Pixel-Wise Ensembled Masked Autoencoder for Multispectral Pansharpening

Yongchuan Cui, *Graduate Student Member, IEEE*, Peng Liu, *Member, IEEE*, Yan Ma, *Member, IEEE*, Lajiao Chen, Mengzhen Xu, and Xingyan Guo

***Abstract*— Pansharpening requires the fusion of a low-spatial-resolution multispectral (LRMS) image and a panchromatic (PAN) image with rich spatial details to obtain a high-spatial-resolution multispectral (HRMS) image. Recently, deep learning (DL)-based models have been proposed to tackle this problem and have made considerable progress. However, most existing methods rely on the conventional observation model, which treats LRMS as a blurred and downsampled version of HRMS. This observation model may lead to unsatisfactory performance and limited generalization ability at full-resolution evaluation, resulting in severe spectral and spatial distortion, as we observed that while DL-based models show significant improvement over traditional models on reduced-resolution evaluation, their performances deteriorate significantly at full resolution. In this article, we rethink the observation model and present a novel perspective from HRMS to LRMS and propose a pixel-wise ensembled masked autoencoder (PEMAE) to restore HRMS. Specifically, we consider LRMS as the result of pixel-wise masking on HRMS. Thus, LRMS can be seen as a natural input of a masked autoencoder. By ensembling the reconstruction results of multiple masking patterns, PEMAE obtains HRMS with both spectral information of LRMS and spatial details of PAN. In addition, we employ a linear cross-attention mechanism to replace the regular self-attention to reduce the computation to linear time complexity. Extensive experiments demonstrate that PEMAE outperforms state-of-the-art (SOTA) methods in terms of quantitative and visual performance at both reduced- and full-resolution evaluations. The codes are available at https://github.com/yc-cui/PEMAE.**

***Index Terms*— Deep learning (DL), image fusion, masked autoencoder, multispectral pansharpening.**

I. Introduction

DUE to physical constraints [1] such as signal-to-noise ratio and diffraction limit, it is difficult to obtain high-spatial-resolution multispectral remote sensing (HRMS) images. Instead, modern satellites, such as WorldView, QuickBird, and GaoFen, can only acquire low-spatial-resolution multispectral (LRMS) images and high-spatial-resolution panchromatic (PAN) images. One way to acquire HRMS is to use PAN to sharpen LRMS, known as pansharpening.

Traditional pansharpening methods leverage prior knowledge to establish robust models and have been widely used in practice. However, these methods are limited in highly nonlinear mapping [3] and their strong assumptions are proved to be unrealistic from the perspective of remote sensing

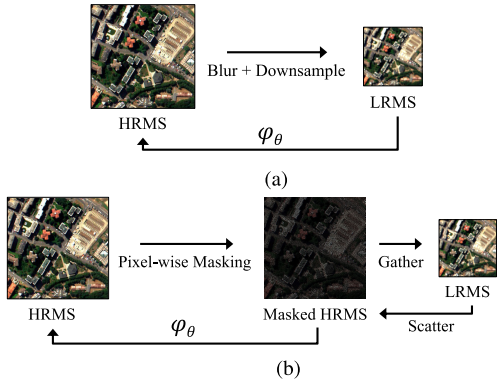

Fig. 1. Comparison of observation models. Traditionally, LRMS is treated as a blurred and downsampled version of HRMS. Thus, researchers usually train a network φ_θ by using LRMS as ground truth and LRMS↓-PAN↓ as inputs and then apply φ_θ to LRMS. In contrast, this article treats LRMS as a pixel-wise masked version of HRMS. We scatter the pixels of LRMS to simulate the masking process and use MAE [2] to recover the masked images. (a) Classical observation model. (b) Proposed observation model.

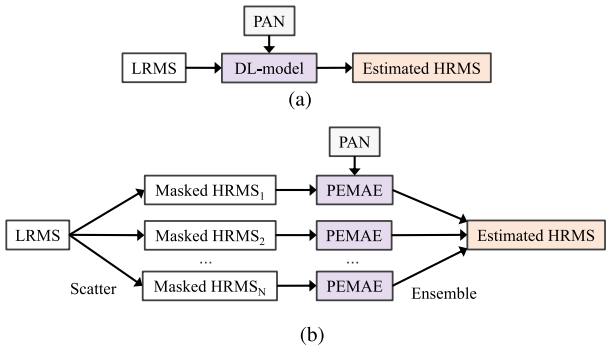

Fig. 2. Comparison of training paradigms. Most DL-based models typically use LRMS and PAN as inputs to obtain an estimation of HRMS directly. However, in this article, LRMS is utilized to obtain multiple masked HRMS through *scattering*. Subsequently, these various reconstructed images are then integrated together to obtain an accurate estimation of HRMS. (a) Classical training paradigm. (b) Proposed training paradigm.

Manuscript received 8 March 2024; revised 6 July 2024; accepted 27 July 2024. Date of publication 27 August 2024; date of current version 25 November 2024. This work was supported by the National Science Foundation of China under Grant U2243222, Grant 41971397, Grant 41471368, Project Y1H103101A, and Project Y5J0100. *(Corresponding author: Peng Liu.)*

Yongchuan Cui, Peng Liu, Yan Ma, and Lajiao Chen are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, and also with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: liupeng202303@aircas.ac.cn).

Mengzhen Xu and Xingyan Guo are with the State Key Laboratory of Hydroscience and Engineering, Department of Hydraulic Engineering, Tsinghua University, Beijing 100190, China.

Digital Object Identifier 10.1109/TGRS.2024.3450688

physics [4], [5], which leads to potential spatial or spectral distortion. In contrast to traditional methods, the recently developed deep learning (DL)-based methods aim to learn the optimal mapping from observations to the ideal fused HRMS from a large number of samples [6] and have achieved the state-of-the-art (SOTA) performance.

Due to the lack of HRMS, DL-based methods employ Wald's protocol [7] to produce simulated data to train networks at reduced-resolution. Wald's protocol requires blurring and downsampling the LRMS-PAN pair to obtain LRMS$\downarrow$-PAN$\downarrow$ ($\downarrow$ represents the operation of blurring and downsampling). Then, DL-based models take LRMS$\downarrow$ and PAN$\downarrow$ as inputs and regard LRMS as the ground truth, seeking the optimal mapping from LRMS$\downarrow$ to LRMS. The network parameterized with θ learns a mapping φ_θ, where $\varphi_\theta(\text{LRMS}\downarrow, \text{PAN}\downarrow) \rightarrow \text{LRMS}$. As a matter of course, it performs well in terms of metrics computed at reduced resolution. However, when evaluating at full resolution to get HRMS through the same network $\varphi_\theta(\text{LRMS}, \text{PAN})$, the result is not satisfactory. Besides, when generalizing φ_θ to new satellites for testing, existing DL-based models exhibit significant degradation in performance at both reduced and full resolutions, even inferior to traditional methods. This implies that these models fail to be interoperable, i.e., lacking satellite-agnostic capabilities. This is known as domain shift, also referred to as a scale-related problem [3], [8], [9].

To address the aforementioned issues, this article made modifications to the observation model and employed a masked autoencoder as a reconstruction network, which significantly enhanced the ability to generalize to full-resolution and new satellites. The traditional observation model and the training paradigm of existing models are built based on Figs. 1(a) and 2(a), respectively, where LRMS is considered as the blurred and downsampled version of HRMS and the reconstruction network φ_θ is designed to directly learn the mapping from low to high spatial resolution. In contrast, this article introduces a novel observation model to tackle the performance gap caused by domain shift. As shown in Fig. 1(b), LRMS is regarded as the result of applying a pixel-wise mask to HRMS (we refer to this step as *scattering*) and *gathering* the unmasked pixels. Based on this assumption, the training paradigm is shown in Fig. 2(b), where pixel-wise ensembled masked autoencoder (PEMAE) is proposed to restore the masked HRMS. Different from vanilla MAE [2], our PEMAE takes pixel-wise inputs instead of patch-wise ones. Also, we modified the attention mechanism by introducing a cross-attention with linear time complexity to reduce calculations. We employed ensembling because the unknowability of HRMS makes it impossible to obtain the subsampling grids. In other words, the pixel positions sampled from HRMS are not unique, which leads to multiple masked HRMS. Different scatter schemes contain distinct spatial information while preserving spectral information intact. By ensembling the recovery results under various scatter schemes, we finally obtain the estimated HRMS with rich spatial information and comprehensive spectral information. The experimental results show that the proposed mask-scatter-ensemble strategy achieved better performance on both full-reference and no-reference metrics, and generalized well to other satellites.

In summary, our contributions can be summarized as the following three points.

1) We provide a novel observation model from HRMS to LRMS from the perspective of pixel-wise masking, rather than simply blurring and downsampling. The masking mechanism models real-world scenarios in a more complicated manner. By ensembling the reconstruction results of multiple masked HRMS, we can obtain more robust results.
2) We propose a pansharpening network based on masked autoencoder [2] with modified attention mechanism to efficiently reconstruct masked HRMS, dubbed as PEMAE. To the best of our knowledge, this is the first work to introduce the pixel-wise masked autoencoder [2] for pansharpening.
3) Extensive experiments verify that our model achieves excellent performance at both reduced- and full-resolution evaluations compared with other SOTA pansharpening approaches, regardless of visual effects, quantitative metrics, and generalization abilities. The deployment codes, pretrained models, and training logs are publicly made available at https://github.com/yc-cui/PEMAE.

The remaining sections of this article are organized as follows. Section II provides a review on pansharpening, covering both traditional and DL-based methods. Section III explains the methods proposed in this article, including pixel-wise masking restoration and ensembling sample spaces. Section IV introduces the experimental setups. Section V presents a detailed comparison of the experimental results and provides an in-depth analysis of the experiments. Finally, Section VI concludes this article and discusses future work.

II. Related Work

With the notable advancements in DL, the focus of pansharpening research is shifting from traditional approaches to deep models. This section provides a concise overview of the relevant methods for pansharpening from both perspectives and then introduces masked autoencoder [2].

A. Traditional Pansharpening Methods

Traditional approaches for pansharpening have been widely used in practice, such as component substitution (CS), multiresolution analysis (MRA), and variational optimization (VO). These methods do not require a large amount of data for training but rely on reasonable assumptions, as well as handcrafted feature extraction and fusion techniques based on domain knowledge of pansharpening. In CS-based methods, HRMS is obtained by a projection-substitution [4], [6] manner: LRMS is projected into a new space and the structural component is substituted with the PAN image. Typical algorithms include intensity–hue–saturation (IHS) fusion [10], principal component analysis (PCA) [11], Gram–Schmidt adaptive (GSA) transform [12], and so on. MRA-based methods apply a multiresolution transformation to the PAN image, extract high-frequency spatial information, and then inject it into the upsampled LRMS to get HRMS. Commonly used

algorithms include smoothing filter-based intensity modulation (SFIM) [13], modulation transfer function generalized Laplace's pyramid (MTF-GLP) [14], [15], and so on. Since the proposal of P + XS [16], VO-based methods have gained ever-increasing attention. VO-based methods consider the relationships between HRMS, LRMS, and PAN to extract priors and formulate pansharpening as an optimization problem. Palsson et al. [17] utilized total variation regularization to address the ill-posed problem inherent in the pansharpening process. Vivone et al. [18] proposed a novel semiblind deconvolution approach to estimate the blur filter. Sun et al. [19] introduced a coupled temporal variation information estimation and resolution enhancement model, which significantly enhanced spectral fidelity. Although these algorithms are widely used in practice, they are prone to spectral and spatial distortion due to limited nonlinear capabilities of feature representation.

B. DL-Based Pansharpening Methods

DL-based methods for remote sensing image fusion [20], [21], [22], [23], [24] have garnered significant attention with the rapid development of deep neural networks. Characterized by a vast number of learnable parameters, such methodologies necessitate extensive datasets for effective training. PNN [25] is a pioneering work utilizing convolutional neural networks (CNNs) for supervised pansharpening. Subsequently, Yang et al. [26] proposed PanNet and explored the impact of different ResNet [27] structures. Following PanNet, a series of CNN-based models were proposed, such as DCFNet [28], GPPNN [29], HyperKite [30], and PGCU [31]. Due to the limitations of the convolutional kernels, CNN-based models cannot learn spatial structural features over large distances. Consequently, several refinements to the kernels of CNN have been proposed. CANConv [32] and KNLConv [33] employ adaptive convolution to filter different regions using distinct kernels, allowing for spatial adaptability. Compared to CNN, Transformer [34], [35] utilizes the multihead self-attention mechanism to effectively capture contextual information for feature extraction. Therefore, many researchers have introduced Transformer as the backbone in pansharpening. For instance, DR-NET [36] used Swin Transformer [37] to extract features and proposed two attention-based modules that allow the model to focus on important information during reconstruction. HyperTransformer [38] regarded LRMS and PAN as queries and keys, respectively. This cross-attention mechanism effectively integrates the spectral information from LRMS and the spatial information from PAN. Other Transformer-based methods include PanFormer [39], CTINN [40], and so on. In addition to CNN- and Transformer-based approaches, generating more realistic and high-quality outputs using deep generative models has also become popular. Most of these are based on generative adversarial networks (GANs) [41], [42], [43] and the recently emerging diffusion models [44], such as MDSSC-GAN [45], HPGAN [46], PSGAN [47], PanDiff [48], and DDRF [49].

Pansharpening is an ill-posed problem, and it may not be reasonable [5], [50], [51] to solely rely on simple blurring and downsampling to simulate the process from HRMS to LRMS. Excessive dependence on simulated data could introduce bias when applied to real-world scenes. As a result, researchers have started exploring network architectures that do not require ground truth, i.e., *unsupervised pansharpening*. For instance, Pan-GAN [5] utilized two discriminators to produce HRMS that possesses both the spectral characteristics of LRMS and the spatial information of PAN. PLRDiff [52] combined the diffusion model [44] and low-rank matrix factorization technique to improve the generalization ability. Z-PNN [50] is a framework specifically designed to train the network at the full resolution. λ-PNN [53] used a novel joint enhancement of spectral and spatial fidelity loss to simultaneously promote the spectral and spatial quality of pansharpened images. Sun et al. [54] proposed an unsupervised 3-D tensor subspace decomposition network for spatial–temporal–spectral fusion. Besides these, other unsupervised models, such as UCGAN [8], MetaPan [55], and UP-SAM [56], also achieved better results at full-resolution evaluation. Despite considerable progress in unsupervised pansharpening, the observation models from HRMS to LRMS in these methods still remain conventional. According to the reported results, unsupervised methods are still not competitive compared with supervised ones.

Although these models achieve good performances on full-reference metrics, they perform poorly when tested on no-reference metrics or generalized to other satellites, due to scale-related issues during training. In this article, we introduce a novel observation model and a reconstruction model based on MAE [2] to tackle these problems.

C. Masked Autoencoder

MAE is a scalable self-supervised learning model proposed by He et al. [2]. The idea of MAE [2] is quite simple. It takes masked images as inputs, encodes visible patches, and then decodes the obtained latent representations along with mask tokens to reconstruct the original images. The encoder and decoder in MAE [2] are asymmetric. The encoder incorporates a higher number of parameters, while the decoder remains lightweight. This is because the training objective is to obtain a powerful and robust encoder for downstream tasks such as image classification and object detection. However, pansharpening needs more accurate reconstruction results rather than generalized representations. Therefore, in our work, we prioritize the decoder as the primary model in order to achieve more precise restoration. The decoder, in comparison to the encoder, is equipped with a greater number of heads and a deeper layer. Although MAE [2] was initially introduced for self-supervised learning, in this article, we integrate this mask image modeling strategy into pansharpening and propose PEMAE, and the experimental results have shown that the proposed approach is highly effective. In the following, we will unleash the potential of MAE [2] for pansharpening.

