise distribution statistics on a reduced-resolution sample from the GaoFen-2 dataset, where two state-of-the-art (SOTA) deep learning (DL)-based methods, i.e., FusionNet [1] and Fourmer [2], are selected for comparison. "R," "G," "B," and "N" represent the red, green, blue, and near-infrared (NIR) bands in the MS image, respectively. The horizontal axis represents the pixel values of each band, while the vertical axis denotes the number of pixels with the corresponding value. We also show the Earth mover's distance (EMD) between the corresponding histogram and reference distribution. The smaller EMD value indicates the higher similarity between the fused band and corresponding reference one (please zoomed-in view to see more details of each distribution pattern).

A. Our Motivation

In the field of pan-sharpening, some attempts have been made to explore the potentials of band-specific characteristics [10], [31], and have achieved promising performance gains. The common strategy of these methods is to simply concatenate the PAN images (or PAN features) with every band (or band features) of MS images to enhance the band-wise information. Despite they have validated the effectiveness of exploiting the band-private characteristics of MS images to benefit the fusion performance, there are some limitations: 1) it is hard to effectively take care of the local features of every band which are precisely the main differences among them since different bands share the similar global structure [13] and 2) they still focus on the image-wise fusion effect while ignoring the differences between every fused spectral band and the corresponding reference band.

To remedy the above issues, in this work, we devise a novel yet effective spatially-adaptive spectral modulation network tailored for pan-sharpening, termed SSMNet. The SSMNet can accurately restore the local characteristics specific to every spectral band of the HRMS image, contributing to the desirable fusion product. Specifically, we first customize a source-aware spectral modulator (SSM) for every band, which predicts a series of spatially-adaptive kernels, conditioning on the PAN guidance, to capture the local information of every band. Then, a cross-band information aggregation (CBIA) module consists of two successive functional operations, i.e., channel mixer and spatial mixer, which are used to facilitate the information communication among various bands, thus enabling the continuous spectral representation. Furthermore, an invertible neural network (INN)-based cross-stage feature integration (CSFI) module is utilized to integrate and refine the output features of every fundamental building block to fuse the final product. In addition, we introduce the histogram loss to constrain the distribution differences between every spectral band of the fused MS images and the corresponding

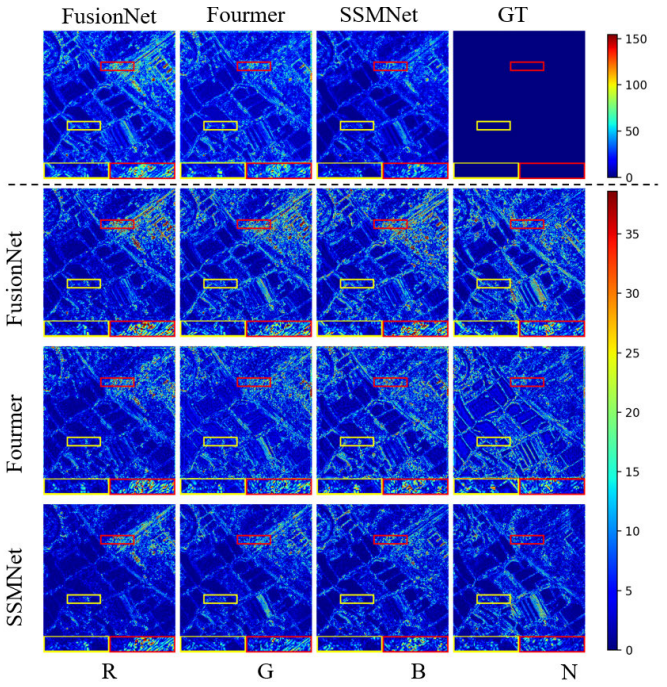

Fig. 2. Visualized absolute error maps corresponding to our motivation. Compared with other SOTA methods, our proposed SSMNet can achieve smaller errors in each band reconstruction (second to fourth row, each column corresponds to one band), thereby achieving lower overall errors (refer to the first row).

reference images. As illustrated in Fig. 1, comparing with another two representative approaches, each band generated by our SSMNet shares a similar pixel distribution pattern with that of reference band and has a smaller EMD value, indicating its excellent band-wise restoration. In addition, we also present the image-wise and band-wise absolute error maps in Fig. 2, from which our method presents the lower and sparser residual maps both in overall or single band than other approaches, indicating high congruence with the ground-truth (GT). From the above observations, our pan-sharpening framework, SSMNet, equipped with the above core designs is capable of accurately estimating every spectral band of MS image, thus obtaining the desirable fusion product.

Overall, our contributions can be summarized as follows.

1) We explore a new solution for pan-sharpening from the perspective of spectral-band modulation. Such design makes the framework consider both image-wise fusion effect and band-wise distribution differences.
2) We propose a novel SSM tailored for effectively learning the band-private characteristics. We also developed a CBIA module to facilitate communication among various band features, thus guaranteeing continuous spectral representation. Besides, an INN-based CSFI module is devised to obtain more informative features for fusing the final product.
3) To the best of our knowledge, this work is the first attempt to introduce the histogram loss into the DL-based pan-sharpening techniques.
4) Extensive experiments over multiple remote sensing datasets show that our SSMNet outperforms other SOTAs, and is well generalized to real-world full-resolution scenes.

The remaining article is organized as follows. In Section II, the related works and motivations are presented. In Section III-B, the proposed pan-sharpening framework will be detailed introduced. Experimental results and related discussions are presented exclusively by Section IV. Finally, the conclusion is drawn in Section V.

II. Related Works

A. Traditional Pan-Sharpening Methods

CS-, MRA-, and VO-based techniques constitute three main large families of the traditional pan-sharpening methods [3], [4], [11]. The CS methods commonly project MS into a new domain, and further substitute its spatial components with PAN. Representative CS approaches include the principal component analysis (PCA) [14], partial replacement adaptive CS (PRACS) [32], hue-intensity-saturation (HIS) [33], and Gram–Schmidt (GS) decomposition [34]. Many efforts have also been devoted to improving the above algorithms. For example, Ghahremani and Ghassemian [35] propose a linear IHS (NIHS) that combines the local and global synthesis strategies to estimate the intensity component, thereby significantly reducing the spectral distortion of IHS. Despite the higher computational efficiency, the CS methods are prone to produce artifacts. In contrast, the MRA approaches present desirable spectral preservation yet suffer from significant spatial distortion. The main principle of this category of the method is to inject the multiscale spatial details of PAN into MS. Smoothing filter-based intensity modulation (SFIM) [16], generalized Laplacian pyramid (GLP) [36], GLP with modulation transfer function matched filter (MTF-GLP) [37], additive wavelet luminance proportional (AWLP) [38], and extracted wavelet transform (DWT) [39] are commonly used MRA methods. Unlike the above two techniques, VO algorithms treat the pan-sharpening task as an ill-posed problem and struggle to minimize the loss function [18], [40], [41]. Methods of this category achieve better tradeoffs between spectral and spatial preservation, while the heavy computational burden hinders their real applications.

B. DL-Based Pan-Sharpening Methods

In recent years, DL techniques have gained popularity in both low-level and high-level vision tasks [22], [42], [43], [44], [45], [46], [47], and made tremendous breakthroughs in comparison to conventional methods. In the field of pan-sharpening, DL-based approaches have shown great superiority to traditional algorithms. PNN [20] is the first attempt, which consists of three convolutional layers, to use CNN for pan-sharpening learning. Though it's a very simple architecture, PNN still achieves favorable performance against conventional counterparts. Subsequently, Yang et al. [21] designed a deep pan-sharpening network based on the residual learning [22]. Since then, a variety of DL-based pan-sharpening models have emerged. Yuan et al. [23] designed an effective multiscale CNN architecture, Deng et al. [1] introduced the traditional fusion schemes into the network design, Jin et al. [48] first employ the adaptive convolution to pan-sharpening task, SFIIN [9] is the pioneering work to deal

with the pan-sharpening problem in both spatial and frequency domains, Zhou et al. [2] devised an efficient yet effective global modeling paradigm from the perspective of the Fourier domain. In addition, some unsupervised techniques [49], [50] which are mainly based on the spectral and spatial constraints have been developed to explore the potential solutions for pan-sharpening. Despite the revolutionary advances, existing DL-based methods can only learn the common features across various spectral bands while simply ignoring the band-specific local characteristics, since they usually focus on the fusion effect upon the entire image. Nevertheless, the abundant information of remote sensing data is dependent on its multiple spectral bands. In other words, an ideal HRMS image should contain both band-shared global structure and band-specific local characteristics.

C. Band Feature Enhancement

Recently, given that the different spectral bands of MS images contain different levels of spatial details, some efforts have been made to explore the intrinsic characteristics specific to each band. Lu et al. [13] developed a unified pan-sharpening model which emphasizes the spectral consistency between every fused band and the corresponding input one, while also studying the relationship between PAN images and every band of MS images in the gradient domain. Yang et al. [31] proposed a novel high-pass modification block to enhance the spatial information of each band in MS images. Zhou et al. [10] designed the so-called MS band-aware feature modulation module to investigate the modality-aware and band-aware characteristics. Despite these works demonstrating that considering the band-private information is beneficial to the model performance, they simply focus on the image-level or feature-level spatial enhancement of each band. Therefore, it is hard to effectively learn the band-private characteristics which are mainly distributed in the local regions. In addition, they rarely take into account the band-wise distribution differences between the fused images and the reference images.

III. Proposed Method

A. Problem Formulation

Pan-sharpening is the technique of super-resolving the low-spatial resolution MS image under the guidance of an HR PAN image. For a better explanation, let us first define some notations used throughout this article. We denote the PAN image and up-sampled MS image as $\mathcal{P} \in \mathbb{R}^{H\times W}$ and $\mathcal{M} \in \mathbb{R}^{H\times W\times c}$, respectively, where H and W are the height and width of image, while c represents the number of MS bands. We formulate the ith band of the up-sampled MS image as $\mathcal{B}_i \in \mathbb{R}^{H\times W}, i = 1, 2, \ldots, c$. Besides, the model output and the corresponding GT are, respectively, remarked as $\mathcal{SR} \in \mathbb{R}^{H\times W\times c}$ and $\mathcal{GT} \in \mathbb{R}^{H\times W\times c}$. The spatial resolution ratio between PAN and MS (denoted as $L \in \mathbb{R}^{h\times w\times c}$) is equal to 4, i.e., $H = 4\times h, W = 4\times w$.

