PanDiff: A Novel Pansharpening Method Based on Denoising Diffusion Probabilistic Model

Qingyan Meng, Wenxu Shi, Sijia Li, and Linlin Zhang

***Abstract*— Pansharpening is a crucial image processing technique for numerous remote sensing downstream tasks, aiming to recover high spatial resolution multispectral images by fusing high spatial resolution panchromatic (PAN) images and low spatial resolution multispectral (LRMS) images. Most current mainstream pansharpening fusion frameworks directly learn the mapping relationships from PAN and LRMS images to high-resolution multispectral (HRMS) images by extracting key features. However, we propose a novel pansharpening method based on the denoising diffusion probabilistic model (DDPM) called PanDiff, which learns the data distribution of the difference maps (DMs) between HRMS and interpolated MS (IMS) images from a new perspective. Specifically, PanDiff decomposes the complex fusion process of PAN and LRMS images into a multistep Markov process, and the U-Net is employed to reconstruct each step of the process from random Gaussian noise. Notably, the PAN and LRMS images serve as the injected conditions to guide the U-Net in PanDiff, rather than being the fusion objects as in other pansharpening methods. Furthermore, we propose a modal intercalibration module (MIM) to enhance the guidance effect of the PAN and LRMS images. The experiments are conducted on a freely available benchmark dataset, including GaoFen-2, QuickBird, and WorldView-3 images. The experimental results from the fusion and generalization tests effectively demonstrate the outstanding fusion performance and high robustness of PanDiff. The results of the proposed method performed on various scenes are shown. In addition, the ablation experiments confirm the rationale behind PanDiff's construction.**

***Index Terms*— Deep learning (DL), denoising diffusion probabilistic model (DDPM), image fusion, pansharpening, remote sensing.**

NOMENCLATURE

$\mathbf{P} \in \mathbb{R}^{H\times W}$	Panchromatic (PAN) image.
$\mathbf{MS} \in \mathbb{R}^{\frac{H}{r}\times\frac{W}{r}\times C}$	Multispectral image (LRMS).
$\widehat{\mathbf{MS}} \in \mathbb{R}^{H\times W\times C}$	Pansharpened image (HRMS).
$\widetilde{\mathbf{MS}} \in \mathbb{R}^{H\times W\times C}$	Interpolated multispectral image (IMS).
$\mathbf{GT} \in \mathbb{R}^{H\times W\times C}$	Ground truth or ideal HRMS.
$\Delta\mathbf{MS} \in \mathbb{R}^{H\times W\times C}$	Difference map (DM), $\Delta\mathbf{MS} = \mathbf{GT} - \widetilde{\mathbf{MS}}$.
r	Spatial resolution ratio.
$q(\cdot \mid \cdot)$	Forward (diffusion) step.
$p_\theta(\cdot \mid \cdot)$	Reverse (denoised) step with network θ.
t	Discrete timesteps t on the range of $[0,\ T]$, where T is the total number.
x_0	Prior distribution of data; in this article, it represents samples from $\mathbf{GT} - \widetilde{\mathbf{MS}}$.
x_T	Random noise after diffusion.
x_t	Diffused data (latent state) at step t.

Manuscript received 12 December 2022; revised 21 March 2023 and 29 April 2023; accepted 8 May 2023. Date of publication 25 May 2023; date of current version 7 June 2023. *(Corresponding author: Wenxu Shi.)*

Qingyan Meng and Linlin Zhang are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, also with the College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the Key Laboratory of Earth Observation of Hainan Province, Hainan Aerospace Information Research Institute, Sanya 572029, China (e-mail: mengqy@radi.ac.cn; zhangll@radi.ac.cn).

Wenxu Shi is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, and also with the College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: shiwenxu20@mails.ucas.ac.cn).

Sijia Li is 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 100049, China (e-mail: lisijia20@mails.ucas.ac.cn).