III. Methodology

The proposed PEMAE is shown in Fig. 3. In this section, we first introduce the proposed novel observation model in Section III-A and then elaborate on details of the pixel-wise

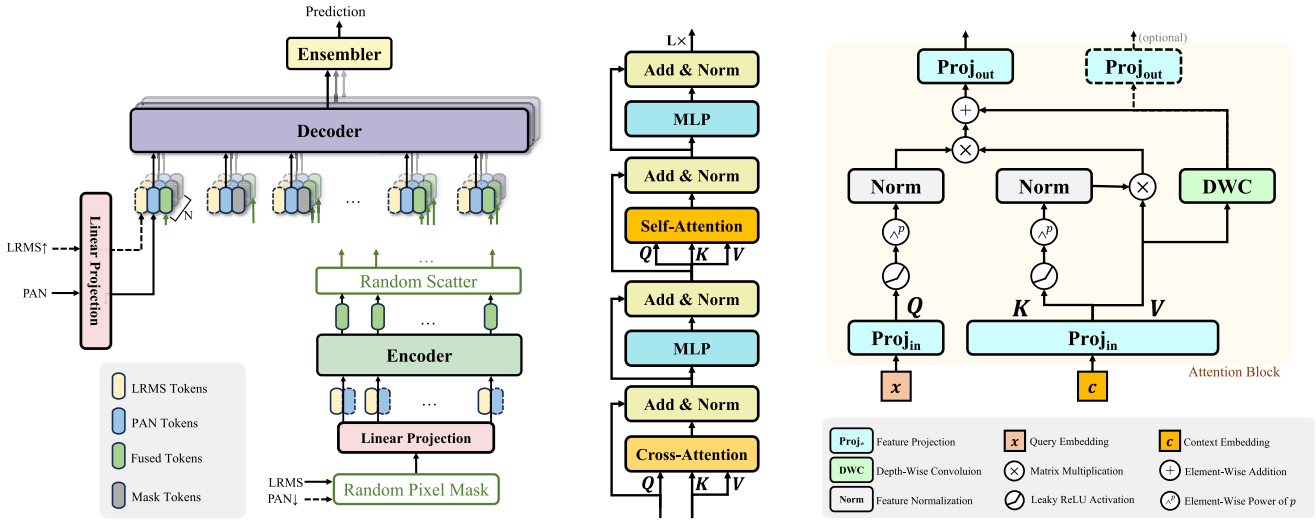

Fig. 3. (a) Overall framework of PEMAE. (b) Detailed block design in encoder and decoder. (c) Modified linear cross-attention. The dashed Proj_{out} will be used as a context embedding input to the subsequent cross-attention block. It is notable that although the network utilizes an ensembling strategy, it is still trained using an end-to-end approach. Our design allows for straightforward control over the number of ensembled nodes by simply adjusting the parameter N. Note that paths with the dashed line are optional. The data flow from $\mathcal{Y}$ and $\mathcal{P}$ to $\hat{\mathcal{X}}$ (symbol definitions can be found in Section III-B) is given as follows: $\mathcal{Y} \xrightarrow{\text{scatter}} \{\widetilde{\mathcal{X}}_1, \widetilde{\mathcal{X}}_2, \ldots, \widetilde{\mathcal{X}}_N\} \xrightarrow[\text{PEMAE}(\widetilde{\mathcal{X}}_n, \mathcal{P})]{\text{reconstruct}} \{\hat{\mathcal{X}}_1, \hat{\mathcal{X}}_2, \ldots, \hat{\mathcal{X}}_N\} \xrightarrow{\text{ensemble}} \hat{\mathcal{X}} \overset{\text{loss}}{\longleftrightarrow} \mathcal{X}$.

masking restoration and ensembling sample spaces operation in Sections III-B and III-C, respectively.

A. Proposed Observation Model

Let $\boldsymbol{Y} \in \mathbb{R}^{hw \times C}$ represent an observed LRMS image, where C is the number of channels, and h and w are the height and width, respectively. $\boldsymbol{P} \in \mathbb{R}^{HW \times 1}$ represents the corresponding PAN image of $\boldsymbol{Y}$ and $\boldsymbol{X} \in \mathbb{R}^{HW \times C}$ is the HRMS image that needs to be reconstructed. Let $\hat{\boldsymbol{X}} \in \mathbb{R}^{HW \times C}$ represent the estimated output of HRMS. Typically, traditional observation models assume that LRMS is blurred and downsampled by HRMS [29], [57], [58], which can be expressed as follows:

$$\boldsymbol{Y} = \boldsymbol{D}\boldsymbol{K}\boldsymbol{X} + \boldsymbol{N} \tag{1}$$

where $\boldsymbol{D} \in \mathbb{R}^{hw \times HW}$ denotes a downsampling matrix and $\boldsymbol{K}$ is a (low-passing) circular convolution matrix [29], which is equivalent to the process of blurring. The variable $\boldsymbol{N}$ represents unpredictable random noise. Most DL-based methods take simulated $\boldsymbol{Y}$ as the input and learn the inverse process of downsampling and blurring via a network φ_θ with learnable parameters θ. After obtaining the prediction $\hat{\boldsymbol{X}}$, the loss is calculated with the ground truth $\boldsymbol{X}$, and then, backward propagation is used to optimize the parameters θ, which can be formulated as

$$\begin{aligned} \hat{\boldsymbol{X}} &= \varphi_\theta(\boldsymbol{Y}, \boldsymbol{P}) \\ \mathcal{L} &= \text{Loss}\big(\hat{\boldsymbol{X}}, \boldsymbol{X}\big). \end{aligned} \tag{2}$$

This process corresponds to Figs. 1(a) and 2(a). However, this observation model and training paradigm suffer from poor generalization ability. The performance of φ_θ is notably subpar when applied to full resolution or data from other satellites.

In contrast, the observation model proposed in this article is formulated as

$$\boldsymbol{Y} = G(\boldsymbol{M}\boldsymbol{K}\boldsymbol{X}) + \boldsymbol{N} \tag{3}$$

where $\boldsymbol{M} \in \mathbb{R}^{HW \times HW}$ is a diagonal matrix, which represents randomly masking pixels across the channel axis, and the form of $\boldsymbol{M}$ is as follows:

$$\boldsymbol{M} = \begin{bmatrix} m_1 & 0 & 0 & \cdots & 0 \\ 0 & m_2 & 0 & \cdots & 0 \\ 0 & 0 & m_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & m_{HW} \end{bmatrix}. \tag{4}$$

The diagonal elements of matrix $\boldsymbol{M}$ can only be either 0 or 1, i.e., $m_i \in \{0, 1\}, i = 1, 2, \ldots, HW$. The sum of the diagonal elements in matrix $\boldsymbol{M}$ must be equal to hw, i.e., $\sum_{i=1}^{HW} \boldsymbol{M}_{\text{ii}} = hw$. $G(\cdot)$ is a utility function used to gather all nonzero rows. Pansharpening is an ill-posed problem, indicating that it is impossible to determine the value of $\boldsymbol{M}$. From the perspective of our proposed observation model, (2) is unreasonable because the abovementioned DL-based models only have a single output, which means φ_θ implicitly set a single estimate for $\boldsymbol{M}$, while in reality, $\boldsymbol{M}$ can take on $C_{HW}^{hw} = (HW!)/((HW - hw)!hw!)$ different situations. This observation model also explains why these networks have poor generalization ability, as when conducting full-resolution evaluation or testing on new satellites, $\boldsymbol{M}$ for the new data may undergo significant changes, rendering these models unable to adapt to the new data distribution.

To ensure that the model can adapt to any variations in the mask matrix $\boldsymbol{M}$ when the data distribution changes, in this article, our proposed solution is to employ an ensembling strategy, i.e., randomly sampling multiple instances of $\boldsymbol{M}$

and aggregating the reconstructed results obtained under each sampling mode. The process can be expressed as follows:

$$\begin{aligned}\hat{\boldsymbol{X}} &= E\big[\varphi_\theta\big(S_{\boldsymbol{M}_1}(\boldsymbol{Y}), \boldsymbol{P}\big), \varphi_\theta\big(S_{\boldsymbol{M}_2}(\boldsymbol{Y}), \boldsymbol{P}\big), \ldots, \\ &\qquad \varphi_\theta\big(S_{\boldsymbol{M}_N}(\boldsymbol{Y}), \boldsymbol{P}\big)\big] \\ \mathcal{L} &= \mathrm{Loss}\big(\hat{\boldsymbol{X}}, \boldsymbol{X}\big) \end{aligned} \tag{5}$$

where $E(\cdot)$ represents ensembling its inputs and $S_{\boldsymbol{M}_n}(\cdot)$ is the process of scatter $\boldsymbol{Y}$ under a random sampled $\boldsymbol{M}_n$ $(n = 1, 2, \ldots, N)$ to obtain a masked HRMS. N is the number of masking patterns. This process corresponds to Figs. 1(b) and 2(b). $S_{\boldsymbol{M}_n}(\cdot)$ is the inverse process of $G(\cdot)$, where $G(\cdot)$ gathers unmasked pixels, while $S_{\boldsymbol{M}_n}(\cdot)$ randomly disperses the pixels of $\boldsymbol{Y}$. In (5), φ_θ does not need to estimate the mask matrix implicitly because we utilize a random policy to explicitly generate multiple masks. This allows the network to overcome the limitation of being confined to only one pattern of $\boldsymbol{M}$, thereby enhancing its generalization ability. Since the proposed observation model involves masked image reconstruction, we adopt a masked autoencoder [2] as the reconstruction network φ_θ. In the following, we will elaborate on details of the process of scattering and ensembling.

B. Design of PEMAE

Based on the proposed observation model and the preceding analysis, this article introduces PEMAE to reconstruct the masked HRMS.

1) Overall Architecture: The overall architecture of PEMAE is depicted in Fig. 3(a). Vanilla MAE [2] aims to extract high-level semantic features and thus designs the encoder more heavily. However, pansharpening is a low-level pixel restoration task focusing on minimizing reconstruction error, and thus, we have designed the decoder with a greater capacity. Fig. 3(b) illustrates the detailed design of the blocks within the encoder and decoder. Different from the original vision transformer (ViT) [35] block, to enhance the integration of information between the two distinct modalities of PAN and LRMS, we employ a two-step attention mechanism. Initially, cross-attention is utilized to enhance the interaction between the two modalities. In this process, the LRMS embeddings serve as queries and the PAN embeddings act as keys and values, facilitating spatial matching between LRMS and PAN. Subsequently, self-attention is applied to further fuse the information and refine the spectral integrity.

Due to the pixel-level masking operation, the sequence length varies directly with the number of pixels. Performing self-attention on thousands or even millions of sequences becomes impractical and unbearable in terms of time and memory consumption. PEMAE modifies the original self-attention mechanism that is quadratically proportional to the sequence length into a linear attention mechanism, thereby addressing computational efficiency concerns. Specifically, we adopted focused linear attention [59] with some necessary modifications to achieve efficient computation on extremely large sequences. Fig. 3(c) displays the modified linear cross-attention, with detailed specifics provided in Section III-B3.

2) Mask and Scatter: As depicted in Fig. 3(a), LRMS undergoes a masking operation before entering the encoder. Upon reaching the bottleneck before the decoder, extracted features are scattered using N different schemes to model various spatial relationships. Multiple decoder outputs are generated through these schemes, which are subsequently ensembled. In our initial design, LRMS served as the sole input to the encoder, while PAN was introduced as an additional input at the bottleneck. However, experimental results revealed suboptimal performance with this configuration. Simultaneous input of LRMS↑ and PAN at the bottleneck yielded superior model performance. Therefore, optional LRMS↑ and PAN↓ paths were incorporated to explore the optimal embedding input, as depicted by the dashed arrows in Fig. 3(a). Further ablation experiments are elaborated in Section V-D1.

During model training, the original LRMS imagery serves as a reference, with the objective of reconstructing the LRMS from the masked version. It should be noted that masking the LRMS is infeasible during full-resolution inference. Consequently, the masking step is omitted in full-resolution inference, and the LRMS is employed directly as the input to the encoder. After obtaining the embedded features of masked LRMS, the next step is to scatter $\boldsymbol{Y}$ to get a masked version of HRMS in the bottleneck, i.e., the process of $S_{\boldsymbol{M}_n}(\boldsymbol{Y})$. With abuse of notations, we use calligraphic typeface to represent the tensor shapes of the aforementioned $\boldsymbol{X}$, $\hat{\boldsymbol{X}}$, $\boldsymbol{Y}$, and $\boldsymbol{P}$ in the neural network, namely, $\mathcal{X} \in \mathbb{R}^{H\times W\times C}$, $\hat{\mathcal{X}} \in \mathbb{R}^{H\times W\times C}$, $\mathcal{Y} \in \mathbb{R}^{h\times w\times C}$, and $\mathcal{P} \in \mathbb{R}^{H\times W}$. Note that for simplicity, $\mathcal{Y}$ is assumed to be an embedding that has been masked, gathered, and extracted by the encoder, i.e., fused tokens, as shown in Fig. 3(a). The scattering process is given as follows. Suppose that we have a zero matrix $\widetilde{\mathcal{X}}_n$, and its shape is the same as $\mathcal{X}$

$$\widetilde{\mathcal{X}}_n = \mathbf{0}. \tag{6}$$

For every pixel position (i, j) in $\mathcal{Y}$, we sample its position index (p, q) in $\widetilde{\mathcal{X}}_n$

$$\begin{aligned} p &\sim U[i \times s, i \times s + s) \\ q &\sim U\big[j \times s, j \times s + s\big) \end{aligned} \tag{7}$$

where U represents sampling from a uniform distribution of discrete integers within its given interval. s is the scale factor, i.e., the scale ratio of the spatial resolution between HRMS and LRMS. Note that we assume that an HRMS image is uniformly masked, and thus, (7) restricts the sampling grids to the nearest neighbor sampling pattern. After sampling the position index, the next step is to assign a value to $\widetilde{\mathcal{X}}_n$

$$\widetilde{\mathcal{X}}_n(p, q) \leftarrow \mathcal{Y}(i, j). \tag{8}$$

The scattering mechanism is similar to the inverse of nearest neighbor downsampling. It restricts the pixels in $\mathcal{Y}$ at position (i, j) to s^2 grids in $\mathcal{X}$ corresponding to the same location. Fig. 4 displays the scattering process, and by repeating this process, we obtain $\{\widetilde{\mathcal{X}}_1, \widetilde{\mathcal{X}}_2, \ldots, \widetilde{\mathcal{X}}_N\}$ under different $\boldsymbol{M}$'s with unique spatial information. For $\widetilde{\mathcal{X}}_n$, it will be sent to the

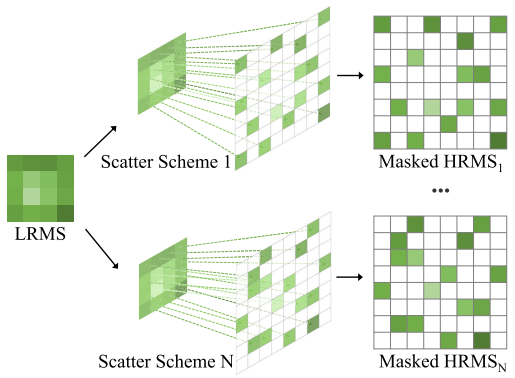

Fig. 4. Illustration of randomly scattering. One scatter scheme corresponds to a process of obtaining simulated masked HRMS.

decoder of PEMAE

$$\hat{\mathcal{X}}_n = \text{Decoder}\left(\widetilde{\mathcal{X}}_n, \mathcal{P}\right). \tag{9}$$

Up to this point, we have obtained the reconstructed $\hat{\mathcal{X}}_n$ from $\widetilde{\mathcal{X}}_n$, which is one estimate of $\mathcal{X}$.

3) Efficient Computation: The computational complexity of the decoder in ViT [35] used by the naive MAE [2] is quadratical to the sequence length. The attention operation is given as

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d}}\right)V \tag{10}$$

where $Q \in \mathbb{R}^{HW\times d}$, $K \in \mathbb{R}^{HW\times d}$, and $V \in \mathbb{R}^{HW\times d}$ denote keys, queries, and values, respectively; HW is the pixel counts of $\mathcal{X}$; and d is the dimension of the input vector. In (10), the computational complexity is approximately $\mathcal{O}(2H^2W^2d)$, including $\mathcal{O}(H^2W^2d)$ from QK^{T}, and $\mathcal{O}(H^2W^2d)$ from $\text{Softmax}(\cdot)V$. To reduce the computation, we employ the focused linear attention proposed in [59] to replace the naive attention. The idea of focused linear attention is to break the computation constraints brought by $\text{Softmax}(\cdot)$, which traditionally requires a quadratic number of computations with respect to the sequence length. The modified linear attention is shown in Fig. 3(c). To effectively approximate the softmax function while maintaining a linear time complexity, given a feature vector $x \in \mathbb{R}^d$, a focused function ϕ_p is introduced as

$$\phi_p(x) = f_p(\text{LeakyReLU}(x)) \tag{11}$$

where $f_p(x) = (\|x\|)/(\|x^p\|)x^p$ is a mapping function to adjust the direction of feature vectors, x^p represents the element-wise power p of x, $\|x\|$ is the norm of x, and p is a focused factor to enlarge the distinguished difference between similar query and key pairs. The similarity between query Q_i and key K_j is then computed as

$$\text{Sim}\left(Q_i, K_j\right) = \phi_p(Q_i)\phi_p\left(K_j\right)^{\text{T}}. \tag{12}$$

By decomposing $\text{Softmax}(\cdot)$ into $\phi_p(\cdot)$, the alternative attention can be expressed as

$$\text{Attention}(Q, K, V) = \phi_p(Q)\phi_p(K)^{\text{T}}V. \tag{13}$$

However, (12) cannot produce sharp distribution as the original $\text{Softmax}(\cdot)$ due to the low rank of the attention matrix. Thus, to increase the rank and restore the sharp attention, a depth-wise convolution (DWC) is applied to the value V as follows:

$$\text{Attention}(Q, K, V) = \phi_p(Q)\phi_p(K)^{\text{T}}V + \text{DWC}(V). \tag{14}$$

The additional DWC layer increases the effective rank of the attention matrix, thus restoring the diversity of the output features. Note that in [59], ReLU activation is used to enhance expressive capability and inhibit unimportant features. However, the use of the ReLU activation function in pansharpening can result in suboptimal outcomes due to its tendency to produce a large number of zero activations, which can hinder the convergence of the network. Thus, we employ LeakyReLU as an alternative activation function to adequately capture the complex relationships between LRMS and PAN.