B. Overall Framework

The commonly used strategy of existing pan-sharpening methods is to directly concatenate the MS (or MS features) and PAN (or PAN features) images to learn the cross-modality complementary information, thus producing the desired MS images. Despite the remarkable advances gained by this fusion paradigm, it usually focuses on the whole fusion effect while simply ignoring the band-private local characteristics, which hinders the further improvement of pan-sharpening performance and suffers from the limited generalization capability. Unlike prior methods, we attempt to accurately estimate the local characteristics of every spectral band thus reconstructing a desired HRMS image, since it is jointly determined by its various bands. To implement this target, in this article, we develop an SSMNet for pan-sharpening, which includes three core designs: SSM, CBIA module, and CSFI module. The first performs the multiscale feature extraction of every spectral band; the second is devised to facilitate the information communication among various bands to guarantee continuous spectral representations; the third integrates the cross-stage output feature to obtain the informative features for producing the final fusion image. In addition, we also redevelop the histogram loss for pan-sharpening learning, which narrows the distribution differences between every fused band and the corresponding reference one. Fig. 3 clearly presents the proposed pan-sharpening framework.

1) Structure Flow: As illustrated in Fig. 3, the input MS image L is first up-sampled to the PAN scale. Then, we split the up-sampled MS image $\mathcal{M}$ into c single bands along the spectral dimension. Next, we apply the simple convolution block to every spectral band $\mathcal{B}_i$ and the corresponding PAN image $\mathcal{P}$ to obtain their shallow feature representations. The obtained feature maps are jointly fed into the backbone network which is assembled by several fundamental building blocks, and then the cross-stage features are gradually refined to generate an ideal MS image $\mathcal{SR}$.

2) Supervision Flow: Orthogonal to model design, we develop effective loss functions to facilitate the network optimization in the training process, thus producing visually pleasing and numerically favorable pan-sharpening outcomes. As illustrated in Fig. 3, our loss functions consist of two components: image loss (the pixel loss, i.e., $\mathcal{L}_1$ loss) and histogram loss. Prior works usually focus on the fusion effect of the entire image, simply ignoring the band-wise differences. In fact, however, the band-wise distribution of the fused images should be consistent with the real images (i.e., $\mathcal{GT}$). To this end, we introduce the histogram loss to minimize the distribution differences between each spectral band of the fused $\mathcal{SR}$ image and the corresponding $\mathcal{GT}$ image. The well-designed loss functions enable our model to generate more desirable MS images since it is capable of considering both the entire fusion effect and band-specific local characteristics.

In the following, we will elaborate on the detailed structure of our model's core designs, including SSM, CBIA, and CSFI, and the newly-developed loss functions.

C. Core Building Designs

1) Source-Aware Spectral Modulator: From Fig. 3, each fundamental building block contains two core building designs. Fig. 4(a) shows the detailed structure of the first functional module, i.e., SSM, which includes multiple kernel

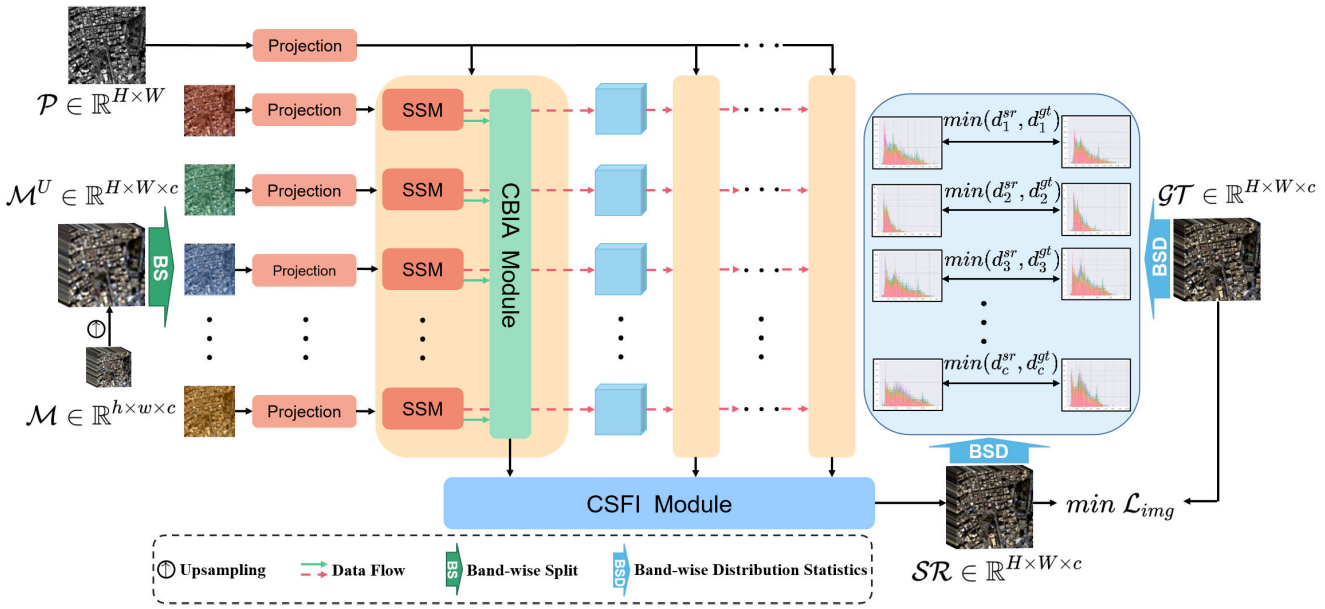

Fig. 3. Pipeline of the proposed pan-sharpening framework whose backbone consists of several fundamental building blocks. Each block includes two core designs: SSM and CBIA module. The output features of the backbone are progressively fused through the CSFI module. ⇢: Data forward propagation; →: Data flow toward CBIA module.

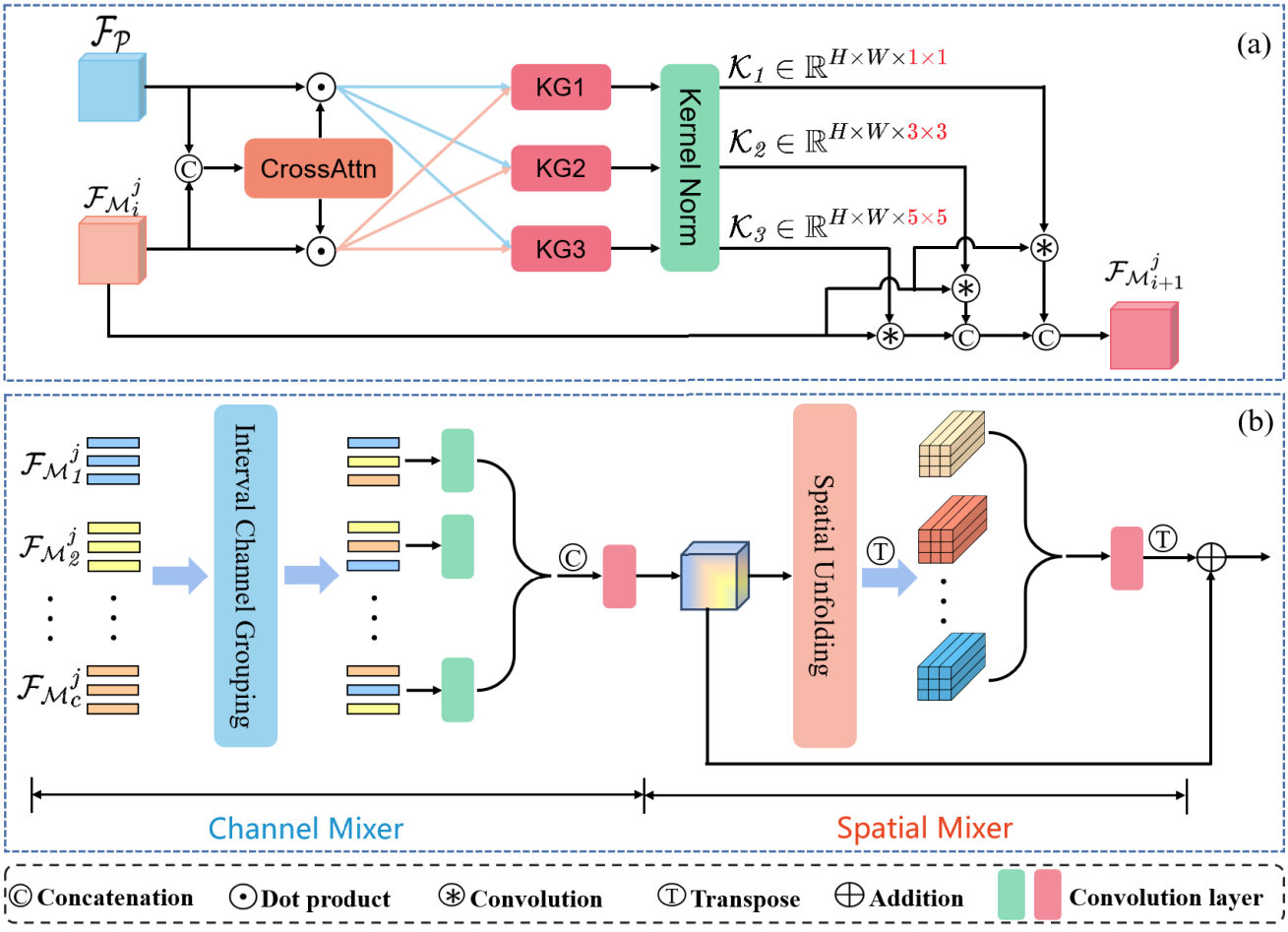

Fig. 4. Detailed structure of our proposed (a) SSM which includes multiple KGs, and (b) CBIA module which consists of two sequential operations, i.e., channel mixer and spatial mixer. →: Scaled PAN features; →: Scaled band features.