Digital Object Identifier 10.1109/TGRS.2023.3279864

I. INTRODUCTION

REMOTE sensing images with high spatial and spectral resolution are required in a wide variety of fields, ranging from scene classification [1], [2], semantic segmentation [3], [4], comprehensive land use mapping [5], urban fine 3-D reconstruction [6], [7], environmental monitoring [8], and urban planning [9]. Unfortunately, due to physical limitations in current sensor technology, it is challenging for remote sensing data acquired by a single satellite sensor to meet this high-quality standard. To address this issue, pansharpening algorithms have been developed to fuse high spatial resolution panchromatic (PAN) images, which possess high spatial resolution but lack spectral information, with low spatial resolution multispectral (LRMS) images that capture spectral information across multiple bands to generate high-resolution multispectral (HRMS) images. Thus, preprocessing techniques, including pansharpening, are required in the field of remote sensing.

Numerous pansharpening methods have been developed, which can be broadly categorized into four groups based on their underlying principles: 1) component substitution (CS)-based approaches; 2) multiresolution analysis (MRA)-based approaches; 3) variational optimization (VO)-based approaches; and 4) deep learning (DL)-based approaches. Categories 1)–3) represent the traditional pansharpening approaches.

The CS-based pansharpening approaches aim to inject spatial information by projecting LRMS images into a new feature space, stripping LRMS images of their structural components, and replacing them with the corresponding components of PAN images. The main methods include the principal component transform (PCA)-based method [10], [11], [12], the intensity-hue-saturation (IHS)-based method [13], and the Gram–Schmidt (GS) method [14], with distinctions lying in their projection rules. The CS-based approaches can

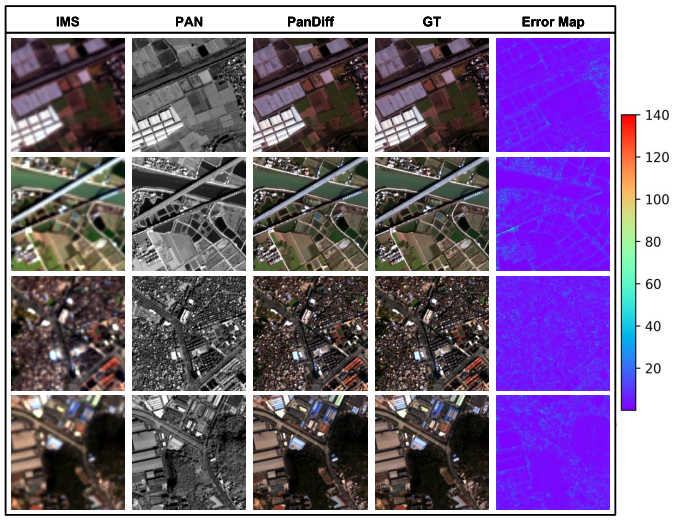

Fig. 1. Visual presentation based on the reduced-resolution images captured by the GaoFen-2 satellite, depicting various land covers. Line 1 illustrates farmland, Line 2 represents a mixed area of farmland and water, Line 3 portrays a complex urban area, and Line 4 showcases a mixed area of urban and forest land.

achieve superior spatial quality but are still subject to spectral distortion.

The MRA-based pansharpening approaches assume that the spatial disparity between LRMS and HRMS images is due to the absence of high-frequency spatial information and inject high-frequency information from PAN images via multiscale decomposition. The conventional multiscale decomposition algorithms include wavelet transform, pyramid Laplacian transform, and more. The typical MRA-based methods encompass modulation transfer function-generalized Laplacian pyramid (MTF-GLP) [15], [16], [17], àtrous wavelet transform (ATWT) [18], and proportional additive wavelet intensity method (AWLP) [19]. This category of approaches can produce HRMS with high spectral fidelity, but the spatial quality is relatively low.

The VO-based pansharpening approaches treat the pansharpening process as an ill-posed inverse problem and recast the problem as a VO problem with a hypothetical image connection model. Two stages comprise the VO-based methods: 1) constructing the objective function and 2) optimally solving the objective function, mainly based on iterative optimization algorithms [20]. In general, the objective function includes a fidelity term and a regularization term that contains prior knowledge. According to the classification of the objective function, the representative VO-based works include P + XS methods [21], Bayesian-based methods [22], [23], and sparse representation-based methods [24]. Although compared to the aforementioned two methods, VO-based approaches may effectively balance spatial augmentation and spectral preservation, they are computationally costly and may produce poor fusion outcomes if the underlying assumptions do not match the fusion situation.