C. Ensembling Sample Spaces

Because scattering is a random operation, the reconstructed result from $\widetilde{\mathcal{X}}_n$ only contains unique spatial information associated with its scatter scheme, as shown in Fig. 4. This process with inherent uncertainty leads to multiple random mask results. Due to variations in the masked pixels, the reconstructions obtained from different masked images contain diverse spatial information. The reconstructed $\hat{\mathcal{X}}_n$ from a single scatter scheme is not accurate since it is impossible to make sure whether the pixels of $\mathcal{Y}$ after scattering are in the true positions of $\mathcal{X}$. Thus, to enhance the stability and precision of the final model and capture a more comprehensive representation, this article employs an ensembled strategy.

Specifically, we propose to use multiple scatter schemes for ensembling to obtain HRMS combining a variety of different spatial information modes. As shown in Fig. 3(a), we obtain $\{\widetilde{\mathcal{X}}_1, \widetilde{\mathcal{X}}_2, \ldots, \widetilde{\mathcal{X}}_N\}$ in the bottleneck after N scatter schemes. After each $\widetilde{\mathcal{X}}_n$ undergoes reconstruction, $\{\hat{\mathcal{X}}_1, \hat{\mathcal{X}}_2, \ldots, \hat{\mathcal{X}}_N\}$ are obtained as the estimation of $\mathcal{X}$. They are concatenated together by the channel axis and sent into the ensembling model $E(\cdot)$ to get $\hat{\mathcal{X}}$. The ensemble model $E(\cdot)$ in this article is quite simple, which only contains three layers of residual block [27] for convolution operation. The cross-attention in PEMAE has already done most of the reconstruction work and there is no reason for a more complex ensembling model.

D. Relationship Between PEMAE and MAE

Our proposed PEMAE is fundamentally rooted in MAE [2]. However, PEMAE distinguishes itself through several innovative modifications tailored to the specific demands of pansharpening. The primary differences are given as follows.

1) Our approach introduces an ensemble mechanism specifically crafted for the observation model proposed in this article. Contrary to the vanilla MAE [2], which presumes a priori knowledge of the position of masked patches during the restoration, we employ random scattering to recover masked HRMS by sampling multiple possible positions.
2) Different from the single-image modality input of vanilla MAE [2], PEMAE incorporates the dual modalities of LRMS and PAN imagery. Therefore, distinct inputs were

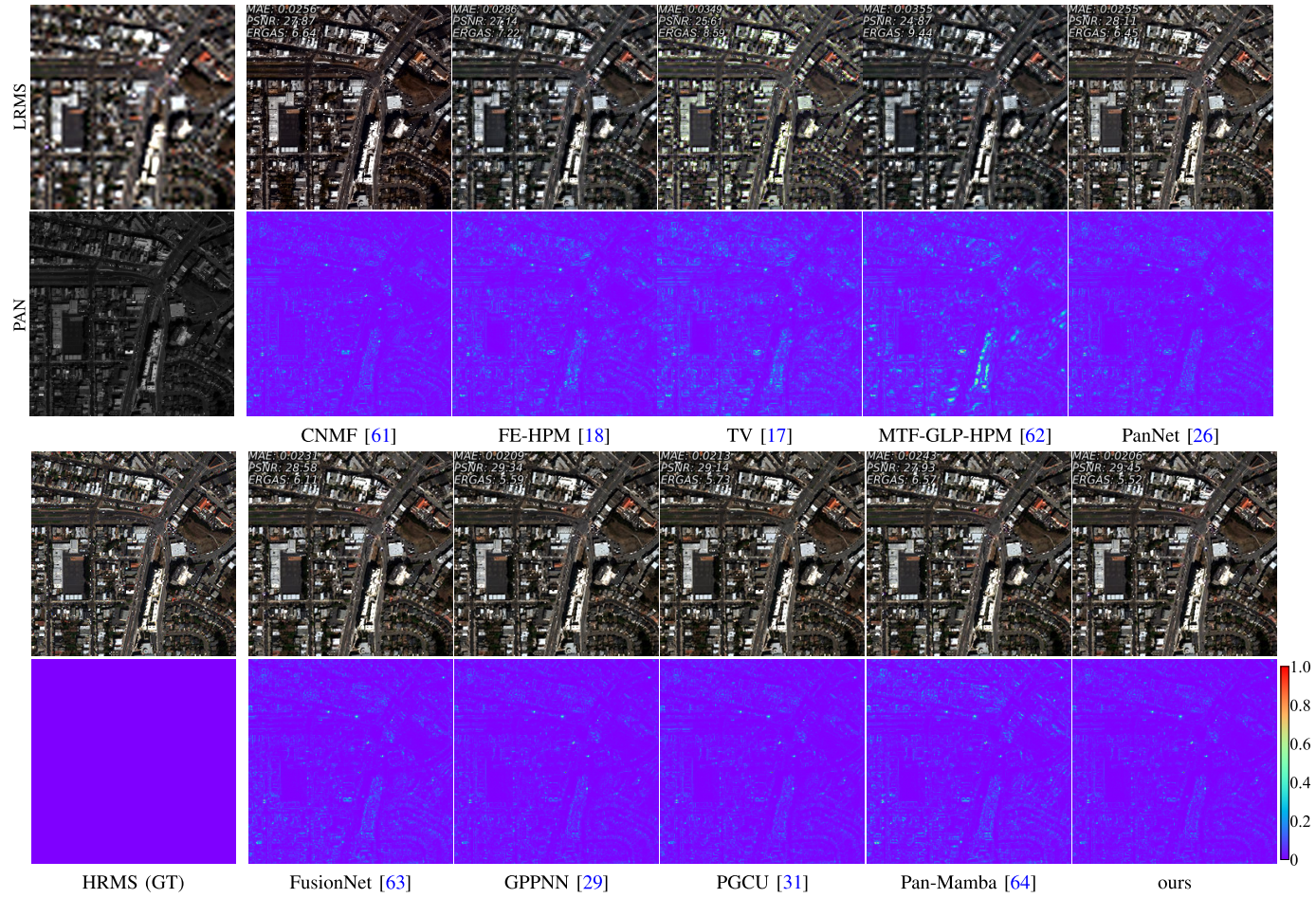

Fig. 5. Visual comparison on WorldView-2 at reduced resolution. Please zoom in for best view.

devised for the encoder and decoder to investigate how these two modalities influence the reconstruction results.

3) In order to learn semantic-level information, vanilla MAE [2] chooses a patch size of 16 × 16 and explores the influence of the mask ratio through fine-tuning and linear probing. In contrast, the patch size of the PEMAE proposed in this article is set to pixel-level which is 1. This is because pansharpening is a low-level vision task of pixel reconstruction, which requires each pixel to be accurately reconstructed. Besides, the mask ratio is fixed and related to the ratio of high to low spatial resolutions, which is constrained by the task specification.
4) To address the issue of inadequate feature fusion and excessive computational demands, we introduced cross-attention and modified linear attention mechanism in the encoder and decoder blocks of PEMAE. Within the linear attention mechanism, we opted LeakyReLU activation function over ReLU to enhance the model's ability to capture complex patterns. In addition, we substituted standard convolutions with deformable convolutions (DCNs) [60] for feature projection, thereby further enhancing the model's capacity for representation and improving the overall performance.

IV. Experimental Setups

A. Datasets and Implementation Details

We use the large-scale NBU_PansharpRSData[1] datasets [6] to evaluate the performance of all models. NBU_PansharpRSData contains LRMS-PAN image pairs collected from diverse satellites, which encompasses a wide variety of land uses and land covers. Each training pair comprises one LRMS image measuring 256 × 256 pixels and one PAN image measuring 1024 × 1024 pixels. In this article, we validate our proposed method using GaoFen-1, WorldView-2, WorldView-3, and IKONOS data, including both reduced- and full-resolution validations, as well as cross-sensor generalization testing (assessing the performance of models trained on source satellite imagery when applied to target satellite data).

Although the proposed model is trained in a supervised manner, we do not rely on Wald's protocol [7] to construct the training pairs. The original MS image still serves as an HRMS reference. However, the model does not receive the downsampled MS and PAN images as input. Instead, as depicted in Fig. 3, the MS image is randomly masked, and the nonmasked pixels are used as encoder inputs. The PAN image is linearly projected to an embedding of the same

[1]https://github.com/starboot/NBU_PansharpRSData

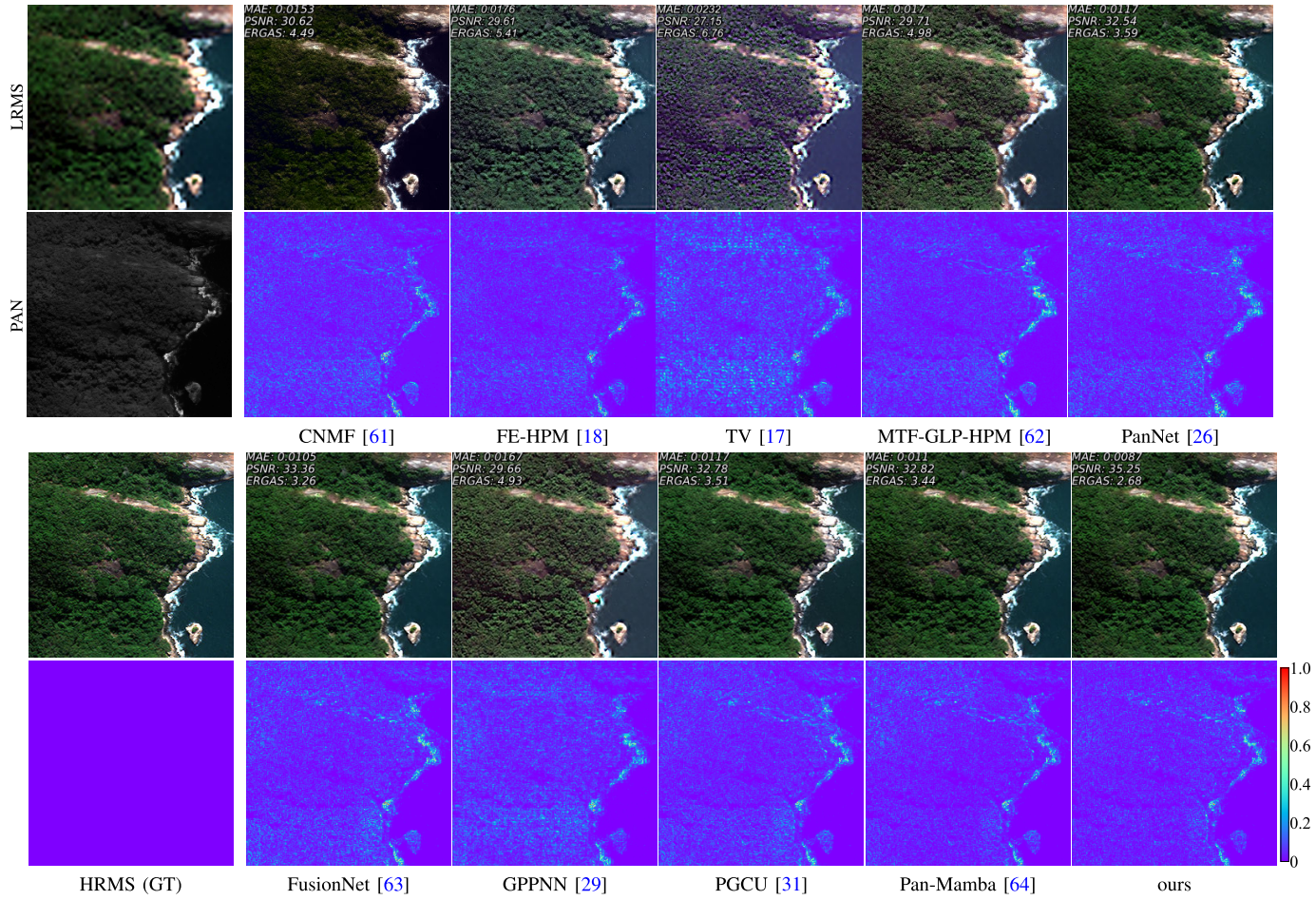

Fig. 6. Visual comparison on WorldView-3 at reduced-resolution. Please zoom in for best view.

size as HRMS before being fed into the decoder. Consistent with the proposed observation model, the masking strategy enables reconstructing the original image from a diverse subset of pixels, thereby effectively mitigating the information loss typically associated with downsampling solely based on Wald's protocol [7].

The network architecture is implemented with PyTorch version v2.0.1. We use the AdamW [65] optimizer (with $\beta_1 = 0.9$ and $\beta_2 = 0.999$) to update the weights of the network iteratively. The network is trained with a learning rate of 0.001 until convergence. During training, other than the mentioned parameters, all the hyperparameters are set to the same values as provided by the official MAE [2] implementation.[2] All experiments were conducted using an NVIDIA GeForce RTX 4090 D with 24 GB of memory. Codes, pretrained weights, and training logs are available at https://github.com/yc-cui/PEMAE.

B. Comparative Methods

We compare the proposed PEMAE with several competitive methods containing nine commonly recognized traditional methods, including Brovey [70], IHS [10], SFIM [13], GS [71], MTF-GLP-HPM [62], GSA [12], CNMF [61], TV [17], and FE-HPM [18]; and 11 SOTA DL-based methods, including PanNet [26], FusionNet [63], GPPNN [29], MDCUN [66], PGCU [31], S^{2}DBPN [67], UAPN [68], UTSN [69], Pan-Mamba [64], Z-PNN [50], and λ-PNN [53]. The traditional methods are from the benchmark toolbox[3] [72], [73]. The DL-based methods are listed in Table I. SL refers

TABLE I
DL-BASED COMPARATIVE METHODS. SL: SUPERVISED LEARNING. UL: UNSUPERVISED LEARNING

Model	Type	Published in	Year
PanNet [26]	SL	ICCV	2017
FusionNet [63]	SL	TGRS	2021
GPPNN [29]	SL	CVPR	2021
MDCUN [66]	SL	CVPR	2022
PGCU-PanNet [31]	SL	CVPR	2023
S2DBPN [67]	SL	TGRS	2023
UAPN-B [68]	SL	TGRS	2023
UTSN [69]	SL	TGRS	2023
Pan-Mamba [64]	SL	ECCV	2024
Z-PNN [50]	UL	TGRS	2022
λ-PNN [53]	UL	TGRS	2023