generators (KGs) to predict a series of spatially-adaptive convolution kernels, enabling it effectively capture the multiscale features of every spectral band. Note that, each SSM is independent yet they share the same structure. Since pan-sharpening is the process of spatially super-resolving LRMS image under the guidance of PAN image, cross-

modality information integration is beneficial for obtaining the informative representation. As shown in Fig. 4(a), we first apply a cross-attention module consisting of several convolution layers coupled with sigmoid activation function to obtain the spatial attention maps (A_P, A_M) of the input feature pair $(\mathcal{F}_P, \mathcal{F}_{\mathcal{M}_i}^{\ j})$ as

$$\mathcal{A}_{\mathcal{P}}, \mathcal{A}_{\mathcal{M}_i}^{\ j} = \text{CrossAttn}\left(\text{cat}\left(\mathcal{F}_{\mathcal{P}}, \mathcal{F}_{\mathcal{M}_i}^{\ j}\right)\right) \tag{1}$$

where cat$(\cdot)$ denotes the concatenation operation. $i \in \{1, 2, \ldots, c\}$ and $j \in \{1, 2, \ldots, n\}$, where c and n are the number of the spectral bands of MS image and fundamental building blocks, respectively. $\mathcal{F}_{\mathcal{M}_i}^{\ j}$ represents the feature maps of the ith band in the jth block. $\mathcal{F}_{\mathcal{P}}$ is the feature maps of the PAN image, which is shared across all bands and blocks. CrossAttn$(\cdot)$ is the mapping function of cross attention. Following feature information scaling by attention maps, we further adopt a group of KGs to predict the private adaptive kernels for every spectral band as:

$$\begin{aligned}
\mathcal{K}_i &= \text{Norm}\left(\text{KG1}\left(\mathcal{A}_{\mathcal{P}} \odot \mathcal{F}_P, \mathcal{A}_{\mathcal{M}_i}^{\ j} \odot \mathcal{F}_{\mathcal{M}_i}^{\ j}\right)\right) \\
\mathcal{K}_2 &= \text{Norm}\left(\text{KG2}\left(\mathcal{A}_{\mathcal{P}} \odot \mathcal{F}_P, \mathcal{A}_{\mathcal{M}_i}^{\ j} \odot \mathcal{F}_{\mathcal{M}_i}^{\ j}\right)\right) \\
\mathcal{K}_3 &= \text{Norm}\left(\text{KG3}\left(\mathcal{A}_{\mathcal{P}} \odot \mathcal{F}_P, \mathcal{A}_{\mathcal{M}_i}^{\ j} \odot \mathcal{F}_{\mathcal{M}_i}^{\ j}\right)\right)
\end{aligned} \tag{2}$$

where $\odot$ denotes the dot product. Norm$(\cdot)$ is the kernel normalization that is performed to scale the kernel elements for facilitating network optimization. $\mathcal{K}_1 \in \mathbb{R}^{H\times W\times 1\times 1}$, $\mathcal{K}_2 \in \mathbb{R}^{H\times W\times 3\times 3}$, and $\mathcal{K}_3 \in \mathbb{R}^{H\times W\times 5\times 5}$ are the current-stage predicted $k \times k$ multiscale kernels specific to every spectral. Notably, the designed KGs produce pixel-wise weights dependent on the input features, thus enabling a spatially-adaptive manner to capture the multiscale feature information corresponding to every spectral band. All KGs share a similar structure

$$\begin{aligned}
\mathcal{F}_{\text{Spa}} &= \text{SA}\left(\mathcal{F}_{\mathcal{P}}\right) \\
\mathcal{F}_{\text{Spe}} &= \text{CA}\left(\mathcal{F}_{\mathcal{M}_i}^{\ j}\right) \\
\text{KG}\left(\cdot\right) &= \mathcal{F}_{\text{Spa}} \odot \mathcal{F}_{\text{Spe}}
\end{aligned} \tag{3}$$

where SA$(\cdot)$ and CA$(\cdot)$ represent the spatial-wise and spectral-wise attention, respectively, both of them consist of several convolution layers. $\mathcal{F}_{\text{Spa}}$ and $\mathcal{F}_{\text{Spe}}$ are the corresponding outputs. Note that, without loss of generality, we have omitted the subscripts of KG$(\cdot)$.

By performing the predicted spatially-adaptive convolution kernels on the feature maps of every spectral band $\mathcal{F}_{\mathcal{M}_i}^{\ j}$, we can effectively learn its multiscale feature information as

$$\mathcal{F}'_{\mathcal{M}_i}^{\ j} = \text{cat}\left(\left(\mathcal{K}_1 \circledast \mathcal{F}_{\mathcal{M}_i}^{\ j}\right), \left(\mathcal{K}_2 \circledast \mathcal{F}_{\mathcal{M}_i}^{\ j}\right), \left(\mathcal{K}_3 \circledast \mathcal{F}_{\mathcal{M}_i}^{\ j}\right)\right) \tag{4}$$

where $\circledast$ denotes the convolution operation. $\mathcal{F}'_{\mathcal{M}_i}^{\ j}$ is the output of the ith SSM in the jth block.

2) Cross-Band Information Aggregation: After obtaining the informative band-specific features from every SSM, we further devise a CBIA module to enhance their channel and spatial interactions adequately, thereby enabling the continuous spectral representation. As shown in Fig. 4(b), the proposed CBIA module includes two successive operations: channel mixer and spatial mixer.

a) Channel mixer: To reduce the intergroup feature redundancy while fully exploiting the complementary information across various bands, we first perform the group shuffle [28] to divide the output features of every SSM module into m groups in the channel dimension as

$$\begin{aligned}
\left[g_1^1, g_1^2, \ldots, g_1^m\right] &= \text{GS}\left(\mathcal{F}'_{\mathcal{M}1}^{\ j}\right) \\
\left[g_2^1, g_2^2, \ldots, g_2^m\right] &= \text{GS}\left(\mathcal{F}'_{\mathcal{M}2}^{\ j}\right) \\
&\ldots \\
\left[g_c^1, g_c^2, \ldots, g_c^m\right] &= \text{GS}\left(\mathcal{F}'_{\mathcal{M}c}^{\ j}\right)
\end{aligned} \tag{5}$$

where GS$(\cdot)$ is the group shuffle operation. $g_i^l \in \mathbb{R}^{H\times W\times C_g}, i \in \{1, 2, \ldots, c\}, l \in \{1, 2, \ldots, m\}$ denotes the lth group of feature maps corresponding to the ith band, C_g is the channels of g_i^l. Next, we recombine these features to form new groups, which can achieve cross-band feature interaction, thus enhancing the diversity of spectral representation. In concrete terms, we can formulate this process as follows:

$$\begin{aligned}
\mathcal{F}''_{\mathcal{M}i}^{\ 1} &= \text{Re}\left(g_1^1, g_2^1, \ldots, g_c^1\right) \\
\mathcal{F}''_{\mathcal{M}i}^{\ 2} &= \text{Re}\left(g_1^2, g_2^2, \ldots, g_c^2\right) \\
&\ldots \\
\mathcal{F}''_{\mathcal{M}i}^{\ c} &= \text{Re}\left(g_1^m, g_2^m, \ldots, g_c^m\right)
\end{aligned} \tag{6}$$

where Re$(\cdot)$ denotes recombining the grouped cross-band features. Then, we integrate these shuffled features through convolution layers as

$$\mathcal{F}_{\text{CM}} = \text{Conv}\left(\text{cat}\left(\text{Conv}\left(\mathcal{F}''_{\mathcal{M}i}^{\ 1}, \mathcal{F}''_{\mathcal{M}i}^{\ 2}, \ldots, \mathcal{F}''_{\mathcal{M}i}^{\ c}\right)\right)\right) \tag{7}$$

where Conv$(\cdot)$ denotes the convolution layer, $\mathcal{F}_{\text{CM}}$ is the obtained informative features from the channel mixer.

b) Spatial mixer: Following the above process, we further perform the spatial unfolding operation which is expressed as follows:

$$[P_1, P_2, \ldots, P_N] = \text{SU}(\mathcal{F}_{\text{CM}}) \tag{8}$$

where SU$(\cdot)$ represents the spatial unfolding operation and has parameters stride and padding to control the overlap between the divided patches. $P_k \in \mathbb{R}^{p\times p\times C}$ $(k \in \{1, 2, \ldots, N\})$ represents the kth feature patch with size of p, while N is the number of patches. Next, we employ a convolution layer to obtain the spatially-shuffled features as

$$\mathcal{F}_{\text{SM}} = \text{Conv}(\text{Reshape}(P_1, P_2, \ldots, P_N)) \tag{9}$$

where Reshpae$(\cdot)$ transforms the patch dimension into the channel dimension such that the following convolution will operate on each divided patch and obtain residual information for global fusion. $\mathcal{F}_{\text{SM}}$ denotes the output feature of the spatial mixer.

By adding the $\mathcal{F}_{\text{CM}}$ and $\mathcal{F}_{\text{SM}}$, we can obtain the final output feature of the jth fundamental buliding block as

$$\mathcal{O}_j = \text{Add}(\mathcal{F}_{\text{CM}}, \mathcal{F}_{\text{SM}}) \tag{10}$$

where Add$(\cdot)$ is the element-wise addition.

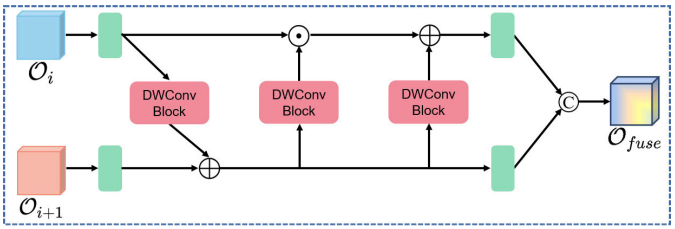

Fig. 5. Detailed flowchart of our proposed INN-based CSFI module.

3) Cross-Stage Feature Integration: Considering the lossless information of INN [51], [52], we apply an INN-based CSFI module to obtain more informative features for producing the desirable MS images. As illustrated in Fig. 5, inside the CSFI module, the affine coupling layers are implemented using depth-wise separable convolution blocks given the tradeoff between feature integration and computational consumption. Suppose that $\mathcal{O}_1$ and $\mathcal{O}_2$ are the output features corresponding to the first and second fundamental building blocks. In each INN layer, the detailed feature transformation is as follows:

$$\begin{aligned}\mathcal{I}_{\text{inn}}^{I} &= \text{DW}\left(\text{Conv}\left(\mathcal{O}_1\right)\right) + \text{Conv}\left(\mathcal{O}_2\right)\\ \mathcal{I}_{\text{inn}}^{II} &= \text{Conv}\left(\mathcal{O}_1\right) \odot \text{DW}\left(\mathcal{I}_{\text{inn}}^{I}\right) + \text{DW}\left(\mathcal{I}_{\text{inn}}^{I}\right)\\ \mathcal{I}_{\text{inn}}^{III} &= \text{cat}\left(\text{Conv}\left(\mathcal{I}_{\text{inn}}^{I}\right), \mathcal{I}_{\text{inn}}^{II}\right)\end{aligned} \tag{11}$$

where $\text{DW}(\cdot)$ denotes the depth-wise separable convolution block. Based on the above definition, we can formulate the whole process of CSFI as follows:

$$\begin{aligned}\mathcal{O}_{\text{fuse}}^{1} &= \text{INN}\left(\mathcal{O}_1, \mathcal{O}_2\right)\\ \mathcal{O}_{\text{fuse}}^{2} &= \text{INN}\left(\mathcal{O}_{\text{fuse}}^{1}, \mathcal{O}_3\right)\\ &\cdots\\ \mathcal{O}_{\text{fuse}}^{L} &= \text{INN}\left(\mathcal{O}_{\text{fuse}}^{n-1}, \mathcal{O}_n\right)\end{aligned} \tag{12}$$

where $\text{INN}(\cdot)$ denotes the function in (11). In the last, we perform a tail convolution layer on the refined features $\mathcal{O}_{\text{fuse}}^{L}$ to obtain the final fused image $\mathcal{SR}$.

By far, all the constituent modules of SSMNet have been elucidated comprehensively. To facilitate an enhanced comprehension, we also provide a formal representation of our model that is delineated as Algorithm 1.

Algorithm 1 SSMNet

Input: The number of blocks n, PAN image $\mathcal{P}$, up-sampled MS image $\mathcal{M}$ that contains c spectral bands $\mathcal{B}_i$

$\mathcal{F}_{\mathcal{M}_i^0} = Proj(\mathcal{B}_i)$, $\mathcal{F}_{\mathcal{P}} = Proj(\mathcal{P})$
for $j = 1, 2, 3, \ldots, n$ **do**
  for $i = 1, 2, 3, \ldots, c$ **do**
    $\mathcal{F}_{\mathcal{M}_i^j} = SSM(\mathcal{F}_{\mathcal{M}_i^{j-1}}, \mathcal{F}_{\mathcal{P}})$
  end
  $\mathcal{O}_j = CBIA(\{\mathcal{F}_{\mathcal{M}_i^j}, j = 1, 2, 3, \ldots, c\})$
  $\mathcal{O}_{fuse}^{j} = CSFI(\mathcal{O}_j, \mathcal{O}_{j-1})$
end
$\mathcal{SR} = Conv(\mathcal{O}_{fuse}^{n})$
return $\mathcal{SR}$

D. Training Optimization

As seen in Fig. 3, we adopt two loss terms, including the reconstruction loss $\mathcal{L}_r$ and the histogram loss $\mathcal{L}_h$, as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_r + \alpha\mathcal{L}_h \tag{13}$$

where α denotes the hyperparameter that is used to balance the spatial reconstruction details and overall distribution for each spectral channel. Specifically, we choose a widely used L_1 loss to serve as the reconstruction loss $\mathcal{L}_r$ to measure the image-wise differences between the fused $\mathcal{SR}$ images and $\mathcal{GT}$

$$\mathcal{L}_r = \left\|\mathcal{SR} - \mathcal{GT}\right\|_1. \tag{14}$$

The histogram loss $\mathcal{L}_h$ is derived from the EMD [53]. For simplicity, we calculate the Manhattan distance of the cumulative histograms between every fused spectral band and the corresponding reference band to obtain the EMD [54], which is defined as follows:

$$\mathcal{L}_h = \frac{1}{c} * \sum_{i=0}^{c-1} \left\|\mathcal{H}_i^{\mathcal{SR}} - \mathcal{H}_i^{\mathcal{GT}}\right\|_1 \tag{15}$$

where $\mathcal{H}_i^{\mathcal{SR}}$ and $\mathcal{H}_i^{\mathcal{GT}}$ are the estimated 1-D cumulative histogram vector of the ith fused band and the corresponding reference one, respectively.

Since the process of computing a histogram is nondifferentiable, we introduce the differentiable histogram with hard-binning trick [55] to enable end-to-end training. For convenience, we reuse the subscripts i and j again for representation. Suppose that μ_j and B_j represent the center value and bandwidth of the jth bin. By enumerating each pixel p in the image I, we approximate the hard-binned 1-D cumulative histogram vector using the following equations:

$$\begin{aligned}h_{i,j} &= \sum_{p \in I} \Psi\left(1.01^{B_j - |p - \mu_j|}, 1, 0\right)\\ \mathcal{H}_{i,j} &= \text{cdf}_j\left(h_i\right)\end{aligned} \tag{16}$$

where

$$\Psi(x, 1, 0) = \begin{cases} x, & \text{if } x > 1\\ 0, & \text{otherwise}\end{cases}$$

is a threshold function, and $h_{i,j}$ denotes the jth count value of the 1-D histogram vector corresponding to the ith band. By applying the cumulative density function (cdf) to the vector h_i, we can obtain each element $\mathcal{H}_{i,j}$ of the final outcome. Given that B_j $(0 < B_j < 1)$ is usually a tiny value, thus $\forall p \in (\mu_j - B_j, \mu_j + B_j)$, the output value of $\Psi(\cdot)$ will be close to 1, while 0 for others. Based on the above analysis, we can easily implement the differentiable histogram loss through CNN layers with the fixed parameters μ and B. In practice, to stabilize the training process, we will normalize the $h_{i,j}$ with the number of pixels $H \times W$.

IV. Experimental Results

In this section, we conduct a series of experiments on multiple satellite datasets to demonstrate the effectiveness

and generalization capability of our pan-sharpening framework. Specifically, we compare our model with several SOTA pan-sharpening approaches in both simulated and real-world scenes. Furthermore, we thoroughly discuss the proposed network architecture through ablation experiments.

A. Baseline Models

PNN [20], DiCNN [56], FusionNet [1], GPPNN [57], LAGNet [48], INNFormer [58], and Fourmer [2] are seven representative DL techniques that we select to demonstrate the superiority of our model. In addition, we also compare the proposed framework with four traditional algorithms: BT-H [59], BDSD-PC [60], MTF-GLP-HPM-P [17], and LRTCFPan [19].

B. Experiment Settings

1) Datasets: In this section, we validate the effectiveness of our SSMNet over multiple remote sensing datasets, including WorldView-3 (WV3), GaoFen-2 (GF2), and WorldView-2 (WV2). Specifically, both the WV3 and WV2 datasets contain eight-band (coastal, blue, green, yellow, red, red edge, NIR 1, and NIR 2) MS images and single-channel PAN images, while the GF2 dataset includes four-band (red, green, blue, and NIR) MS images and the corresponding PAN images, respectively. Notably, each MS and PAN image pairs are collected from the same scene. Due to the absence of GT images, we simulate the training set by leveraging Wald's protocol [61]. As a result, each training set contains thousands of PAN/LRMS/GT (i.e., the original MS images) images pairs with the sizes of $64 \times 64 \times 1$, $64 \times 64 \times 8$, and $64 \times 64 \times 8$, respectively. In addition, all training and testing data used in this work can be available on the public website.[1]

2) Metrics: We follow the research standards of the pan-sharpening community, and select the spectral angle mapper (SAM) [62], the relative dimensionless global error in synthesis (ERGAS) [63], the spatial correlation coefficient (SCC) [64], and the $Q2^n$ (Q8 for eight-band datasets while Q4 for four-band datasets) [65] as the reduced-resolution image quality assessment (IQA) metrics. For the full-resolution evaluation, we adopt three no-reference indicators, including the hybrid quality with no reference (HQNR) index [66], the spectral distortion D_λ index and the spatial distortion D_s index [67]. Specifically, the detailed mathematical definition of these metrics is presented as follows.

a) Spectral angle mapper: The SAM metric whose ideal value is 0 is widely used to measure spectral distortions of fused images compared with GT. The mathematical expression of SAM is defined as

$$\mathrm{SAM} = \frac{1}{C}\sum_{i=1}^{C} \arccos\left(\frac{x_i \cdot \hat{x}_i}{\|x_i\|_2 \|\hat{x}_i\|_2}\right) \tag{17}$$

where C represents the number of spectral bands, x_i and $\hat{x}_i$ are the ith spectral vector of GT and pan-sharpened image. $\|\cdot\|_2$ means the L2 norm.

[1] https://liangjiandeng.github.io/PanCollection.html

b) Relative dimensionless global error in synthesis: The ERGAS, a global index, is commonly used to assess the overall distortions of fused images. The optimal value for ERGAS is equal to 0, mathematically, it can be expressed as

$$\mathrm{ERGAS}(x, \hat{x}) = 100 \times r \sqrt{\frac{1}{C}\sum_{i=1}^{C} \frac{\mathrm{RMSE}(x_i, \hat{x}_i)}{\mu_{x_i}}} \tag{18}$$

where r is the spatial resolution ratio between PAN and MS images, while x and $\hat{x}$ denote GT and fused image, respectively. $\mathrm{RMSE}(x_i, \hat{x}_i)$ represents the root mean square error between the ith band of GT and fused image. μ_{x_i} is the mean value of the ith band of GT.

c) Spatial correlation coefficient: The SCC index with an ideal value of 1 characterizes the similarity between the spatial details of GT and fused images. The specific calculation of SCC includes two steps: 1) using a high-pass filter to extract the high frequencies of images and 2) calculating the correlation coefficient (CC) between the high frequencies to obtain the SCC. The commonly used Laplacian filter has the following form:

$$F = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}. \tag{19}$$

The CC is another widely used spectral indicator which is defined as follows:

$$\mathrm{CC} = \frac{\sum_{i=1}^{w}\sum_{j=1}^{h}\left(x_{i,j} - \mu_x\right)\left(\hat{x}_{i,j} - \mu_{\hat{x}}\right)}{\sqrt{\sum_{i=1}^{w}\sum_{j=1}^{h}\left(x_{i,j} - \mu_x\right)^2\left(\hat{x}_{i,j} - \mu_{\hat{x}}\right)^2}} \tag{20}$$

where w and h are the width and height of the image, while x and $\hat{x}$ denote the GT and fused image, respectively. μ_* denotes the mean value of the image.

d) Quality index ($Q2^n$): The $Q2^n$ is extended from the universal image quality index (UIQI). Analogously to UIQI, the $Q2^n$ is defined as the product of three terms including CC, contrast distortion, and mean bias, the specific expression is as follows:

$$Q2^n = \frac{|\sigma_{x\hat{x}}|}{\sigma_x \cdot \sigma_{\hat{x}}} \cdot \frac{2\sigma_x \cdot \sigma_{\hat{x}}}{\sigma_x^2 + \sigma_{\hat{x}}^2} \cdot \frac{2|\bar{x}| \cdot |\bar{\hat{x}}|}{|\bar{x}|^2 \cdot |\bar{\hat{x}}|^2} \tag{21}$$

where $x = x(i, j)$ and $\hat{x} = \hat{x}(i, j)$ are two hypercomplex numbers that characterize the GT and fused image at pixel (i,j). $\sigma_{x\hat{x}}$ is the covariance between x and $\hat{x}$. σ_*^2 and $\bar{*}$ denote the variance and mean. The ideally fused image has a $Q2^n$ of 1. Specifically, for the eight-band data (e.g., WorldView-3), the $Q2^n$ is Q8, while Q4 corresponds to the four-band data (e.g., GaoFen-2 dataset).

e) Spectral distortion index (D_λ): D_λ is a representative spectral metric denoting the difference of interband Q values, which is calculated from the fused MS bands and LR MS bands. Specifically, it is defined as follows:

$$D_\lambda = \sqrt[q]{\frac{1}{C(C-1)}\sum_{i=1}^{C}\sum_{j=1(j\neq i)}^{C} |d_{i,j}(\hat{x}, y)|^q} \tag{22}$$

where $\hat{x}$ and y denote the fused MS image and LR MS image, respectively. C is the number of spectral bands in the MS image, and $d_{i,j} = Q(\hat{x}_i, \hat{x}_j) - Q(y_i, y_j)$.