[2]https://github.com/facebookresearch/mae

[3]https://github.com/liangjiandeng/DLPan-Toolbox

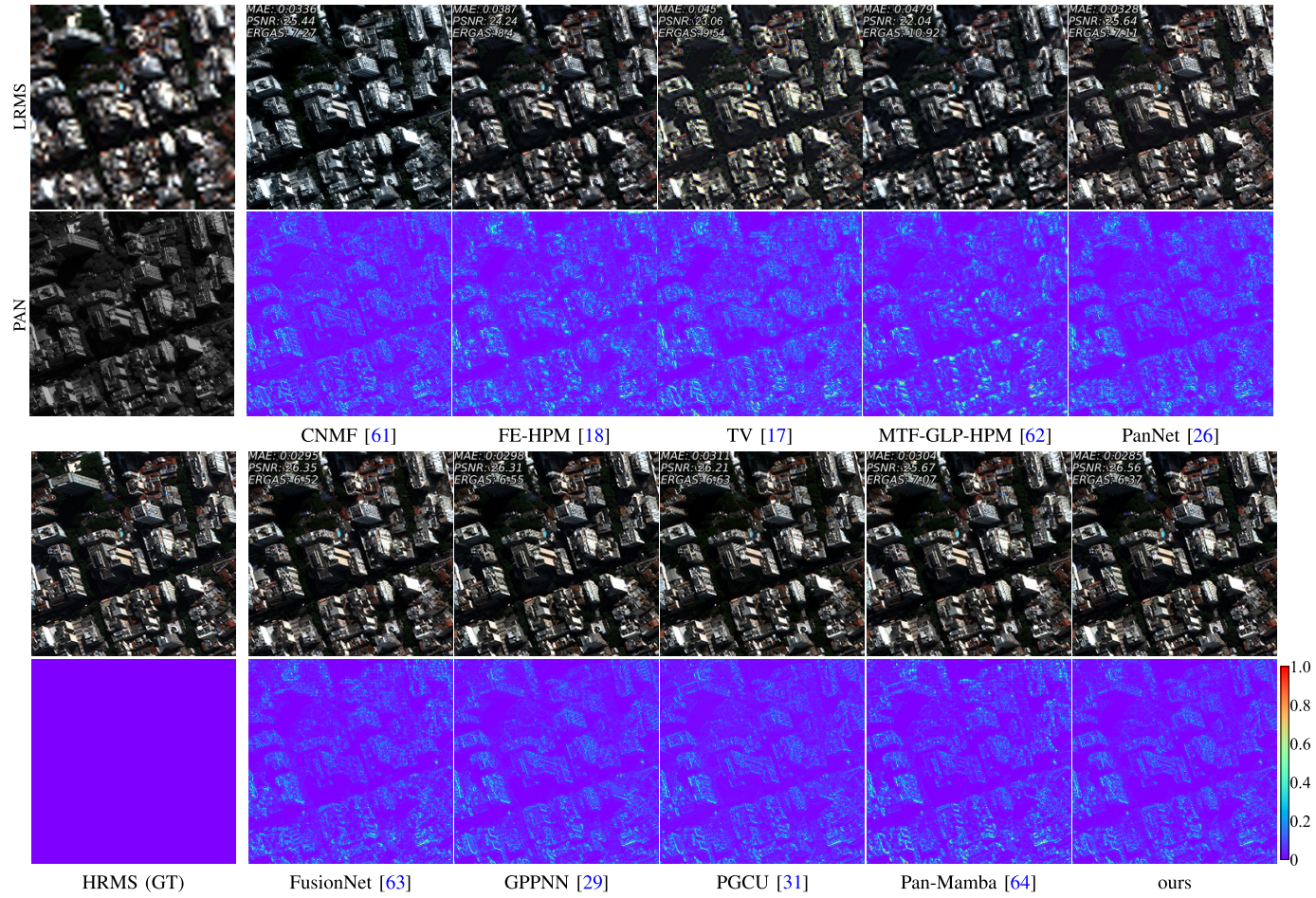

Fig. 7. Visual comparison of results generated by representative methods to test the generalization capability on the WorldView-3 satellite at reduced resolution. The DL-based models are trained using WorldView-2 data. Please zoom in for best view.

to supervised learning, where models are trained on labeled data following Wald's protocol [7]. UL refers to unsupervised learning, where models are trained at full resolution without ground-truth labels.

C. Evaluation Metrics

For quantitative comparisons, the mean absolute error (MAE), the erreur relative globale adimensionnelle de synthèse (ERGAS) [7], the peak signal-to-noise ratio (PSNR), and the correlation coefficients (CCs) are employed as reduced-resolution metrics. The quality without reference (QNR) and its two components, including spectral distortion index D_λ and spatial distortion index D_S, and the correlation distortion index D_ρ [74] are employed as no-reference metrics, which will be evaluated at full resolution. For qualitative evaluation, the reconstructed images and the error heatmaps are provided for visual comparison.

1) The MAE metric measures the average magnitude of the errors between the predicted and the actual values. It is defined as

$$\mathrm{MAE} = \frac{1}{C}\sum_{i=1}^{C}\frac{1}{HW}\sum_{j=1}^{HW}\left|\mathbf{A}_i^j - \mathbf{B}_i^j\right| \tag{15}$$

where C is the number of bands; H and W are the height and width of the inputs, respectively; and $\mathbf{A}_i^j$ and $\mathbf{B}_i^j$ are the actual and predicted values for the jth pixel in the ith band.

2) The CC metric measures the geometric similarity between two signals and is defined as

$$\mathrm{CC} = \frac{1}{C}\sum_{i=1}^{C}\frac{\sum_{j=1}^{HW}\left(\mathbf{A}_i^j - \mu_{\mathbf{A}_i}\right)\left(\mathbf{B}_i^j - \mu_{\mathbf{B}_i}\right)}{\sqrt{\sum_{j=1}^{HW}\left(\mathbf{A}_i^j - \mu_{\mathbf{A}_i}\right)^2\sum_{j=1}^{HW}\left(\mathbf{B}_i^j - \mu_{\mathbf{B}_i}\right)^2}} \tag{16}$$

where C, H, and W are the number of bands, height, and width of inputs, respectively; and $\mu_{\mathbf{A}_i}$ calculates the mean value of the ith band of $\mathbf{A}$.

3) The PSNR metric measures the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of the signal. It is calculated using the formula

$$\mathrm{PSNR} = \frac{1}{C}\sum_{i=1}^{C}10\log_{10}\left(\frac{\max(\mathbf{B}_i)}{\mathrm{RMSE}(\mathbf{A}_i, \mathbf{B}_i)}\right)^2 \tag{17}$$

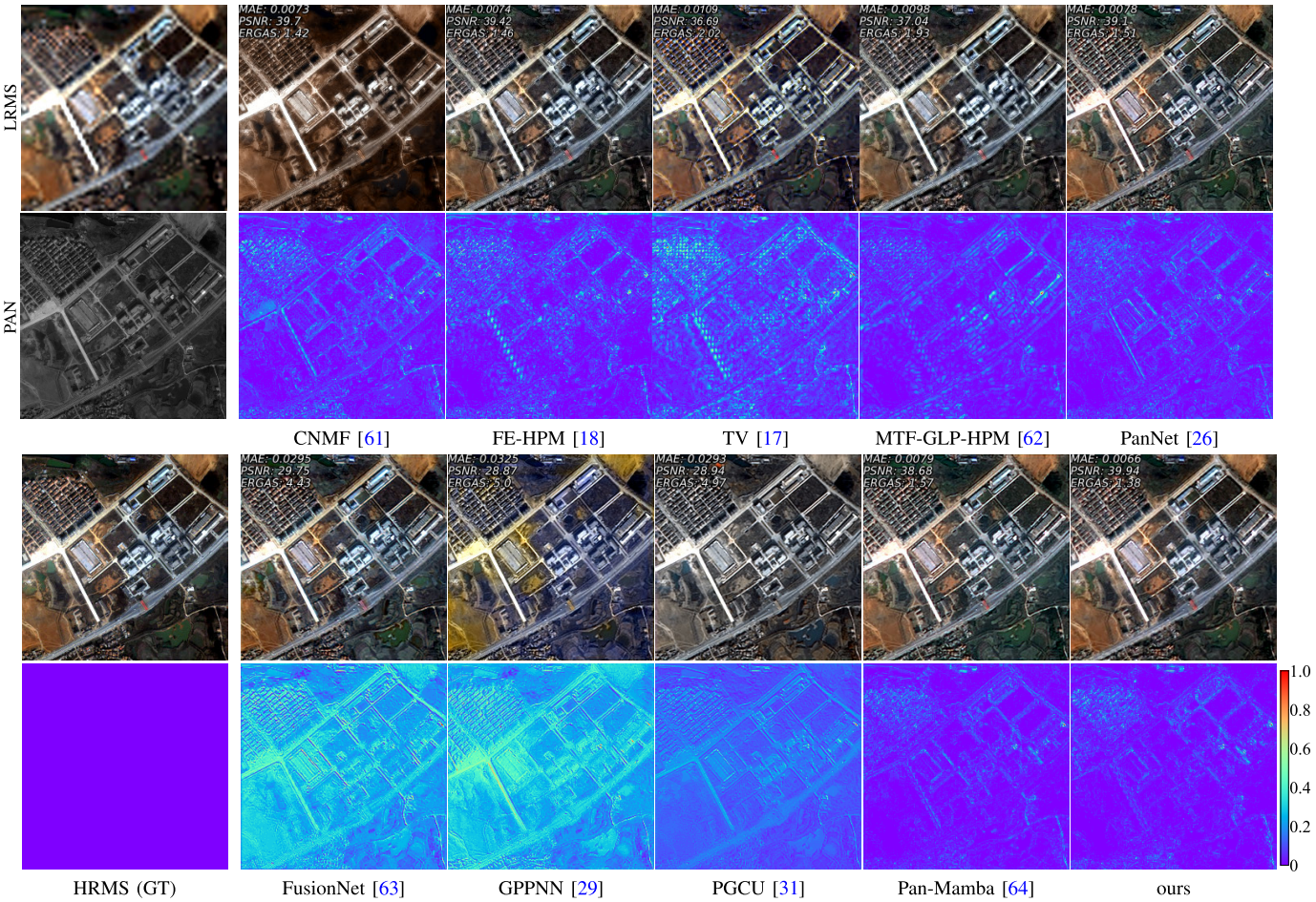

Fig. 8. Visual comparison of results generated by representative methods to test the generalization capability on the IKONOS satellite at reduced resolution. The DL-based models are trained using WorldView-3 data. Please zoom in for best view.

TABLE II

COMPARISON RESULTS WITH SOTA METHODS AT REDUCED RESOLUTION. COLOR CONVENTION: THE BEST AND SECOND BEST AMONG ALL ALGORITHMS. ↓: LOWER IS BETTER. ↑: HIGHER IS BETTER

Model	WorldView-2				WorldView-3				IKONOS				GaoFen-1			
	MAE↓	PSNR↑	CC↑	ERGAS↓	MAE↓	PSNR↑	CC↑	ERGAS↓	MAE↓	PSNR↑	CC↑	ERGAS↓	MAE↓	PSNR↑	CC↑	ERGAS↓
Brovey [70]	0.0190	32.17	0.8911	5.703	0.0242	29.70	0.8787	6.118	0.0130	35.44	0.8592	2.439	0.0159	34.29	0.7638	2.314
IHS [10]	0.0200	32.07	0.8863	5.797	0.0251	29.62	0.8769	6.288	0.0132	35.32	0.8552	2.508	0.0162	34.43	0.7483	2.424
SFIM [13]	0.0209	30.84	0.8625	6.674	0.0266	28.41	0.8590	7.187	0.0126	34.98	0.8997	2.568	0.0154	34.21	0.8478	2.439
GS [71]	0.0186	32.38	0.8995	5.502	0.0236	29.95	0.8950	5.984	0.0120	35.94	0.9130	2.314	0.0164	33.91	0.8012	2.540
MTF-GLP-HPM [62]	0.0239	29.02	0.8510	8.774	0.0304	26.96	0.8517	8.909	0.0146	33.56	0.8884	2.997	0.0178	32.74	0.8173	2.841
GSA [12]	0.0172	32.56	0.9048	5.331	0.0223	29.88	0.8968	6.025	0.0105	36.72	0.9224	2.151	0.0148	34.17	0.8627	2.287
CNMF [61]	0.0169	32.75	0.9090	5.218	0.0222	30.00	0.8986	5.968	0.0107	36.58	0.9193	2.142	0.0120	36.23	0.8881	1.906
TV [17]	0.0238	30.13	0.8650	7.104	0.0295	27.74	0.8588	7.625	0.0165	33.27	0.8762	3.070	0.0180	32.99	0.8145	2.728
FE-HPM [18]	0.0189	31.26	0.8973	6.572	0.0247	28.74	0.8922	7.075	0.0116	35.74	0.9239	2.351	0.0151	34.34	0.8685	2.389
PanNet [26]	0.0153	33.42	0.9195	4.912	0.0191	31.51	0.9243	5.013	0.0087	38.12	0.9448	1.828	0.0065	41.17	0.9623	1.105
FusionNet [63]	0.0124	35.03	0.9429	4.082	0.0152	33.16	0.9480	4.078	0.0077	38.89	0.9541	1.670	0.0054	42.90	0.9722	0.927
GPPNN [29]	0.0127	34.82	0.9401	4.184	0.0158	32.86	0.9444	4.194	0.0083	38.45	0.9476	1.752	0.0067	41.03	0.9581	1.144
MDCUN [66]	0.0145	33.61	0.9225	4.806	0.0180	31.48	0.9246	4.946	0.0090	37.59	0.9373	1.920	0.0077	39.12	0.9418	1.377
PGCU [31]	0.0134	34.34	0.9358	4.384	0.0168	32.36	0.9386	4.402	0.0092	37.63	0.9401	1.901	0.0078	39.47	0.9453	1.332
S^2DBPN [67]	0.0144	33.89	0.9268	4.647	0.0196	30.84	0.9147	5.394	0.0089	37.96	0.9431	1.851	0.0076	39.76	0.9490	1.285
UAPN-B [68]	0.0162	33.12	0.9151	4.966	0.0222	30.03	0.8924	5.721	0.0116	35.62	0.8980	2.410	0.0077	39.50	0.9468	1.316
UTSN [69]	0.0121	35.42	0.9484	3.894	0.0152	33.29	0.9497	3.968	0.0084	38.48	0.9472	1.735	0.0071	40.29	0.9531	1.217
Pan-Mamba [64]	0.0150	32.88	0.9083	5.233	0.0163	32.31	0.9393	4.455	0.0085	37.97	0.9440	1.840	0.0046	44.38	0.9801	0.785
PEMAE(ours)	0.0106	36.63	0.9606	3.389	0.0134	34.32	0.9598	3.510	0.0067	40.23	0.9639	1.469	0.0044	44.90	0.9811	0.758

where RMSE is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{HW}\sum_{j=1}^{HW}\left(\mathbf{A}_i^j - \mathbf{B}_i^j\right)^2}. \tag{18}$$

4) The ERGAS metric provides a comprehensive assessment of the quality of a fused product and is given by

$$\mathrm{ERGAS} = \frac{100}{s}\sqrt{\frac{1}{C}\sum_{i=1}^{C}\left(\frac{\mathrm{RMSE}(\mathbf{A}_i, \mathbf{B}_i)}{\mu_{\mathbf{B}_i}}\right)^2} \tag{19}$$

TABLE III

GENERALIZATION COMPARISON WITH SOTA METHODS AT REDUCED RESOLUTION. COLOR CONVENTION: THE BEST AND SECOND BEST AMONG ALL ALGORITHMS. ↓: LOWER IS BETTER. ↑: HIGHER IS BETTER

Model	WorldView-2 → WorldView-3				WorldView-3 → IKONOS				IKONOS → GaoFen-1				GaoFen-1 → WorldView-2			
	MAE↓	PSNR↑	CC↑	ERGAS↓	MAE↓	PSNR↑	CC↑	ERGAS↓	MAE↓	PSNR↑	CC↑	ERGAS↓	MAE↓	PSNR↑	CC↑	ERGAS↓
PanNet [26]	0.0210	30.57	0.9077	5.594	0.0108	36.20	0.9197	2.284	0.0122	35.52	0.8903	2.015	0.0221	30.48	0.8246	7.003
FusionNet [63]	0.0193	31.19	0.9227	5.186	0.0276	29.69	0.8703	4.526	0.0155	33.91	0.8890	2.310	0.0244	29.66	0.7844	7.608
GPPNN [29]	0.0203	30.81	0.9199	5.393	0.0327	28.73	0.8462	5.137	0.0241	30.72	0.8587	3.131	0.0305	27.56	0.7902	10.30
MDCUN [66]	0.0200	30.60	0.9081	5.504	0.0154	33.91	0.8810	2.975	0.0118	35.87	0.8905	1.932	0.0228	29.88	0.8096	7.344
PGCU [31]	0.0228	29.80	0.9164	5.829	0.0418	26.60	0.8433	7.018	0.0285	30.31	0.8058	3.633	0.0332	27.29	0.7695	9.180
S^2DBPN [67]	0.0247	29.32	0.9061	6.194	0.0337	28.50	0.8611	5.525	0.0203	31.78	0.8991	2.725	0.0257	29.41	0.8244	8.026
UAPN-B [68]	0.0211	30.66	0.9100	5.412	0.0113	35.96	0.9065	2.279	0.0095	37.55	0.9191	1.716	0.0229	30.05	0.8085	7.265
UTSN [69]	0.0226	30.09	0.9112	5.711	0.0386	27.37	0.8137	6.169	0.0304	29.41	0.8138	3.977	0.0427	24.66	0.7440	14.55
Pan-Mamba [64]	0.0202	30.32	0.9023	5.713	0.0153	33.74	0.8512	2.960	0.0075	39.41	0.9461	1.335	0.0226	29.99	0.8135	7.333
PEMAE(ours)	0.0159	32.64	0.9408	4.356	0.0154	34.11	0.8350	2.865	0.0128	34.89	0.8759	2.161	0.0197	31.38	0.8652	6.235

TABLE IV

COMPARISON RESULTS WITH SOTA METHODS AT FULL RESOLUTION. COLOR CONVENTION: THE BEST AND SECOND BEST AMONG ALL ALGORITHMS. ↓: LOWER IS BETTER. ↑: HIGHER IS BETTER

Model	WorldView-2				WorldView-3				IKONOS				GaoFen-1			
	QNR ↑	D_λ ↓	D_S ↓	D_ρ ↓	QNR ↑	D_λ ↓	D_S ↓	D_ρ ↓	QNR ↑	D_λ ↓	D_S ↓	D_ρ ↓	QNR ↑	D_λ ↓	D_S ↓	D_ρ ↓
Brovey [70]	0.7544	0.1051	0.1611	0.2431	0.7397	0.1033	0.1795	0.3094	0.5464	0.2383	0.2969	0.2466	0.4707	0.2104	0.4164	0.3609
IHS [10]	0.7359	0.1175	0.1740	0.2798	0.6995	0.1206	0.2133	0.3431	0.5172	0.2626	0.3150	0.2554	0.4461	0.2332	0.4352	0.3835
SFIM [13]	0.7614	0.1063	0.1546	0.1966	0.7603	0.1175	0.1493	0.2128	0.6387	0.1842	0.2320	0.1698	0.6274	0.1803	0.2472	0.1302
GS [71]	0.7938	0.0752	0.1468	0.2328	0.7488	0.0857	0.1886	0.2993	0.6717	0.1415	0.2293	0.1885	0.5222	0.1560	0.3903	0.3478
MTF-GLP-HPM [62]	0.7522	0.1166	0.1548	0.2222	0.7450	0.1245	0.1593	0.2376	0.6267	0.1970	0.2340	0.1969	0.5424	0.2159	0.3247	0.2593
GSA [12]	0.7497	0.0935	0.1789	0.2338	0.7179	0.1040	0.2082	0.2898	0.6906	0.1345	0.2125	0.1713	0.5768	0.1601	0.3256	0.2610
CNMF [61]	0.7186	0.1211	0.1906	0.2975	0.7081	0.1296	0.1946	0.3241	0.7089	0.1415	0.1835	0.1741	0.6528	0.1400	0.2452	0.1951
TV [17]	0.8848	0.0665	0.0526	0.1370	0.8624	0.0649	0.0769	0.1854	0.7920	0.0871	0.1349	0.1630	0.5398	0.2055	0.3368	0.1812
FE-HPM [18]	0.7490	0.1150	0.1604	0.1510	0.7409	0.1242	0.1642	0.1799	0.6205	0.1942	0.2449	0.1429	0.5063	0.2139	0.3746	0.2040
PanNet [26]	0.8525	0.0556	0.0983	0.1027	0.8501	0.0609	0.0993	0.1799	0.8077	0.0824	0.1218	0.0857	0.7428	0.1258	0.1519	0.0155
FusionNet [63]	0.8124	0.0671	0.1320	0.1847	0.8573	0.0520	0.0974	0.2325	0.7889	0.0888	0.1377	0.1125	0.7945	0.1056	0.1113	0.0426
GPPNN [29]	0.7922	0.0786	0.1453	0.2745	0.8197	0.0839	0.1084	0.3022	0.7523	0.1171	0.1576	0.1573	0.8013	0.0980	0.1114	0.0970
MDCUN [66]	0.8326	0.0594	0.1169	0.1673	0.8438	0.0579	0.1077	0.2224	0.7788	0.0963	0.1436	0.1199	0.7716	0.1161	0.1281	0.0289
PGCU [31]	0.6205	0.2489	0.1947	0.3103	0.6764	0.2182	0.1441	0.3257	0.6658	0.1841	0.1984	0.2141	0.8055	0.0960	0.1119	0.1084
S^2DBPN [67]	0.8257	0.0596	0.1236	0.1735	0.7880	0.0705	0.1584	0.2309	0.7137	0.1286	0.1930	0.1119	0.7721	0.1187	0.1245	0.0061
UAPN-B [68]	0.7705	0.0892	0.1573	0.1771	0.8054	0.0865	0.1226	0.1452	0.6644	0.1715	0.1989	0.0451	0.6996	0.1271	0.2004	0.0056
UTSN [69]	0.8121	0.0714	0.1296	0.2830	0.8102	0.0836	0.1222	0.3231	0.7667	0.1103	0.1497	0.1719	0.8172	0.1053	0.0866	0.1551
Pan-Mamba [64]	0.7902	0.1020	0.1307	0.2738	0.8674	0.0556	0.0835	0.2673	0.8276	0.0922	0.0936	0.2253	0.7211	0.1032	0.1982	0.2748
Z-PNN [50]	0.7985	0.1194	0.0975	0.2102	0.7819	0.1209	0.1140	0.3595	0.7630	0.1052	0.1515	0.3111	0.6414	0.1224	0.2772	0.2354
λ-PNN [53]	0.7858	0.1076	0.1205	0.2047	0.8282	0.0943	0.0920	0.3117	0.6761	0.1863	0.1809	0.3260	0.7741	0.0934	0.1523	0.2959
PEMAE(ours)	0.8782	0.0456	0.0800	0.1408	0.8962	0.0375	0.0694	0.2006	0.8168	0.0769	0.1189	0.1372	0.7373	0.1102	0.1712	0.0654

TABLE V

GENERALIZATION COMPARISON WITH SOTA METHODS AT FULL RESOLUTION. COLOR CONVENTION: THE BEST AND SECOND BEST AMONG ALL ALGORITHMS. ↓: LOWER IS BETTER. ↑: HIGHER IS BETTER

Model	WorldView-2 → WorldView-3				WorldView-3 → IKONOS				IKONOS → GaoFen-1				GaoFen-1 → WorldView-2			
	QNR ↑	D_λ ↓	D_S ↓	D_ρ ↓	QNR ↑	D_λ ↓	D_S ↓	D_ρ ↓	QNR ↑	D_λ ↓	D_S ↓	D_ρ ↓	QNR ↑	D_λ ↓	D_S ↓	D_ρ ↓
PanNet [26]	0.8568	0.0536	0.0989	0.8568	0.6320	0.2216	0.2020	0.2957	0.6954	0.1145	0.2165	0.1168	0.7333	0.0696	0.2162	0.0238
FusionNet [63]	0.7989	0.0734	0.1455	0.7989	0.7061	0.1685	0.1605	0.2434	0.6645	0.1279	0.2419	0.2302	0.7013	0.0638	0.2570	0.0768
GPPNN [29]	0.7726	0.0901	0.1613	0.7726	0.6331	0.2036	0.2165	0.2892	0.6816	0.1141	0.2327	0.2532	0.7617	0.0672	0.1847	0.2129
MDCUN [66]	0.7746	0.0893	0.1566	0.7746	0.5730	0.2673	0.2351	0.2229	0.7484	0.1047	0.1662	0.0965	0.7748	0.0694	0.1709	0.0542
PGCU [31]	0.6501	0.1979	0.2032	0.6501	0.6298	0.2659	0.1452	0.2752	0.4654	0.2485	0.4201	0.3930	0.7375	0.0762	0.2074	0.0612
S^2DBPN [67]	0.8500	0.0460	0.1103	0.8500	0.5531	0.2296	0.3051	0.2463	0.6366	0.1287	0.2737	0.1963	0.7532	0.0710	0.1918	0.0306
UAPN-B [68]	0.7474	0.0990	0.1771	0.7474	0.7226	0.1584	0.1487	0.1113	0.6771	0.2076	0.1474	0.0259	0.7039	0.0705	0.2460	0.0003
UTSN [69]	0.7988	0.0851	0.1371	0.7988	0.6734	0.1763	0.1927	0.3057	0.5802	0.1724	0.3167	0.3282	0.7200	0.0902	0.2068	0.2630
Pan-Mamba [64]	0.8059	0.0959	0.1237	0.8059	0.6276	0.2380	0.1915	0.3546	0.8045	0.0772	0.1283	0.1039	0.7540	0.0860	0.1808	0.0862
Z-PNN [50]	0.6631	0.2115	0.1808	0.6631	0.5044	0.3523	0.2393	0.3544	0.4602	0.2717	0.3762	0.3187	0.7295	0.1356	0.1692	0.1937
λ-PNN [53]	0.6752	0.2042	0.1718	0.6752	0.5310	0.3576	0.1830	0.3874	0.6135	0.2477	0.1883	0.2185	0.6723	0.1831	0.1849	0.3017
PEMAE(ours)	0.8782	0.0456	0.0800	0.8400	0.7797	0.1267	0.1115	0.1350	0.7486	0.1099	0.1607	0.1676	0.7365	0.1203	0.1686	0.1807

where s is the scale factor from high to low resolution (typically, the value of s is 4).

5) The QNR metric can be used when the ground-truth image is not available for comparison. It measures both the spectral and spatial distortion of the predicted image. The QNR is given by

$$\text{QNR} = (1 - D_\lambda)^\alpha (1 - D_S)^\beta \tag{20}$$

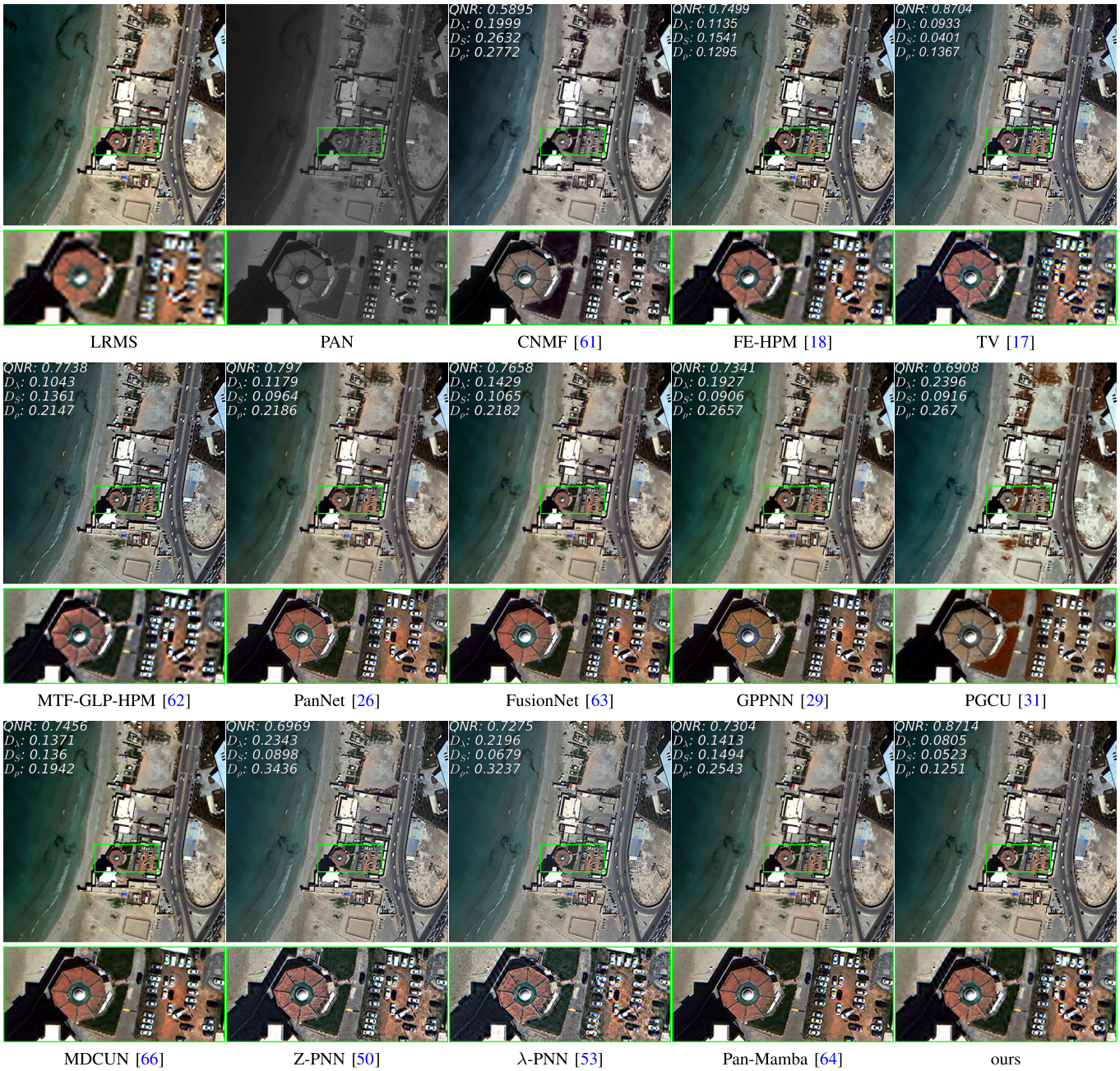

Fig. 9. Visual comparison on WorldView-3 at full resolution. The bright rectangular box represents the area zoomed in for display.

where α and β are two coefficients (usually set to 1). D_λ and D_S measure spectral and spatial distortion, respectively. They are defined as

$$D_\lambda = \left(\frac{1}{C(C-1)} \sum_{i=1}^{C} \sum_{j=1, j\neq i}^{C} \times \left\| Q\left(\mathbf{A}_i, \mathbf{A}_j\right) - Q\left(\mathbf{B}_i, \mathbf{B}_j\right) \right\|^p \right)^{\frac{1}{p}} \tag{21}$$

$$D_S = \left(\frac{1}{C} \sum_{i=1}^{C} \left\| Q(\mathbf{A}_i, \mathbf{P}) - Q\left(\mathbf{B}_i, \mathbf{P}_\downarrow\right) \right\|^q \right)^{\frac{1}{q}} \tag{22}$$

where p and q serve as coefficients to emphasize weights of distorted pixels (usually set to 1). In D_λ and D_S, $\mathbf{A}$ and $\mathbf{B}$ represent the estimated HRMS and the LRMS, respectively. $\mathbf{P}$ and $\mathbf{P}_\downarrow$ represent the PAN image and the scaled version of PAN to match the resolution of LRMS, respectively. $Q(\cdot, \cdot)$ is the Q index to calculate the dissimilarities between two images, and it is defined as

$$Q(\boldsymbol{x}, \boldsymbol{y}) = \frac{4\sigma_{\boldsymbol{x}\boldsymbol{y}}\mu_{\boldsymbol{x}}\mu_{\boldsymbol{y}}}{\left(\sigma_{\boldsymbol{x}}^2 + \sigma_{\boldsymbol{y}}^2\right)\left(\mu_{\boldsymbol{x}}^2 + \mu_{\boldsymbol{y}}^2\right)} \tag{23}$$

in which $\sigma_{\boldsymbol{x}\boldsymbol{y}}$ denotes the covariance between $\boldsymbol{x}$ and $\boldsymbol{y}$; $\mu_{\boldsymbol{x}}$ and $\mu_{\boldsymbol{y}}$ calculate the means; and $\sigma_{\boldsymbol{x}}^2$ and $\sigma_{\boldsymbol{y}}^2$ are the variances of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively.

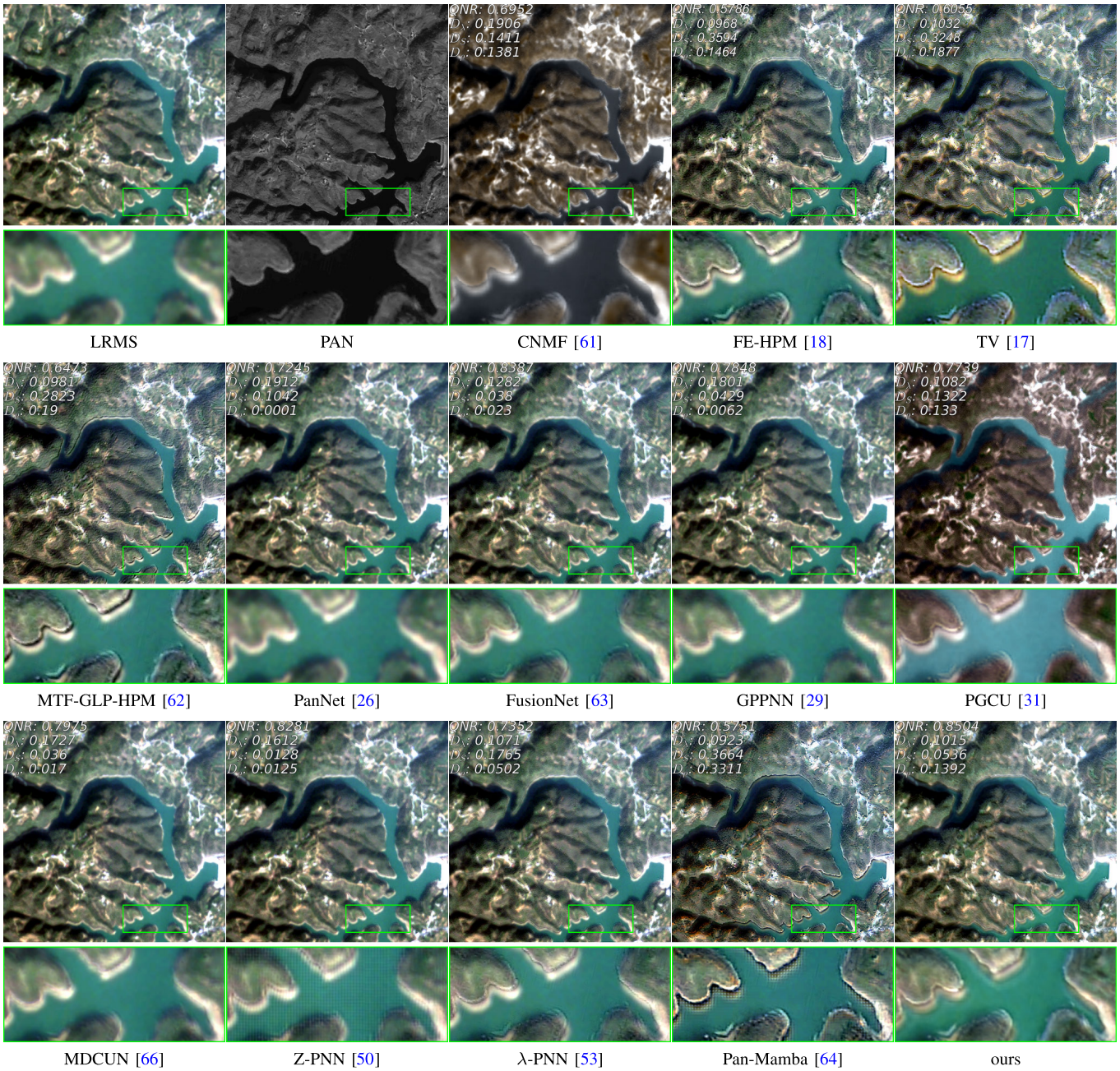

Fig. 10. Visual comparison on GaoFen-1 at full resolution. The bright rectangular box represents the area zoomed in for display.