TABLE I
QUANTITATIVE COMPARISON BETWEEN OUR MODEL AND OTHER SOTA METHODS OVER THE REDUCED-RESOLUTION SAMPLES FROM TWO BENCHMARK DATASETS (BOLD: BEST; UNDERLINE: SECOND BEST)

Method	World-View3				GaoFen2			
	SAM	ERGAS	Q8	SCC	SAM	ERGAS	Q4	SCC
BT-H	4.9198	4.5789	0.8324	0.9252	1.6488	1.5280	0.9177	0.9570
BDSD-PC	5.4293	4.6976	0.8294	0.9081	1.6813	1.6667	0.8922	0.9521
MTF-GLP-HPM-R	5.3383	5.2301	0.8346	0.8892	1.6502	1.5875	0.8996	0.9480
LRTCFPan	4.7367	4.3153	0.8463	0.9273	1.2977	1.2723	0.9350	0.9637
PNN	3.6798	2.6819	0.8929	0.9761	1.0477	1.0572	0.9604	0.9772
DiCNN	3.5929	2.6733	0.9004	0.9763	1.0525	1.0812	0.9594	0.9771
FusionNet	3.3252	2.4666	0.9044	0.9807	0.9735	0.9878	0.9641	0.9806
GPPNN	3.0554	2.3056	0.9143	0.9837	0.7972	0.7107	0.9791	0.9897
LAGNet	3.1042	2.2999	0.9098	0.9838	0.7859	0.6869	0.9804	0.9906
INNFormer	3.2174	2.3604	0.9117	0.9831	0.7875	**0.6619**	0.9801	0.9919
Fourmer	3.2363	2.4189	0.9108	0.9840	0.9757	0.8845	0.9698	0.9870
SSMNet (ours)	**2.9002**	**2.1504**	**0.9193**	**0.9864**	**0.7344**	0.6815	**0.9812**	**0.9926**
Ideal value	**0**	**0**	**1**	**1**	**0**	**0**	**1**	**1**

f) Spatial distortion index (Q_s): D_s measures the spatial distortion complementary to D_λ. It is defined as

$$D_s = \sqrt[q]{\frac{1}{C}\sum_{i=1}^{C} |Q(\hat{x}_i, p) - Q(y_i, \tilde{p})|^q} \tag{23}$$

where p and $\tilde{p}$ and represent PAN and its degraded LR version. q is usually set to 1.

g) Hybrid quality with no reference: The HQNR index measuring the global quality of fused image without GT is the combination of the above two distortions. It is given as

$$\text{HQNR} = \left(1 - D_\lambda^K\right)^\alpha (1 - D_s)^\beta \tag{24}$$

where usually $\alpha = \beta = 1$.

3) Training Details: We implement our SSMNet in PyTorch 2.0 and Python 3.10 using a Linux operating system with an NVIDIA RTX4090 GPU. We train our network for 500 epochs with a batch size of 32. During the training stage, we adopt Adam optimizer with $\beta_1 = 0.9$, $\beta_1 = 0.999$ to update the network parameters using a dynamic learning rate which is set to 5×10^{-4} for the initial 250 epochs and becomes 0.1 times the original for the next 250 epochs.

C. Comparison With SOTA Methods

1) Evaluation on Reduced-Resolution Scene: We first perform the reduced-resolution assessment to measure the difference between the predicted HRMS images and GT images. The quantitative results for all datasets have been collected in Table I, where the best values are highlighted in bold. From Table I, it is clearly shown that our model yields favorable results on all evaluation indexes against other comparison techniques demonstrating its superiority. The SAM and SCC indicators are widely applied to characterize the spectral distortions and spatial similarity of pan-sharpened images compared to the reference images. It is apparent that our model is significantly lower than other approaches in SAM, yet gains higher SCC values over all datasets, demonstrating its desirable spectral and spatial preservation. The best quantitative outcomes of all metrics over multiple datasets favorably prove that our model is capable of recovering precise spatial details while preserving the ideal spectral information. This can be justified because, compared with other DL-based technologies which simply focus on the image-wise fusion effect, our model reconstructs the texture-rich MS images by accurately estimating every spectral band thus better balancing the spectral and spatial qualities. In other words, it is necessary to take into account the band-private characteristics since an ideal HRMS image is collectively determined by its every spectral band.

Our model presents excellent fusion ability. The visual comparison of our model and other cutting-edge DL-based techniques on a representative WV3 example have been depicted in Fig. 6. Note that we have omitted the results of traditional algorithms due to their inferior performance. The first two columns present the RGB visualization of all compared methods, where our SSMNet presents minimal spatial and spectral aberrations in comparison to other competing approaches. In addition to exhibiting the visual results for each method, we also calculate the absolute error maps between the pan-sharpened images and the corresponding reference image. As shown in the final column, our method presents lower and sparser residual maps than other approaches, indicating high congruence with the GT image. This is expected because our model is capable of accurately estimating every band by learning the band-private characteristics, contributing to high-fidelity pan-sharpening. In a word, our method embraces superior fusion capability in comparison to other pan-sharpening methods as evidenced by the quantitative and qualitative comparisons.

We further demonstrate our model's superiority by comparing the visual results on a typical GF2 sample. As illustrated

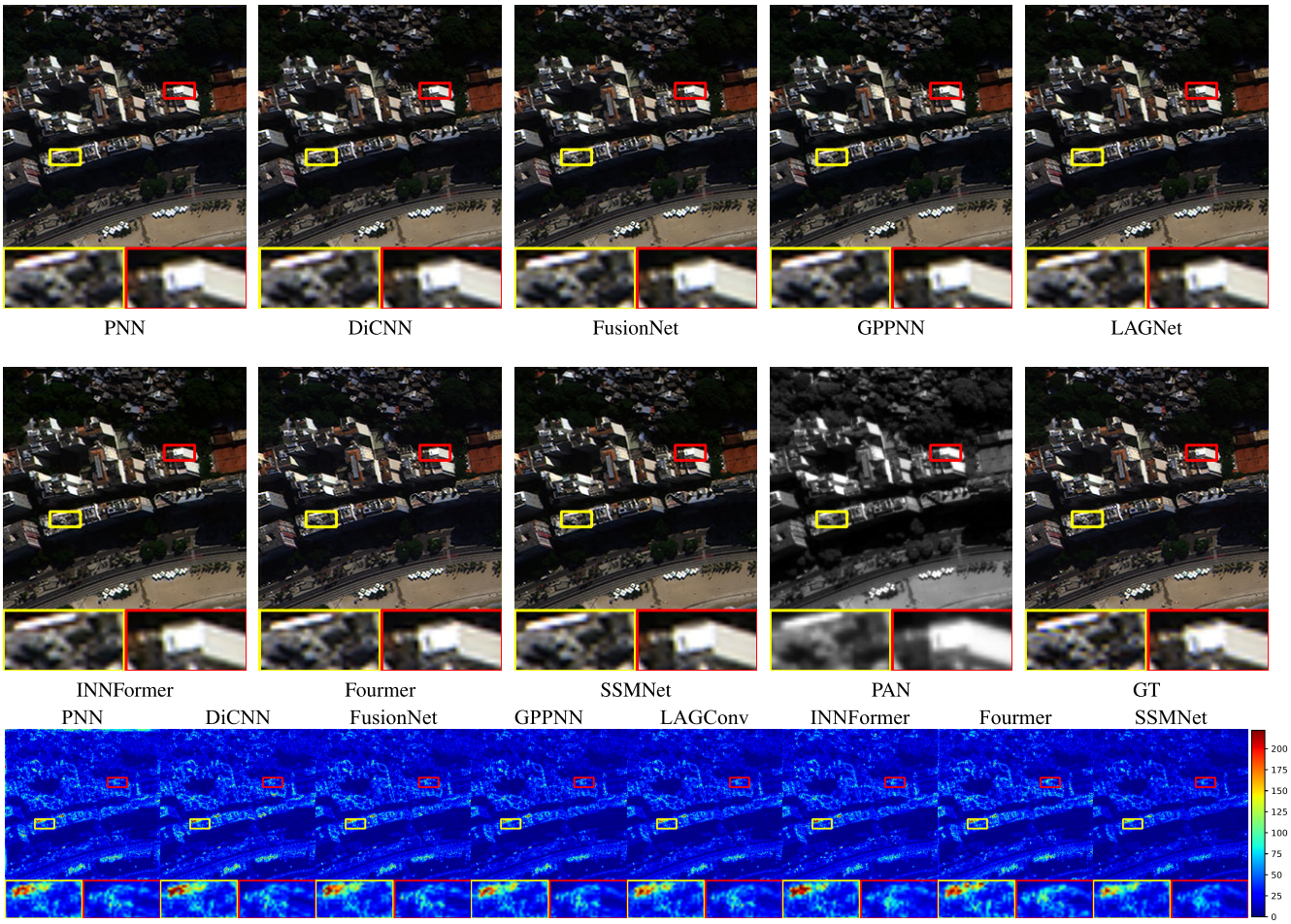

Fig. 6. Qualitative comparison between our model and other SOTA techniques on a typical satellite image from the WorldView-3 dataset. Images in the last row visualize the absolute error maps (mean value of all bands) between the pan-sharpened results and the GT (please zoomed-in view to see more details).

in Fig. 7, our technique preserves more spatial details yet enables desirable spectral fidelity, thereby presenting minimal aberration. By contrast, other comparison approaches demonstrate obvious spatial distortions or inferior spectral quality. For example, the sharpened outcomes of GPPNN and Fourmer exhibit distorted structures, while the results of LAGNet and INNFormer present desired spatial textures but suffer from spectral aberration. Based on the above quantitative and qualitative comparisons, our method renders the numerically favorable and visually pleasing fusion outcomes compared with other approaches, since it kindly takes care of the band-private characteristics.

2) Evaluation on Full-Resolution Scene: We further implement experiments on some full-resolution examples to evaluate the performance of our model in real-world scenes and its generalizability. Specifically, we directly apply the pre-trained model obtained from the reduced-resolution data to some full-resolution samples. Note that we adopt three widely used no-reference evaluation metrics to assess the full-resolution performance due to the unavailability of GT images. The quantitative outcomes for all compared methods over WV3 and GF2 datasets have been compiled in Table II. As presented in Table II, the DL-based techniques still perform favorably against their traditional counterparts. In addition, our proposed framework achieves almost the optimal outcomes for all indexes, confirming its superior generalization capability compared to both traditional and DL-based SOTA pan-sharpening technologies. The visual comparison of all compared DL-based methods on a representative full-resolution test case has been illustrated in Fig. 8. It is easy to observe from Fig. 8, our proposed pan-sharpening framework presents better spatial textures yet shows pleasing spectral preservation. Other competing approaches, by contrast, either contain blurry spatial textures and distorted edges or exhibit significant spectral aberration. In the enlarged area marked by a red rectangle, for example, the roof top generated by FusionNet, INNFormer, and Fourmer displays obvious spatial distortions. From the amplified region in the bottom left corner, moreover, the inferior spectral fidelity can be clearly observed in the edges of buildings produced by INNFormer and Fourmer. Both the quantitative results and qualitative analysis demonstrate that our model can be well applied to real-world scenes.

3) Comparison of Band-Wise Restoration: To better present the pan-sharpened quality per band, we further employ a global indicator, i.e., ERGAS, to measure the overall distortions of fused bands from our method and two cutting-edge DL techniques, e.g., FusionNet and Fourmer. Fig. 9 presents

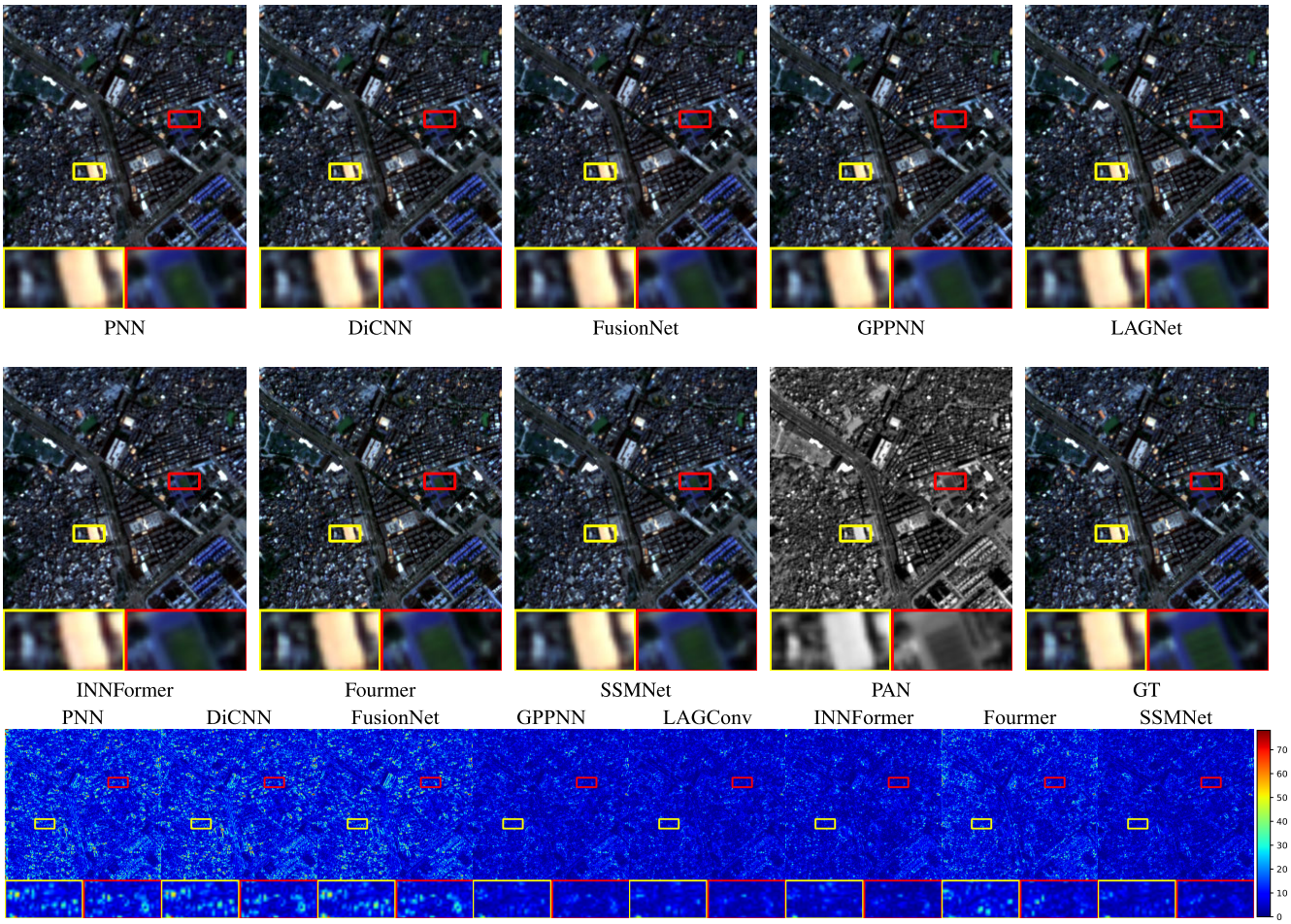

Fig. 7. Qualitative comparison between our model and other SOTA techniques on a typical satellite image from the GaoFen-2 dataset. Images in the last row visualize the absolute error maps (mean value of all bands) between the pan-sharpened results and the GT (please zoomed-in view to see more details).

TABLE II

QUANTITATIVE COMPARISON BETWEEN OUR MODEL AND OTHER SOTA METHODS OVER THE FULL-RESOLUTION SAMPLES FROM TWO BENCHMARK DATASETS (BOLD: BEST; UNDERLINE: SECOND BEST)

Method	World-View3			GaoFen2		
	D_λ	D_s	HQNR	D_λ	D_s	HQNR
BT-H	0.0574	0.0810	0.8670	0.0602	0.1313	0.8165
BDSD-PC	0.0625	0.0730	0.8698	0.0759	0.1548	0.7812
MTF-GLP-HPM-R	<u>0.0206</u>	0.0630	0.9180	0.0336	0.1404	0.8309
LRTCFPan	**0.0176**	0.0528	0.9307	0.0325	0.0896	0.8806
PNN	0.0213	0.0428	0.9369	0.0317	0.0943	0.8771
DiCNN	0.0362	0.0462	0.9195	0.0369	0.0992	0.8675
FusionNet	0.0239	0.0364	0.9406	0.0350	0.1013	0.8673
GPPNN	0.0281	0.0378	0.9352	0.0229	0.0665	0.9122
LAGNet	0.0368	0.0418	0.9230	<u>0.0284</u>	0.0792	0.8947
INNFormer	0.0550	0.0679	0.8815	0.0609	0.1096	0.8360
Fourmer	0.0224	<u>0.0347</u>	<u>0.9437</u>	0.0470	**0.0380**	<u>0.9166</u>
SSMNet (ours)	0.0209	**0.0323**	**0.9475**	**0.0223**	<u>0.0579</u>	**0.9211**
Ideal value	**0**	**0**	**1**	**0**	**0**	**1**

the ERGAS of each fused band on the GaoFen-2 testing dataset that contains four bands including red, green, blue, and NIR. It is clearly seen that every fused band from our method has a smaller ERGAS in comparison to that of another two approaches, demonstrating its improved band-wise restoration.

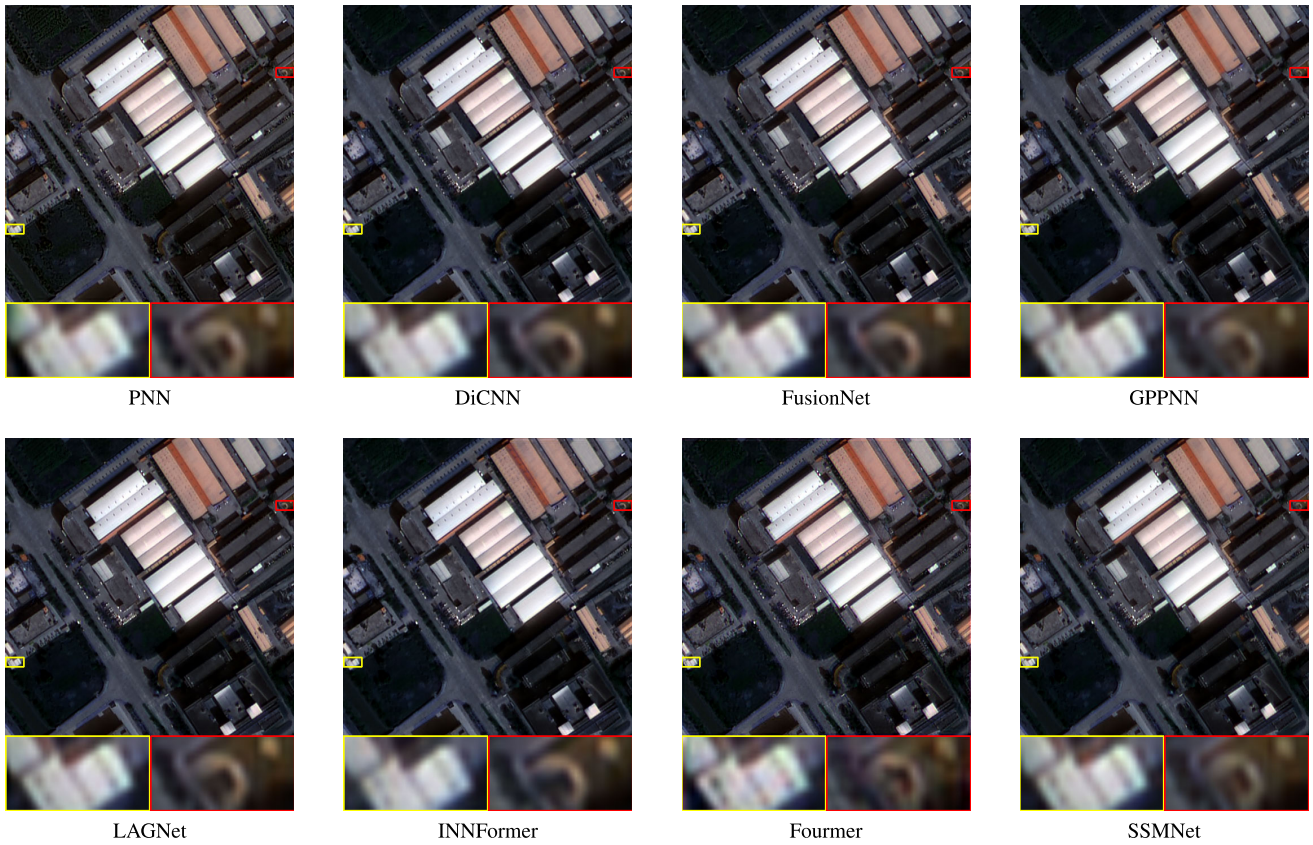

Fig. 8. Qualitative comparison between our model and other SOTA techniques on a typical satellite image from the GaoFen-2 dataset. Notably, the MSE maps are unavailable due to the lack of GT MS images in real-world full-resolution scenes (please zoomed-in view to see more details).