6) The correlation distortion index D_ρ measures the spatial consistency between the pansharpened image and the PAN at full resolution. It is defined as

$$D_\rho = 1 - \rho_{\sigma P \hat{M}} \tag{24}$$

where $\rho_{\sigma P \hat{M}}$ is the average value of the local CCs between the PAN image P and the pansharpened image $\hat{M}$ over space and spectral bands, given by

$$\rho_{\sigma P \hat{M}} = \frac{1}{C}\sum_{i=1}^{C}\left\{\frac{1}{HW}\sum_{x=1}^{H}\sum_{y=1}^{W}\text{corrcoef}\left(P_{\sigma xy}, \hat{M}^i_{\sigma xy}\right)\right\} \tag{25}$$

where $X_{\sigma xy}$ denotes a $\sigma \times \sigma$ patch of image X centered at location (x, y). The function corrcoef$(\cdot)$ calculates the CC between two patches. $P_{\sigma xy}$ is the patch of the PAN image centered at (x, y), and $\hat{M}^i_{\sigma xy}$ is the corresponding patch in the ith band of the pansharpened image. The index D_ρ ranges from 0 (perfect correlation) to 1 (no correlation).

V. Experimental Results

This section evaluates the performance of PEMAE by comparing it with other SOTA pansharpening methods through extensive experiments. Furthermore, we investigate other factors that may affect PEMAE's performance, e.g., the impact

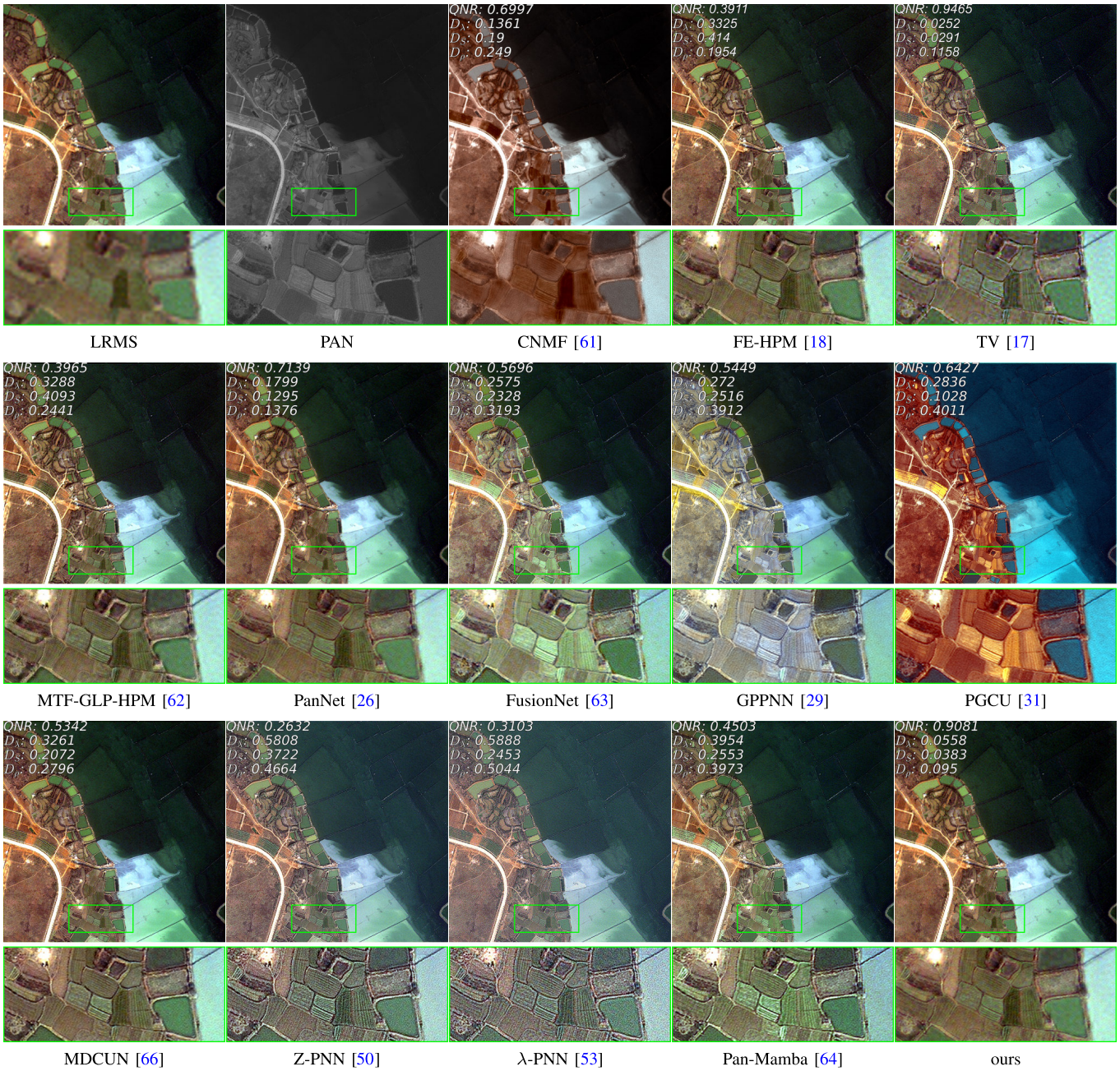

Fig. 11. Visual comparison of results generated by representative methods to test the generalization capability on the IKONOS satellite at full resolution. The DL-based models are trained using WorldView-3 data. The bright rectangular box represents the area zoomed in for display.

of scattering count N, and conduct ablation studies to validate the necessity of components.

A. Reduced-Resolution Assessment

Table II presents the quantitative evaluation results for all methods on four satellite datasets at reduced resolution. Our PEMAE model stands out among all DL-based models, which fully demonstrates the superiority of the method proposed in this article. Among other models, UTSN [69], PanNet [26], and Pan-Mamba [64] also achieve comparable results. Traditional methods are far behind most DL-based models in terms of reduced-resolution indicators. This finding demonstrates the effectiveness of data-driven methods in learning the data distribution and leveraging their powerful feature extraction and fusion capabilities to achieve favorable outcomes. On the WorldView-2, WorldView-3, IKONOS, and GaoFen-1 datasets, PEMAE outperforms the second-best model by 1.31, 1.03, 1.34, and 0.52 dB in PSNR, respectively. The improvements demonstrate that PEMAE can effectively fit data from various satellites, further validating the effectiveness of the proposed method. Figs. 5 and 6 illustrate the visual comparison results of representative methods on WorldView-2 and WorldView-3, respectively. Visually, all methods exhibit excellent fusion performance. However, according to the error maps, PEMAE exhibits lower reconstruction errors.

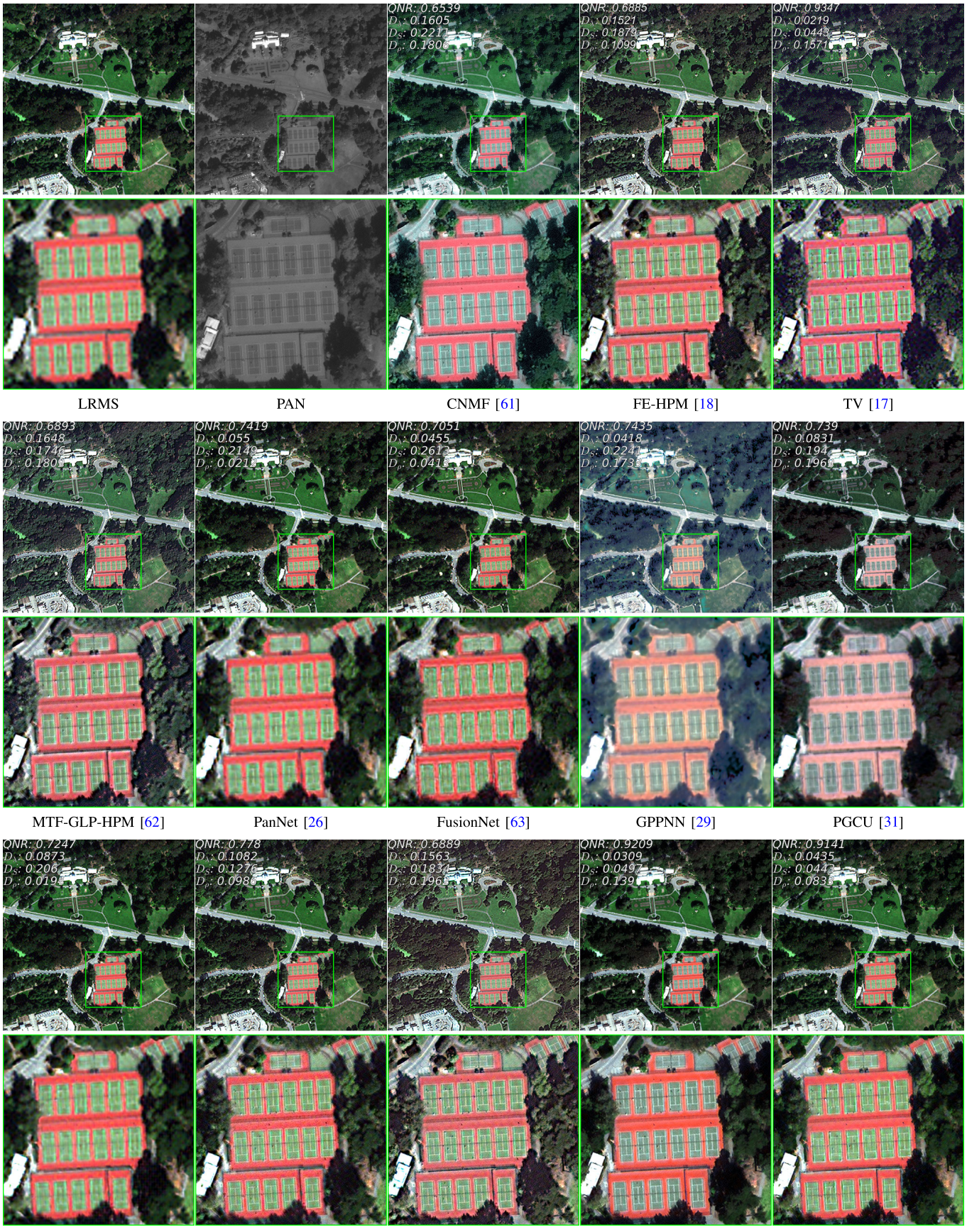

Fig. 12. Visual comparison of results generated by representative methods to test the generalization capability on the WorldView-2 satellite at full resolution. The DL-based models are trained using GaoFen-1 data. The bright rectangular box represents the area zoomed in for display.

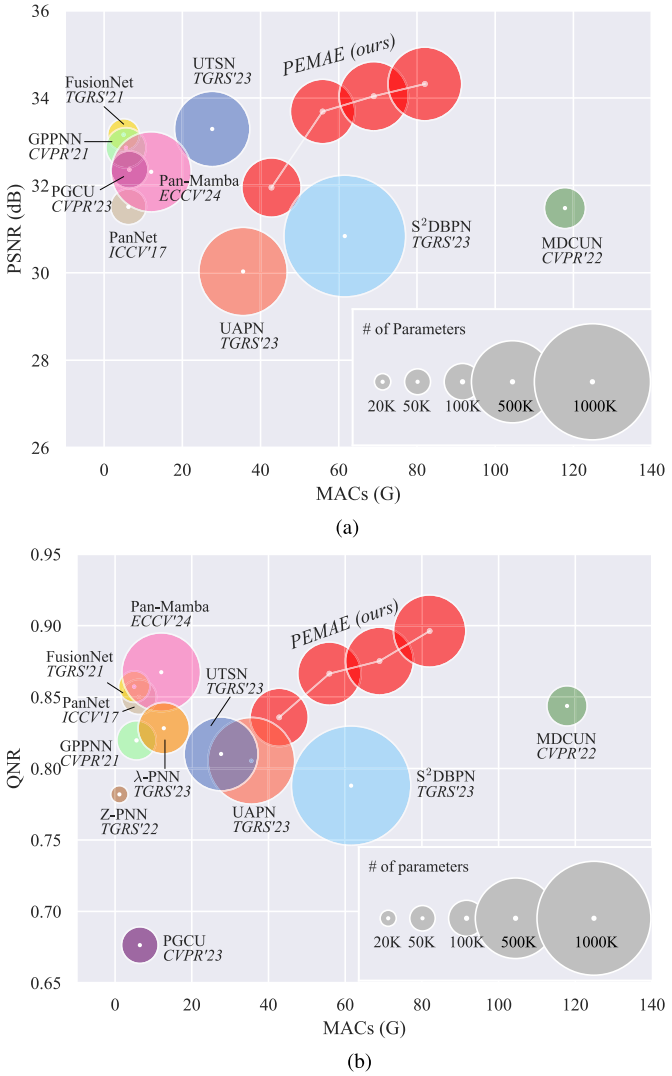

Fig. 13. Comparisons on model volume and performance. The computation of MACs was obtained under the condition of predicting images with four bands 256 × 256 resolution and a batch of 1. Note that original settings of S^{2}DBPN [67] encompass ∼16 M parameters, which is not fair for comparisons because all comparable models in this article have <1 M parameters. Thus, we adjusted the channels of the intermediate layers within S^{2}DBPN [67], reducing its parameter count to ∼1 M. (a) PSNR versus number of parameters versus MACs. (b) QNR versus number of parameters versus MACs.

Table III presents the generalization testing results across different satellite datasets at reduced resolution. PEMAE demonstrates superior performance on the majority of the indicators, followed by PanNet [26] and Pan-Mamba [64]. In contrast, UTSN [69], which performs well on the trained satellite, achieves poorer results on new data, indicating its limited generalization capability. Although the model proposed in this article did not achieve the best results when generalizing from WorldView-3 to IKONOS and from IKONOS to GaoFen-1, it still outperforms most models, such as UTSN [69], GPPNN [29], and S^{2}DBPN [67]. Figs. 7 and 8 show the visual comparison results of representative methods generalizing from WorldView-2 to WorldView-3 and from WorldView-3 to IKONOS, respectively. It can be observed that when generalizing from WorldView-2 to WorldView-3, due to the similar spectral characteristics between these two satellites, models trained on WorldView-2 adapt well to the data distribution of WorldView-3, with comparable visual effects for all models. However, when generalizing from WorldView-3 to IKONOS, where the spectral characteristics change significantly, models, such as FusionNet [63], GPPNN [29], and PGCU [31], exhibit substantial errors, indicating their lack of generalization ability. PEMAE maintains its transferability well in the face of significant changes in data distribution, which underscores the strong generalization capability of the method proposed in this article.

B. Full-Resolution Assessment

The experiments conducted at reduced resolutions primarily assess the method's performance on simulated datasets, yet the method's applicability to real-world data remains to be validated. To comprehensively demonstrate the effectiveness of our approach, this section presents quantitative and qualitative comparisons using full-resolution satellite imagery.