TABLE III

QUANTITATIVE COMPARISON BETWEEN OUR MODEL AND OTHER SOTA METHODS OVER THE REDUCED-RESOLUTION AND FULL-RESOLUTION SAMPLES FROM WORLDVIEW-2 DATASET (BOLD: BEST; UNDERLINE: SECOND BEST)

Method	Reduced Resolution				Full Resolution		
	SAM	ERGAS	Q8	SCC	D_λ	D_s	HQNR
PNN	7.1158	5.6152	0.7619	0.8782	0.1488	0.0771	0.7866
DiCNN	6.9216	6.2507	0.7205	0.8552	0.1416	0.1023	0.7697
FusionNet	6.4257	5.1363	0.7961	0.8746	<u>0.0523</u>	0.0559	<u>0.8944</u>
GPPNN	6.8295	5.1743	0.7942	0.9086	0.1194	0.0550	0.8323
LAGNet	6.9545	5.3262	0.8054	<u>0.9125</u>	0.1308	0.0547	0.8223
INNFormer	6.2793	**4.6189**	<u>0.8237</u>	**0.9170**	0.1403	<u>0.0376</u>	0.8276
Fourmer	<u>6.2241</u>	4.9145	0.8214	0.8980	0.0545	0.0723	0.8769
SSMNet (ours)	**5.9818**	<u>4.7329</u>	**0.8261**	0.9068	**0.0494**	**0.0316**	**0.9207**
Ideal value	**0**	**0**	**1**	**1**	**0**	**0**	**1**

D. Generalization Capability on Cross-Sensor Data

We further investigate the generalization capability and adaptability of our proposed framework and other DL-based techniques on the cross-sensor dataset. Specifically, we use some samples from the WV2 sensor to test all DL-based models that are trained on the WV3 dataset without any fine-tuning. This is justified because both WV3 and WV2 data share the same spectral band but their spatial resolution is slightly different. Table III has summarized the quantitative results of all compared approaches on the reduced- and full-resolution scenes. It is clear that our proposed framework gains almost the best results for all indexes again, indicating its pleasing applicability and robustness.

E. Ablation Experiments

We further investigate the contribution of different core ingredients by performing ablation studies on the GF2 dataset.

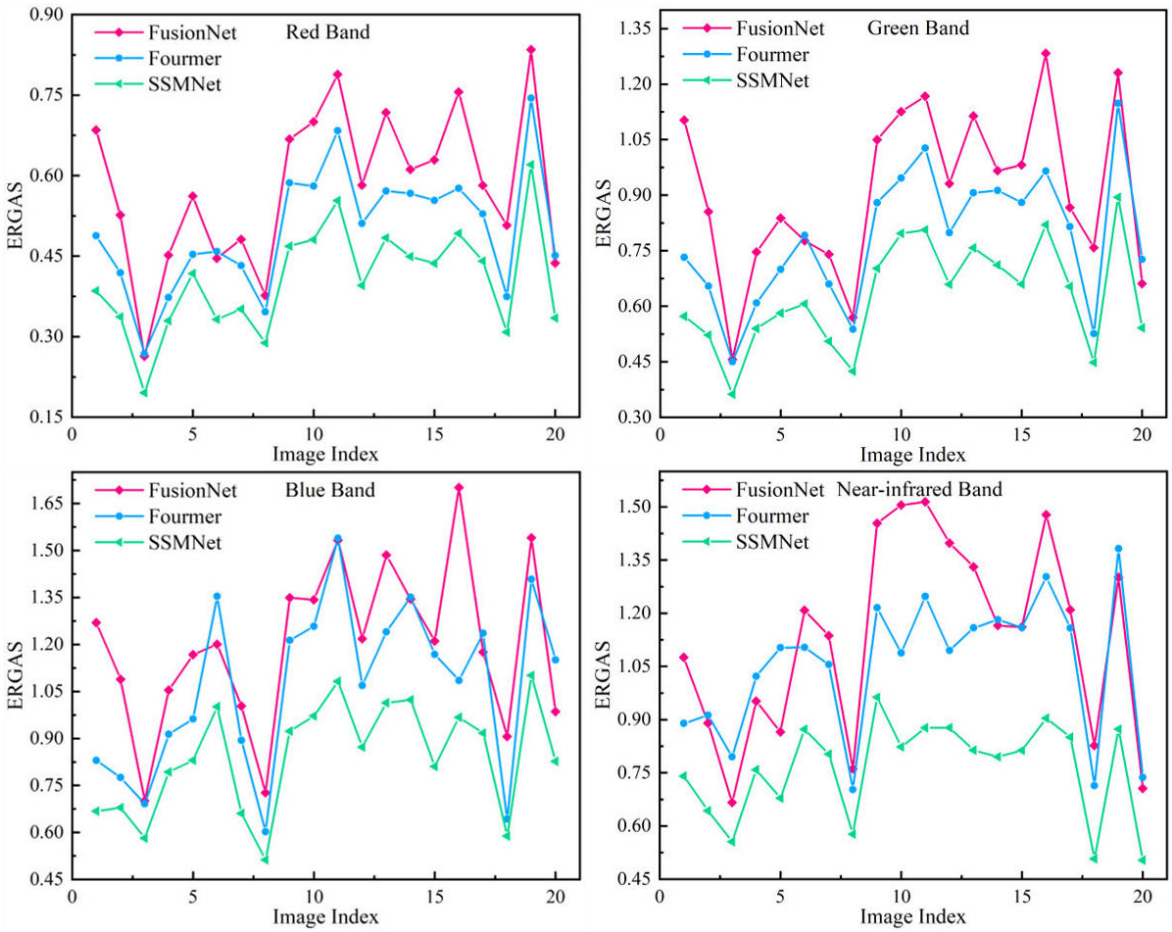

Fig. 9. Comparison of band-wise ERGAS between our method and two representative DL techniques over the GaoFen-2 dataset.

TABLE IV
QUANTITATIVE RESULTS OF ABLATION EXPERIMENTS OVER THE REDUCED-RESOLUTION SAMPLES FROM GAOFEN2 DATASET, WHERE HL DENOTES THE HISTOGRAM LOSS

Config.	SSM	CBIA	CSFI	HL	SAM↓	ERGAS↓	Q4↑	SCC↑
(I)	✓	✓	✓	✓	**0.7344**	**0.6815**	**0.9812**	**0.9926**
(II)	✓	✓	✓	✗	0.7734	0.6918	0.9799	0.9903
(III)	✓	✓	✗	✗	0.7885	0.6969	0.9799	0.9902
(IV)	✓	✗	✗	✗	0.7975	0.7243	0.9790	0.9892
(V)	✗	✗	✗	✗	0.8605	0.7951	0.9747	0.9875

As illustrated in Table IV, we explore four variants of our proposed SSMNet (i.e., Config. I).

1) Effect of Histogram Loss Functions: To the best of our knowledge, we are the first to introduce histogram loss into DL-based pan-sharpening techniques. To verify its effectiveness, we make a comparison between Config. I and Config. II (by removing the histogram loss from Config. I). As compiled in Table IV, by incorporating the histogram loss into the network optimization, our framework achieves better results, proving its effectiveness and applicability.

2) Effect of CSFI Structure: The CSFI is a postfusion module that is used for effectively refining and integrating the cross-stage output features to fuse the final product. We take an INN-based architecture to implement this module. In our design, we ameliorate the affine coupling layers with the depth-wise separable convolution blocks for balancing the feature extraction and computational consumption. To make a fair comparison and demonstrate the effectiveness, we replace the CSFI module with a vanilla convolution module with similar parameters size, and directly take the concatenated

TABLE V
QUANTITATIVE RESULTS OF ABLATION EXPERIMENTS OVER THE REDUCED-RESOLUTION SAMPLES FROM GAOFEN2 DATASET. $X - L_1$ AND $X - L_2$ DENOTE CALCULATING THE CORRESPONDING LOSS TERM THROUGH L_1 AND L_2 DISTANCE, RESPECTIVELY

Config.	$HL - L_1$	$HL - L_2$	$GC - L_1$	$GC - L_2$	SAM↓	ERGAS↓	Q4↑	SCC↑
(I)	✓	✗	✗	✗	0.7344	0.6815	0.9812	**0.9926**
(II)	✗	✓	✗	✗	**0.7298**	**0.6686**	**0.9816**	0.9912
(III)	✗	✗	✓	✗	0.7389	0.6835	0.9799	0.9908
(IV)	✗	✗	✗	✓	0.7351	0.6823	0.9806	0.9909

cross-stage output as input (corresponding to the Config. III). The quantitative results in Table IV have demonstrated the effectiveness of the CSFI module (Config. II) in comparison to the vanilla concatenation and convolution module in Config. III. The performance difference indicates that using our CSFI benefits the model performance gains, suggesting its better ability to integrate and fuse the cross-stage information.

3) Effect of CBIA Structure: Due to its meticulously crafted channel and spatial mixer, the CBIA module is employed to facilitate the cross-band information communication and the exchange of complementary features, thereby guaranteeing a more comprehensive spectral representation. We further investigate the effect of CBIA architecture. Specifically, we replace the channel mixer and spatial mixer in CBIA with several depth-wise separable convolution layers (Config. IV). In Table IV, the CBIA (Config. III) obtains better outcomes which can be attributed to its more effective feature interaction, indicating the rationality of our module design.

4) Effect of SSM Structure: Each individual band of the MS image has its own SSM module, enabling the SSMNet to extract the band-private feature in the early-stage. The SSM module mainly utilizes a series of predicted spatially-adaptive convolution kernels with different kernel sizes, enabling it to effectively capture the multiscale features of every spectral band. To validate the efficacy of our design, we substitute these adaptive convolution kernels with vanilla 2-D convolution and process all bands (Config. V) at one time. Table IV gives the corresponding quantitative results. It is evidently that our proposed multiscale spatially-adaptive kernels yield better results for all metrics, demonstrating its favorable representation learning ability.

5) Further Exploration of Histogram Loss: We further explore other variants and alternatives of histogram loss. To be specific, we first compare the impact of EMD calculation methods on performance. Table V clearly shows that our model achieves improved performance when replacing the L_1 distance with L_2, indicating the promising potential of histogram loss. As many have recognized, high-order statistics are commonly used to enforce the learning of texture features. However, most high-order statistics are nondifferentiable. In light of this, we use the gradient statistic constraint as an approximation of high-order statistics to explore its efficacy. Here, we still employ the L_1 and L_2 distance, respectively, to calculate the gradient difference between the sharpened band and the reference one. Likewise, the gradient statistic constraint with L_2 distance [Config. V(IV)] obtains better performance compared to L_1 [Config. V(III)]. Overall, our model equipped with histogram loss gains better outcomes, moreover, promising performance gains are available when altering the calculation of histogram loss.

F. Limitation and Further Discussion

First, we evaluate the superiority of the proposed framework on pan-sharpening, and we will extend it to other image fusion tasks, e.g., hyperspectral image super-resolution (HISR). Second, though we are the first to introduce the histogram loss in pan-sharpening tasks to the best of our knowledge, it is still worth exploring its effectiveness in other models and analogous tasks. Third, we will attempt to rethink the histogram loss from other aspects, such as gradient-domain and Fourier space.

V. CONCLUSION

In this work, we develop a novel yet effective framework for pan-sharpening since existing SOTAs simply ignore the band-private local characteristics that differ greatly from each other. First, the proposed SSM predicts a series of spatially-adaptive kernels under the HR guidance of PAN image to learn the multiscale local information of every spectral band. Then, the CBIA module is performed to facilitate the CBIA to guarantee the continuous spectral representations. Furthermore, the INN-based CSFI module is configured to progressively integrate the cross-stage outputs to obtain more informative features for generating the desired MS images. In addition, we first introduce a histogram loss term to minimize the distribution differences between the fused spectral bands and the corresponding reference ones. Benefiting from the above core designs, our solution is capable of obtaining numerically favorable while visually pleasing fusion results.

REFERENCES

[1] 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.

[2] M. Zhou, J. Huang, C.-L. Guo, and C. Li, "Fourmer: An efficient global modeling paradigm for image restoration," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 42589–42601.

[3] X. Meng, H. Shen, H. Li, L. Zhang, and R. Fu, "Review of the pansharpening methods for remote sensing images based on the idea of meta-analysis: Practical discussion and challenges," *Inf. Fusion*, vol. 46, pp. 102–113, Mar. 2019.

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

[5] I. R. Farah, W. Boulila, K. S. Ettabaa, and M. B. Ahmed, "Multiapproach system based on fusion of multispectral images for land-cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 12, pp. 4153–4161, Dec. 2008.

[6] T. Berhane et al., "Decision-tree, rule-based, and random forest classification of high-resolution multispectral imagery for wetland mapping and inventory," *Remote Sens.*, vol. 10, no. 4, p. 580, Apr. 2018.

[7] H. Zhang, M. Gong, P. Zhang, L. Su, and J. Shi, "Feature-level change detection using deep representation and feature change analysis for multispectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 11, pp. 1666–1670, Nov. 2016.

[8] L. Mou, L. Bruzzone, and X. X. Zhu, "Learning spectral–spatial–temporal features via a recurrent convolutional neural network for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 924–935, Feb. 2019.

[9] M. Zhou et al., "Spatial-frequency domain information integration for pan-sharpening," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 274–291.

[10] M. Zhou, K. Yan, X. Fu, A. Liu, and C. Xie, "PAN-guided band-aware multi-spectral feature enhancement for pan-sharpening," *IEEE Trans. Comput. Imag.*, vol. 9, pp. 238–249, 2023.

[11] 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.

[12] C. Ballester, V. Caselles, L. Igual, J. Verdera, and B. Rougé, "A variational model for P+XS image fusion," *Int. J. Comput. Vis.*, vol. 69, no. 1, pp. 43–58, Aug. 2006.

[13] H. Lu, Y. Yang, S. Huang, W. Tu, and W. Wan, "A unified pansharpening model based on band-adaptive gradient and detail correction," *IEEE Trans. Image Process.*, vol. 31, pp. 918–933, 2022.

[14] P. S. Chavez and A. Y. Kwarteng, "Extracting spectral contrast in Landsat thematic mapper image data using selective principal component analysis," *Photogramm. Eng. Remote Sens.*, vol. 55, no. 3, pp. 339–348, 1989.

[15] A. Garzelli, F. Nencini, and L. Capobianco, "Optimal MMSE pan sharpening of very high resolution multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 1, pp. 228–236, Jan. 2008.

[16] 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.

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

[18] J.-L. Xiao, T.-Z. Huang, L.-J. Deng, Z.-C. Wu, X. Wu, and G. Vivone, "Variational pansharpening based on coefficient estimation with nonlocal regression," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023, Art. no. 5406115, doi: 10.1109/TGRS.2023.3305296.

[19] Z.-C. Wu et al., "LRTCFPan: Low-rank tensor completion based framework for pansharpening," *IEEE Trans. Image Process.*, vol. 32, pp. 1640–1655, 2023.

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

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

[22] 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.

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

[24] C. Jin, L.-J. Deng, T.-Z. Huang, and G. Vivone, "Laplacian pyramid networks: A new approach for multispectral pansharpening," *Inf. Fusion*, vol. 78, pp. 158–170, Feb. 2022.

[25] M. Zhou, J. Huang, F. Zhao, and D. Hong, "Modality-aware feature integration for pan-sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, 2023, Art. no. 5400312, doi: 10.1109/TGRS.2022.3232384.

[26] R. Ran, L.-J. Deng, T.-X. Jiang, J.-F. Hu, J. Chanussot, and G. Vivone, "GuidedNet: A general CNN fusion framework via high-resolution guidance for hyperspectral image super-resolution," *IEEE Trans. Cybern.*, vol. 53, no. 7, pp. 4148–4161, Jul. 2023, doi: 10.1109/TCYB.2023.3238200.

[27] K. Zhang and H. Yang, "Semi-supervised multi-spectral land cover classification with multi-attention and adaptive kernel," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 1881–1885.

[28] X. Wang, Y. Cheng, X. Mei, J. Jiang, and J. Ma, "Group shuffle and spectral–spatial fusion for hyperspectral image super-resolution," *IEEE Trans. Comput. Imag.*, vol. 8, pp. 1223–1236, 2022.

[29] D. Zhao, Q. Wang, J. Zhang, and C. Bai, "Mine diversified contents of multispectral cloud images along with geographical information for multilabel classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023, Art. no. 4102415, doi: 10.1109/TGRS.2023.3270204.

[30] J. Hou, Q. Cao, R. Ran, C. Liu, J. Li, and L.-J. Deng, "Bidomain modeling paradigm for pansharpening," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 347–357.

[31] G. Yang, M. Zhou, K. Yan, A. Liu, X. Fu, and F. Wang, "Memory-augmented deep conditional unfolding network for pansharpening," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1778–1787.

[32] 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.

[33] W. Carper, T. Lillesand, and R. Kiefer, "The use of intensity-hue-saturation transformations for merging SPOT panchromatic and multispectral image data," *Photogramm. Eng. Remote Sens.*, vol. 56, no. 4, pp. 459–467, Apr. 2004.

[34] 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.

[35] M. Ghahremani and H. Ghassemian, "Nonlinear IHS: A promising method for pan-sharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 11, pp. 1606–1610, Nov. 2016.

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

[37] L. Alparone, A. Garzelli, and G. Vivone, "Intersensor statistical matching for pansharpening: Theoretical issues and practical solutions," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4682–4695, Aug. 2017.

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

[39] 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.

[40] L.-J. Deng, G. Vivone, W. Guo, M. D. Mura, and J. Chanussot, "A variational pansharpening approach based on reproducible kernel Hilbert space and heaviside function," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4330–4344, Sep. 2018.

[41] X. Fu, Z. Lin, Y. Huang, and X. Ding, "A variational pan-sharpening with local gradient constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10257–10266.

[42] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[43] H. Shen, Z.-Q. Zhao, and W. Zhang, "Adaptive dynamic filtering network for image denoising," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 2, 2023, pp. 2227–2235.

[44] X. Chen, J. Pan, J. Lu, Z. Fan, and H. Li, "Hybrid CNN-transformer feature fusion for single image deraining," in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, no. 1, pp. 378–386.

[45] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[46] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[47] Q. Xu, Y. Li, M. Zhang, and W. Li, "COCO-Net: A dual-supervised network with unified ROI-loss for low-resolution ship detection from optical satellite image sequences," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022, Art. no. 5629115, doi: 10.1109/TGRS.2022.3201530.

[48] Z.-R. Jin, T.-J. Zhang, T.-X. Jiang, G. Vivone, and L.-J. Deng, "LAGConv: Local-context adaptive convolution kernels with global harmonic bias for pansharpening," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 1, pp. 1113–1121.

[49] J. Ma, W. Yu, C. Chen, P. Liang, X. Guo, and J. Jiang, "Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion," *Inf. Fusion*, vol. 62, pp. 110–120, Oct. 2020.

[50] Q. Xu, Y. Li, J. Nie, Q. Liu, and M. Guo, "UPanGAN: Unsupervised pansharpening based on the spectral and spatial loss constrained generative adversarial network," *Inf. Fusion*, vol. 91, pp. 31–46, Mar. 2023.

[51] M. Zhou, J. Huang, X. Fu, F. Zhao, and D. Hong, "Effective pansharpening by multiscale invertible neural network and heterogeneous task distilling," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022, Art. no. 5411614, doi: 10.1109/TGRS.2022.3199210.

[52] L. Ardizzone, C. Lüth, J. Kruse, C. Rother, and U. Köthe, "Guided image generation with conditional invertible neural networks," 2019, *arXiv:1907.02392*.

[53] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, Nov. 2000.

[54] M. Werman, S. Peleg, and A. Rosenfeld, "A distance metric for multidimensional histograms," *Comput. Vis., Graph., Image Process.*, vol. 32, no. 3, pp. 328–336, Dec. 1985.

[55] I. Yusuf, G. Igwegbe, and O. Azeez, "Differentiable histogram with hard-binning," 2020, *arXiv:2012.06311*.

[56] L. He et al., "Pansharpening via detail injection based convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 4, pp. 1188–1204, Apr. 2019.

[57] S. Xu, J. Zhang, Z. Zhao, K. Sun, J. Liu, and C. Zhang, "Deep gradient projection networks for pan-sharpening," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1366–1375.

[58] M. Zhou, J. Huang, Y. Fang, X. Fu, and A. Liu, "Pan-sharpening with customized transformer and invertible neural network," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 3, pp. 3553–3561.

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

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

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

[62] R. H. Yuhas, A. F. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm," in *Proc. JPL, Summaries 3rd Annu. JPL Airborne Geosci. Workshop.*, vol. 1, 1992, pp. 1–3.

[63] L. Wald, *Data Fusion: Definitions and Architectures: Fusion of Images of Different Spatial Resolutions*. Paris, France, 2002.

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

[65] 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.

[66] B. Aiazzi, L. Alparone, S. Baronti, R. Carla, A. Garzelli, and L. Santurri, "Full-scale assessment of pansharpening methods and data products," *Proc. SPIE*, vol. 9244, Oct. 2014, Art. no. 924402.

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