Table IV presents the quantitative results at full-resolution evaluation. Table V assesses the generalization performance at full resolution. For DL-based models, despite achieving overwhelming results over traditional algorithms at reduced resolution as shown in Tables II and III, models, such as PGCU [31], only achieve comparable or even worse results at full resolution. Conversely, traditional models, such as TV [17] and FE-HPM [18], significantly outperform DL-based models on some metrics such as QNR and D_S on the WorldView-2 dataset. Despite the unsupervised methods, Z-PNN and λ-PNN being trained directly at full resolution, they exhibit suboptimal performance in spectral preservation but achieve relatively better results in spatial preservation. PEMAE maintains satisfactory results on most satellites, which indicates that the proposed model trained with the observational model presented in this article exhibits excellent generalization capabilities. However, there exists a slight deficiency on the Gaofen-1 satellite. We analyze causes as follows. The Gaofen-1 dataset [6] primarily consists of land covers characterized by low-frequency information, such as oceans, lakes, and agricultural lands. Previous research [75], [76] has demonstrated that neural networks exhibit a prior for reconstructing the low-frequency components. Consequently, it may result in overfitting when training PEMAE using GaoFen-1 dataset because of the data characteristics. Thus, PEMAE yields superior performance at reduced resolution but suboptimal results at full resolution and in cross-satellite assessments. In future research, we will further explore the relationship between dataset distribution and model complexity.

Figs. 9 and 10 display the visual comparison results on the WorldView-3 and GaoFen-1 datasets, respectively. Figs. 11 and 12 show the visual comparison results when generalizing from WorldView-3 to IKONOS and from GaoFen-1 to WorldView-2, respectively. Clearly, PEMAE is capable of preserving more spatial details while avoiding spectral distortion. Unsupervised learning methods, such as Z-PNN [50] and λ-PNN [53], perform well in preserving spatial details but are inferior in spectral preservation. This may be attributed

TABLE VI

ABLATION RESULTS OF INPUTS FOR ENCODER AND DECODER AT REDUCED RESOLUTION. COLOR CONVENTION: THE BEST AND SECOND BEST AMONG ALL ALGORITHMS. ↓: LOWER IS BETTER. ↑: HIGHER IS BETTER

PAN↓	LRMS↑	WorldView-2				WorldView-3				IKONOS				GaoFen-1			
		MAE↓	PSNR↑	CC↑	ERGAS↓	MAE↓	PSNR↑	CC↑	ERGAS↓	MAE↓	PSNR↑	CC↑	ERGAS↓	MAE↓	PSNR↑	CC↑	ERGAS↓
✗	✗	0.0142	33.72	0.9254	4.726	0.0142	34.02	0.9565	3.659	0.0106	36.00	0.9183	2.335	0.0158	34.16	0.8283	2.545
✗	✓	0.0133	33.91	0.9269	4.612	0.0135	34.21	0.9589	3.549	0.0152	34.28	0.8327	2.780	0.0237	31.39	0.6448	3.467
✓	✗	0.0140	33.86	0.9279	4.655	0.0141	34.05	0.9563	3.639	0.0103	36.17	0.9242	2.304	0.0146	34.54	0.8460	2.388
✓	✓	0.0130	34.11	0.9301	4.531	0.0134	34.32	0.9598	3.510	0.0154	34.11	0.8350	2.865	0.0214	32.03	0.6829	3.147

TABLE VII

ABLATION RESULTS OF INPUTS FOR ENCODER AND DECODER AT FULL RESOLUTION. COLOR CONVENTION: THE BEST AND SECOND BEST AMONG ALL ALGORITHMS. ↓: LOWER IS BETTER. ↑: HIGHER IS BETTER

PAN↓	LRMS↑	WorldView-2				WorldView-3				IKONOS				GaoFen-1			
		QNR ↑	D_λ ↓	D_S ↓	D_ρ ↓	QNR ↑	D_λ ↓	D_S ↓	D_ρ ↓	QNR ↑	D_λ ↓	D_S ↓	D_ρ ↓	QNR ↑	D_λ ↓	D_S ↓	D_ρ ↓
✗	✗	0.8826	0.0455	0.0755	0.0953	0.8727	0.0416	0.0890	0.1499	0.7500	0.1606	0.1100	0.1589	0.6261	0.2146	0.2098	0.1799
✗	✓	0.8985	0.0542	0.0505	0.1281	0.9044	0.0411	0.0576	0.2020	0.6265	0.2307	0.2046	0.3077	0.4412	0.2494	0.4354	0.5019
✓	✗	0.8988	0.0389	0.0651	0.0844	0.9016	0.0319	0.0686	0.1363	0.7708	0.1298	0.1168	0.1406	0.6441	0.1840	0.2165	0.1519
✓	✓	0.8889	0.0524	0.0626	0.1358	0.8962	0.0375	0.0694	0.2006	0.7797	0.1267	0.1115	0.1350	0.4447	0.2328	0.4386	0.4684

to the specially designed loss function that maintains spatial structure. Among the compared traditional methods, TV [17] is the most competitive. PEMAE can produce visual effects comparable to or better than those of TV [17], while other models trained under Wald's protocol [7] struggle to achieve comparable results. This further illustrates that the method proposed in this article breaks the limitations of Wald's protocol [7], exhibiting excellent performance on both full-resolution and cross-satellite generalization capabilities.

C. Comparison of Model Size

Fig. 13(a) and (b) illustrates the tradeoffs between the number of layers L in PEMAE ranging from 1 to 4 and the corresponding parameter numbers and multiply–accumulate operations (MACs) [77], [78] against the values of PSNR and QNR on the WorldView-3 dataset. In terms of model parameters, PanNet [26], FusionNet [63], PGCU [31], and Z-PNN [50] are more lightweight, with FusionNet [63] achieving commendable results. The S^{2}DBPN model possesses a larger parameter volume, while MDCUN [66] exhibits higher computational costs. The contrasting performance of the PGCU [31] model in terms of PSNR and QNR indicates a deficiency in its generalization capability. PEMAE's parameter count and computational load are at a moderate level among all models, yet its PSNR and QNR values are in the upper echelons. With an equivalent or even lesser amount of parameters, our model outperforms Pan-Mamba [64], S^{2}DBPN [67], and UAPN [68], which further demonstrates the superiority of the method proposed in this article.

D. Ablation Study

This section conducts ablation studies on the model's inputs and components to verify the necessity of the modifications.

1) Ablation of Model Input: Tables VI and VII present the results of combination experiments for the optional LRMS↑ and PAN↓ as depicted in Fig. 3(a) at reduced and full resolutions, respectively. We applied upsampling to LRMS and downsampling to PAN to ensure compatibility with the model's required input size. The network is trained using the WorldView-3 dataset and generalized to other sensors. The results indicate that when both LRMS↑ and PAN↓ are utilized, the best results are achieved at reduced resolution. This may be attributed to PAN↓ providing additional spatial details for the masked LRMS, while LRMS↑ preserves complete spectral information. The model can effectively integrate the information provided by both to better fit the data. However, it is surprising that when both LRMS↑ and PAN↓ are used, the best results are not obtained at full-resolution evaluation. The optimal performance is achieved

TABLE VIII

REDUCED-RESOLUTION ABLATION RESULTS OF MODEL COMPONENTS. COLOR CONVENTION: IMPROVED AND DETERIORATED BOLD INDICATES THE BEST RESULTS. ↓: LOWER IS BETTER. ↑: HIGHER IS BETTER

Method	Metric							
	MAE↓	Δ_{MAE}	PSNR↑	Δ_{PSNR}	CC↑	Δ_{CC}	ERGAS↓	Δ_{ERGAS}
Baseline	**0.0134**	–	**34.3217**	–	**0.9598**	–	**3.5103**	–
LeakyReLU → ReLU	0.0153	+0.0019	32.9863	−1.3355	0.9445	−0.0153	4.1061	+0.5958
DCN → Conv	0.0151	+0.0017	33.3343	−0.9874	0.9488	−0.0110	3.9333	+0.4230
− Cross Attention	0.0154	+0.0020	33.0919	−1.2298	0.9452	−0.0146	4.0152	+0.5049
+ Positional Embedding	0.0134	+0.0001	34.3209	−0.0009	0.9594	−0.0004	3.5161	+0.0058

TABLE IX

FULL-RESOLUTION ABLATION RESULTS OF MODEL COMPONENTS. COLOR CONVENTION: IMPROVED AND DETERIORATED **BOLD** INDICATES THE BEST RESULTS. ↓: LOWER IS BETTER. ↑: HIGHER IS BETTER

Method	Metric							
	QNR↑	Δ_{QNR}	D_λ ↓	Δ_{D_λ}	D_S ↓	Δ_{D_S}	D_ρ ↓	Δ_{D_ρ}
Baseline	0.8962	–	**0.0375**	–	0.0694	–	0.2006	–
LeakyReLU → ReLU	0.8630	−0.0332	0.0579	+0.0204	0.0853	+0.0159	0.2230	+0.0225
DCN → Conv	0.8666	−0.0297	0.0627	+0.0252	0.0769	+0.0075	0.2534	+0.0529
− Cross Attention	0.8596	−0.0366	0.0630	+0.0255	0.0853	+0.0159	0.2464	+0.0458
+ Positional Embedding	**0.8966**	+0.0003	0.0379	+0.0004	**0.0691**	-0.0003	**0.1979**	−0.0027

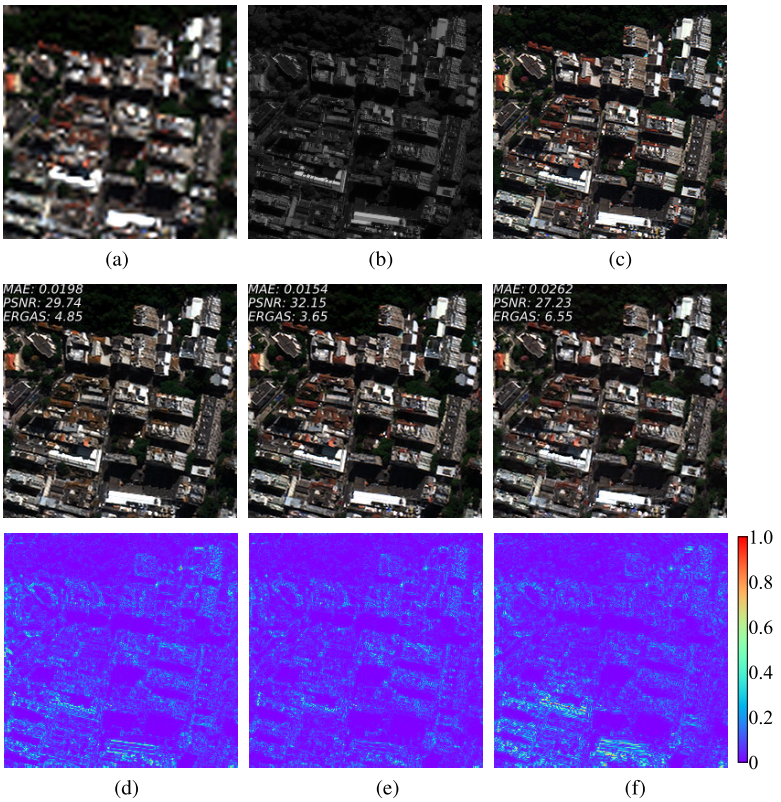

Fig. 14. Visual comparison of model performance under different ensemble numbers N. (a) LRMS. (b) PAN. (c) HRMS (GT). (d) $N = 1$. (e) $N = 4$. (f) $N = 8$.

TABLE X

PARAMETER ANALYSIS OF FOCUSED FACTOR p. BOLD INDICATES THE BEST RESULTS. ↓: LOWER IS BETTER. ↑: HIGHER IS BETTER

p	Metric							
	MAE↓	PSNR↑	CC↑	ERGAS↓	QNR ↑	D_λ ↓	D_S ↓	D_ρ ↓
2	0.01350	34.23	0.9592	3.544	**0.9068**	0.0383	0.0579	0.1985
4	0.01347	34.25	0.9588	3.541	0.8995	0.0401	0.0637	0.1929
6	**0.01336**	**34.32**	**0.9598**	**3.510**	0.8962	**0.0375**	0.0694	0.2006
8	0.01345	34.30	0.9591	3.515	0.8994	0.0396	0.0643	**0.1920**
10	0.01343	34.29	0.9595	3.534	0.9049	0.0388	**0.0525**	0.1973
16	0.01353	34.22	0.9587	3.547	0.9009	0.0392	0.0632	0.1952
32	0.01353	34.25	0.9588	3.510	0.8738	0.0520	0.0797	0.2301

TABLE XI

PARAMETER ANALYSIS OF SCATTERING NUMBER N. BOLD INDICATES THE BEST RESULTS. ↓: LOWER IS BETTER. ↑: HIGHER IS BETTER

N	Metric							
	MAE↓	PSNR↑	CC↑	ERGAS↓	QNR ↑	D_λ ↓	D_S ↓	D_ρ ↓
w/o scatter	0.01450	33.59	0.9520	3.797	0.8461	0.0847	0.0780	**0.1531**
w/o ensemble	0.01464	33.60	0.9516	3.812	0.8715	0.0489	0.0852	0.2280
2	0.01416	33.79	0.9531	3.709	**0.9095**	0.0385	**0.0548**	0.1876
4	**0.01336**	**34.32**	**0.9598**	**3.510**	0.8962	**0.0375**	0.0694	0.2006
6	0.01342	34.29	0.9595	3.520	0.8920	0.0405	0.0710	0.2015
8	0.01443	33.54	0.9534	3.817	0.8612	0.0677	0.0777	0.2148
10	0.01515	33.16	0.9471	3.971	0.8325	0.0850	0.0918	0.2246

when PAN↓ is used solely as an additional encoder input, without LRMS↑ as an additional decoder input. This discrepancy may be attributed to the decoder being more heavy, and when LRMS↑ is used as decoder input, the model may directly ignore the features extracted from unmasked features. In this case, the proposed mask-scatter-ensemble mechanism does not function at all, which degrades the model to one trained under Wald's protocol [7]. Through this experiment, we validated the effectiveness of the proposed observation model and the carefully designed reconstruction approach.

2) Ablation of Model Component: Tables VIII and IX present the ablation studies for components of the model at reduced and full resolutions, respectively. The following conclusions can be drawn.

1) In the linear cross-attention mechanism introduced in this article, LeakyReLU is superior to ReLU. This is because LeakyReLU does not produce overly sparse results, thereby ensuring the diversity of features.
2) The use of DCN [60] is more advantageous than conventional convolutions, as DCN [60] can dynamically capture the most relevant local features, which complements the random masking mechanism proposed in this article, allowing for more effective reconstruction of the image from unmasked features.
3) The cross-attention mechanism is effective compared to using only self-attention mechanisms, which effectively fuses information from two modalities.
4) Positional embedding is negligible for the pansharpening task, as we observed that the addition of positional embedding does not significantly alter the model's performance.

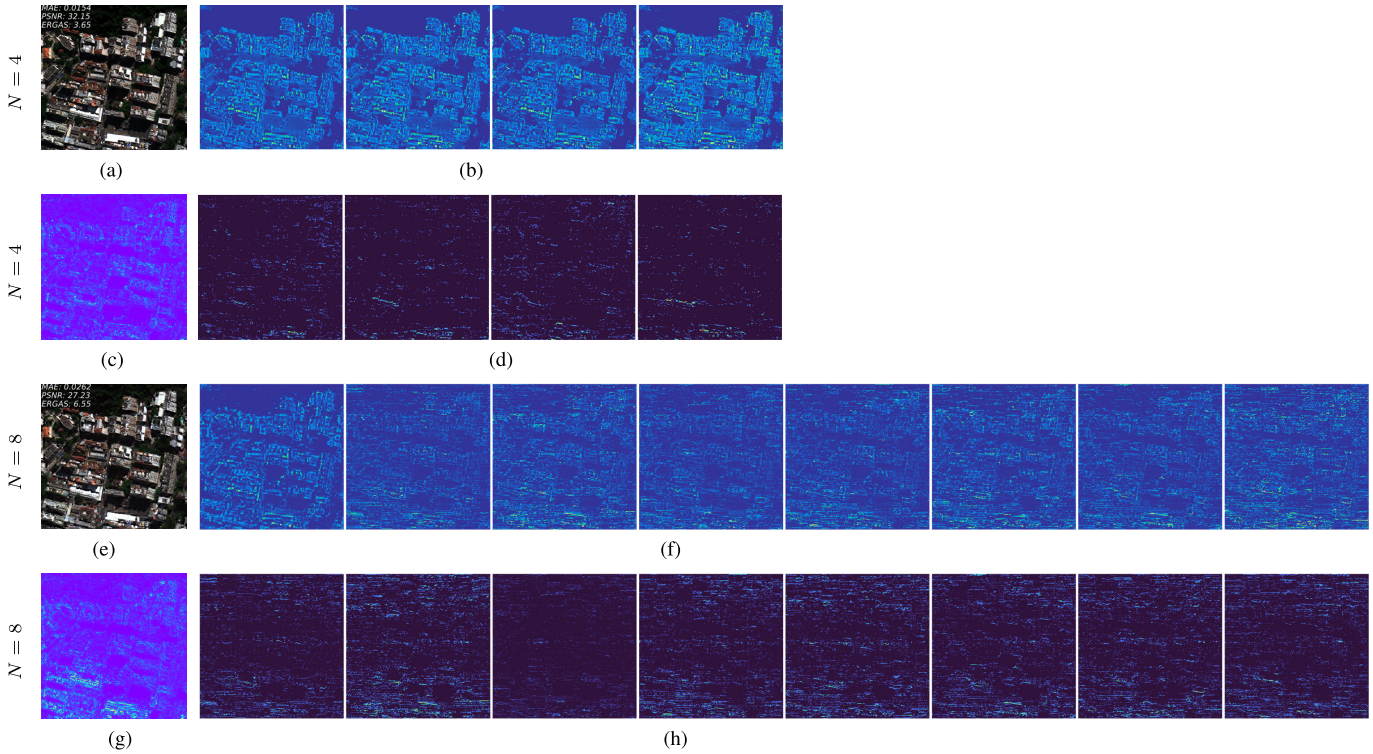

Fig. 15. Visualization of decoded features under the ensemble number of 4 and 8. Clearly, when $N = 4$, the decoded features exhibit subtle differences in spatial details. However, at $N = 8$, the features become chaotic, with only the first feature maintaining completeness. The increase in ensembles has introduced excessive randomness, impeding the model's capacity to decode the features accurately. (a) Prediction. (b) $N = 4$ decoded features. (c) Error. (d) $N = 4$ difference between adjacent features. (e) Prediction. (f) $N = 8$ decoded features. (g) Error. (h) $N = 8$ difference between adjacent features.

E. Parameter Analysis

We examine the effects of the two tunable hyperparameters, including the focused factor p and the number of ensembles N in the proposed model. When p varies, N is fixed to 4, and the number of layers L is fixed to 4. When N varies, p is fixed at 6, and the number of layers L remains constant at 4.

1) Analysis of Focused Factor p: The focused factor p controls to what extent the features are nonlinearly stretched. With an appropriate p, the model practically achieves a more pronounced distinction between similar and dissimilar query–key pairs, restoring the sharp attention distribution akin to the original Softmax$(\cdot)$. Table X illustrates the impact of varying p from 2 to 32 on the model performances. The metrics show no significant changes, indicating that the hyperparameter p of the focused linear attention is robust for the pansharpening task.

2) Analysis of Ensembling Count N: Table XI illustrates the impact of varying the number of ensembles N from 2 to 10 on model performance. To further investigate the functionality of scattering and ensembling, we also conducted experiments without scattering [omitting *fused tokens* in Fig. 3(a)] and without ensemble (setting $N = 1$). The results indicate that the optimal performance is achieved when N is 2 or 4. Conversely, the least favorable outcomes occur when scattering or ensembling is not enabled. This suggests that the mask-scatter-ensemble mechanism proposed in this article effectively enhances the model performance when N is appropriately set.

Surprisingly, a larger N is not always better. An increase in N from 6 to 10 does not enhance the model performance but rather results in a degradation. Fig. 14 presents the visualizations of prediction on the WorldView-3 dataset for $N = \{1, 4, 8\}$. The minimum reconstruction error is observed at $N = 4$, with a noticeable increase as N reaches 8. To elucidate the cause of this phenomenon, we analyzed the decoder outputs for $N = 4$ and $N = 8$, as depicted in Fig. 15. The features decoded at $N = 4$ are presented in Fig. 15(b), with the differences between adjacent feature maps illustrated in Fig. 15(d) to emphasize the distinctions between features. For $N = 8$, the corresponding decoded features and adjacent feature differences are displayed in Fig. 15(f) and (h). Although the features at $N = 4$ are not markedly distinct visually, their differences are closely tied to the error maps, suggesting that diverse scattering schemes reconstruct distinct spatial details, thereby enabling accurate HRMS estimation through the ensembling process. This finding corroborates the effectiveness of the mask-scatter-ensemble mechanism proposed in this article. In contrast, at $N = 8$, the excessive number of integrations results in significant discrepancies in decoded features, as evident in Fig. 15(f), where only the first feature remains effective. This observation suggests that the model struggles to reconcile the excessive random embeddings, losing sight of the essential patterns required for precise reconstruction from masked images, which consequently degrades the model's performance.

VI. Conclusion and Discussion

In this article, we proposed a PEMAE for multispectral pansharpening. Our approach leverages masked autoencoder

to achieve superior performance in both quantitative and qualitative evaluations. Key contributions of this article include the novel observation model that considers LRMS as the result of pixel-wise masking, the scattering mechanism that captures unique spatial information, and the ensembling of multiple scatter schemes to obtain more accurate and robust HRMS. We have also introduced an efficient computational strategy by introducing a modified linear cross-attention, significantly reducing computational complexity. Extensive experiments demonstrated that PEMAE outperforms SOTA methods in terms of reconstruction accuracy and generalization ability. In addition, the parameter analysis in this article provided insights into the influence of different parameters and configurations on the performance of PEMAE.

Future research directions can focus on further exploring the potential of PEMAE. For instance, investigating alternative scattering mechanisms and exploring the use of different loss functions may offer interesting avenues for improving the performance even further. Moreover, integrating other advanced DL techniques and architectures can be explored to enhance the overall capabilities of the proposed method. On the other hand, the ensemble strategy, which can lead to an increase in computational costs, may not be the optimal choice. Paradoxically, a large number of ensembles may lead to inferior performance. Future research could delve into alternative approaches to ensemble strategy, e.g., incorporating uncertainty estimations for pixels at different spatial positions.

REFERENCES

[1] Y. Zhang, "Understanding image fusion," *Photogramm. Eng. Remote Sens.*, vol. 70, pp. 657–661, Jun. 2004.

[2] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15979–15988.

[3] H. Zhou, Q. Liu, and Y. Wang, "PGMAN: An unsupervised generative multiadversarial network for pansharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6316–6327, 2021.

[4] C. Thomas, T. Ranchin, L. Wald, and J. Chanussot, "Synthesis of multispectral images to high spatial resolution: A critical review of fusion methods based on remote sensing Physics," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1301–1312, May 2008.

[5] 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.

[6] 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.

[7] 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, pp. 691–699, Jun. 1997.

[8] 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.

[9] S. Luo, S. Zhou, Y. Feng, and J. Xie, "Pansharpening via unsupervised convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4295–4310, 2020.

[10] S. Rahmani, M. Strait, D. Merkurjev, M. Moeller, and T. Wittman, "An adaptive IHS pan-sharpening method," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 4, pp. 746–750, Oct. 2010.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "MTF-tailored multiscale fusion of high-resolution MS and pan imagery," *Photogrammetric Eng. Remote Sens.*, vol. 72, no. 5, pp. 591–596, May 2006.

[16] 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.

[17] 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.

[18] G. Vivone et al., "Pansharpening based on semiblind deconvolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 1997–2010, Apr. 2015.

[19] W. Sun, J. Zhou, X. Meng, G. Yang, K. Ren, and J. Peng, "Coupled temporal variation information estimation and resolution enhancement for remote sensing spatial–temporal–spectral fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5532118.

[20] W. Han et al., "Geological remote sensing interpretation using deep learning feature and an adaptive multisource data fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4510314.

[21] R. Fan, R. Feng, W. Han, and L. Wang, "Urban functional zone mapping with a bibranch neural network via fusing remote sensing and social sensing data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 11737–11749, 2021.

[22] Y. Lu et al., "Remote-sensing interpretation for soil elements using adaptive feature fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4505515.

[23] Y. Zhang, P. Liu, L. Chen, M. Xu, X. Guo, and L. Zhao, "A new multi-source remote sensing image sample dataset with high resolution for flood area extraction: GF-FloodNet," *Int. J. Digit. Earth*, vol. 16, no. 1, pp. 2522–2554, Oct. 2023.

[24] P. Liu, L. Wang, R. Ranjan, G. He, and L. Zhao, "A survey on active deep learning: From model driven to data driven," *ACM Comput. Surveys*, vol. 54, no. 10s, pp. 1–34, Sep. 2022.

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

[26] 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.

[27] 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.

[28] X. Wu, T.-Z. Huang, L.-J. Deng, and T.-J. Zhang, "Dynamic cross feature fusion for remote sensing pansharpening," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14667–14676.

[29] 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.

[30] W. G. C. Bandara, J. M. J. Valanarasu, and V. M. Patel, "Hyperspectral pansharpening based on improved deep image prior and residual reconstruction," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5520816.

[31] Z. Zhu, X. Cao, M. Zhou, J. Huang, and D. Meng, "Probability-based global cross-modal upsampling for pansharpening," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14039–14048.

[32] Y. Duan, X. Wu, H. Deng, and L.-J. Deng, "Content-adaptive non-local convolution for remote sensing pansharpening," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2024, pp. 27738–27747.

[33] R. Ran, L.-J. Deng, T.-J. Zhang, J. Chang, X. Wu, and Q. Tian, "KNLConv: Kernel-space non-local convolution for hyperspectral image super-resolution," *IEEE Trans. Multimedia*, vol. 26, pp. 8836–8848, 2024.

[34] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inform. Process. Syst.*, vol. 30, 2017, pp. 1–24.

[35] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–29.

[36] X. Su, J. Li, and Z. Hua, "Transformer-based regression network for pansharpening remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5407423.
[37] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
[38] W. G. C. Bandara and V. M. Patel, "HyperTransformer: A textural and spectral feature fusion transformer for pansharpening," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1757–1767.
[39] 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.
[40] 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.*, vol. 36, 2022, pp. 3553–3561.
[41] I. Goodfellow et al., "Generative adversarial networks," in *Proc. Adv. Neural Inform. Process. Syst.*, vol. 3, 2014, pp. 1–26.
[42] Y. Cui, P. Liu, B. Song, L. Zhao, Y. Ma, and L. Chen, "Reconstruction of large-scale missing data in remote sensing images using extend-GAN," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.
[43] B. Song et al., "MLFF-GAN: A multilevel feature fusion with GAN for spatiotemporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4410816.
[44] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. NIPS*, vol. 33, 2020, pp. 6840–6851.
[45] A. Gastineau, J.-F. Aujol, Y. Berthoumieu, and C. Germain, "Generative adversarial network for pansharpening with spectral and spatial discriminators," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4401611.
[46] W. Xie, Y. Cui, Y. Li, J. Lei, Q. Du, and J. Li, "HPGAN: Hyperspectral pansharpening using 3-D generative adversarial networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 463–477, Jan. 2021.
[47] 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.
[48] Q. Meng, W. Shi, S. Li, and L. Zhang, "PanDiff: A novel pansharpening method based on denoising diffusion probabilistic model," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5611317.
[49] Z. Cao, S. Cao, X. Wu, J. Hou, R. Ran, and L.-J. Deng, "DDRF: Denoising diffusion model for remote sensing image fusion," 2023, *arXiv:2304.04774*.
[50] M. Ciotola, S. Vitale, A. Mazza, G. Poggi, and G. Scarpa, "Pansharpening by convolutional neural networks in the full resolution framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5408717.
[51] J. Li et al., "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 112, Aug. 2022, Art. no. 102926.
[52] X. Rui, X. Cao, L. Pang, Z. Zhu, Z. Yue, and D. Meng, "Unsupervised hyperspectral pansharpening via low-rank diffusion model," 2023, *arXiv:2305.10925*.
[53] M. Ciotola, G. Poggi, and G. Scarpa, "Unsupervised deep learning-based pansharpening with jointly enhanced spectral and spatial fidelity," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5405417.
[54] W. Sun, K. Ren, X. Meng, G. Yang, J. Peng, and J. Li, "Unsupervised 3-D tensor subspace decomposition network for spatial–temporal–spectral fusion of hyperspectral and multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5528917.
[55] D. Wang, P. Zhang, Y. Bai, and Y. Li, "MetaPan: Unsupervised adaptation with meta-learning for multispectral pansharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[56] Y. Qu, R. K. Baghbaderani, H. Qi, and C. Kwan, "Unsupervised pansharpening based on self-attention mechanism," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3192–3208, Apr. 2021.
[57] P. Liu, J. Li, L. Wang, and G. He, "Remote sensing data fusion with generative adversarial networks: State-of-the-art methods and future research directions," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 295–328, Jun. 2022.
[58] Q. Xie, M. Zhou, Q. Zhao, D. Meng, W. Zuo, and Z. Xu, "Multispectral and hyperspectral image fusion by MS/HS fusion net," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1585–1594.
[59] D. Han, X. Pan, Y. Han, S. Song, and G. Huang, "FLatten transformer: Vision transformer using focused linear attention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 1–24.
[60] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets V2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9300–9308.
[61] 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.
[62] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "An MTF-based spectral distortion minimizing model for pan-sharpening of very high resolution multispectral images of urban areas," in *Proc. Joint Workshop Remote Sens. Data Fusion Over Urban Areas*, 2003, pp. 90–94.
[63] 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.
[64] X. He et al., "Pan-mamba: Effective pan-sharpening with state space model," 2024, *arXiv:2402.12192*.
[65] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–11.
[66] 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.
[67] K. Zhang, A. Wang, F. Zhang, W. Wan, J. Sun, and L. Bruzzone, "Spatial-spectral dual back-projection network for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5402216.
[68] K. Zheng, J. Huang, M. Zhou, D. Hong, and F. Zhao, "Deep adaptive pansharpening via uncertainty-aware image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5403715.
[69] Z. Sheng, F. Zhang, J. Sun, Y. Tan, K. Zhang, and L. Bruzzone, "A unified two-stage spatial and spectral network with few-shot learning for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5403517.
[70] A. R. Gillespie, A. B. Kahle, and R. E. Walker, "Color enhancement of highly correlated images. II. Channel ratio and 'chromaticit' transformation techniques," *Remote Sens. Environ.*, vol. 22, no. 3, pp. 343–365, Aug. 1987.
[71] 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.
[72] 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.
[73] 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.
[74] G. Scarpa and M. Ciotola, "Full-resolution quality assessment for pansharpening," *Remote Sens.*, vol. 14, no. 8, p. 1808, Apr. 2022.
[75] V. Lempitsky, A. Vedaldi, and D. Ulyanov, "Deep image prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9446–9454.
[76] A. Qayyum, I. Ilahi, F. Shamshad, F. Boussaid, M. Bennamoun, and J. Qadir, "Untrained neural network priors for inverse imaging problems: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 6511–6536, May 2023.
[77] G. Frantz, "Digital signal processor trends," *IEEE Micro*, vol. 20, no. 6, pp. 52–59, May 2000.
[78] S. W. Smith, "The scientist and engineer's guide to digital signal processing," California Tech. Publishing, CA, USA, 1997.

Yongchuan Cui (Graduate Student Member, IEEE) received the B.E. degree in data science and big data technology from China University of Geosciences, Wuhan, China, in 2023. He is currently pursuing the master's degree in signal and information processing with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

His research interests include image processing, deep learning, multimodal, unified foundation models for Earth vision, and geospatial artificial intelligence (AI).

Peng Liu (Member, IEEE) received the M.S. and Ph.D. degrees in signal processing from Chinese Academy of Sciences, Beijing, China, in 2004 and 2009, respectively.

From May 2012 to May 2013, he was with the Department of Electrical and Computer Engineering, George Washington University, Washington, DC, USA, as a Visiting Scholar. He is currently an Associate Professor at the Aerospace Information Research Institute, Chinese Academy of Sciences. His research is focused on big data, sparse representation, compressive sensing, deep learning, and their applications to remote sensing data processing.

Yan Ma (Member, IEEE) received the M.S. and Ph.D. degrees in signal and information processing from the University of Chinese Academy of Sciences, Beijing, China, in 2007 and 2013, respectively.

She is currently an Associate Professor at the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing.

Lajiao Chen is currently an Associate Professor at the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. Her research interests are focused on geographic information systems and remote sensing techniques and their application to water resource management.

Mengzhen Xu is currently a tenured Associate Professor and the Vice Head of the School of Civil Engineering, Tsinghua University, Beijing, China. Her research explores the interfaces between hydraulics, sediment, structures, and aquatic organisms, and their application in river uses and ecological restoration.

Xingyan Guo received the Ph.D. degree in physical geography from Chinese Academy of Sciences, Beijing, China, in September 2020.

She is currently a Post-Doctoral Researcher at Tsinghua University, Beijing. Her research interests primarily focus on fluvial geomorphology and morphodynamics in diverse environmental settings.