Due to the outstanding feature extraction and aggregation capabilities of deep neural networks, DL-based approaches have become a significant research trend in recent years. Convolutional neural networks (CNNs) [25] have gained considerable attention for their excellent performance in pansharpening tasks. Various CNN-based methods have been proposed, including the first DL-based pansharpening method [26], the first full CNN-based PNN [27], PanNet [28], DRPNN [29], TFNet [30], and DSSN [31]. However, CNN-based methods can suffer from feature smoothing caused by vanilla convolution, leading to poor fusion results at the boundaries [32].

Recently, transformer models [33], [34] have gained popularity in pansharpening due to their improved capability for long-range modeling. Examples include the pure transformer-based method [35] and the transformer-based regression network [36]. It is worth mentioning that some methods integrate CNN and transformer for both local and global features extraction [37].

Generative models, such as generative adversarial networks (GANs) [38], have also been utilized for pansharpening, aiming to obtain higher quality HRMS images through the mutual adversarial of generators and discriminators [39], [40]. However, balancing these models remains a challenging task.

Unsupervised pansharpening fusion has also seen recent breakthroughs, with methods such as UP-SAM [41] using an unsupervised self-attention mechanism to explicitly extract details and inject them in a spatially varying manner. Furthermore, PanGAN [42] and UPanGAN [43] propose spectral and textural losses constrained GAN for unsupervised learning.

Overall, the field of DL-based pansharpening has been enriched by numerous contributions, including CNNs, transformers, GAN, invertible neural networks [44], and other methods that continue to advance the state of the art in pansharpening fusion.

Upon conducting an exhaustive review and analysis, we observed that: 1) traditional methods are constrained by the linear transformation process; 2) DL methods that rely on convolutional neural networks (CNNs) frequently grapple with feature smoothing or require spatial detail and gradient augmentation to mitigate this issue; and 3) GAN-based methods struggle with stable training. In addition, both conventional and DL-based techniques center on the standard pansharpening framework, which primarily entails the processing and fusion of PAN and LRMS images. These images facilitate the extraction of crucial features, which are subsequently utilized to learn the mapping from PAN and LRMS images to HRMS images. To surmount the inherent limitations of traditional and DL-based methods, we introduce a novel denoising diffusion probabilistic model (DDPM)-based pansharpening model, PanDiff. Our experimental results reveal that the proposed PanDiff and its distinct pansharpening fusion framework offer the following advantages: 1) exceptional pansharpening results with elevated metrics; 2) robust pansharpening performance with reduced variance; and 3) superior spatial details. However, PanDiff also confronts the challenge of low runtime efficiency, which necessitates further research. The main contributions of PanDiff are given as follows.

1) PanDiff is the first generative model based on the DDPM specifically designed for pansharpening applications.
2) PanDiff alters the learning objective of the traditional fusion networks. It decomposes the complex fusion

process of PAN and LRMS images into a multistep Markov process and primarily learns the data distribution of the difference map (DM) between HRMS and interpolated MS (IMS), rather than the spatial and spectral information of HRMS.

3) PanDiff deviates from the traditional approach of treating input PAN and MS as the primary objects for feature extraction. Instead, it injects the PAN and MS images, intercalibrated by a modal intercalibration module (MIM), as guiding conditions for the U-Net to learn the data distribution of the DM between HRMS and IMS.

The remainder of this article is organized as follows. Section II introduces the background of pansharpening and the basic principles of DDPM. Section III details our proposed PanDiff. Section IV provides the information on the experimental settings. In Section V, we compare PanDiff with state-of-the-art methods and present the results of the ablation experiments. Finally, in Section VI, we draw conclusions and outline future research directions.

II. Background and Preliminary

In this section, we review the basic theory of pansharpening and the fundamentals of DDPM. To facilitate reading and reduce ambiguity, we first define the common notation in the Nomenclature.

A. Pansharpening

The primary objective of pansharpening is to identify a stable mapping function $\mathcal{F}_\theta(\cdot)$ from LRMS and PAN to the ideal HRMS. Consequently, the general definition of pansharpening is given in the following equation:

$$\widehat{\mathbf{MS}} = \mathcal{F}_\theta(\mathbf{MS}, \mathbf{P}). \tag{1}$$

In order to solve this mapping function, many studies have developed classical algorithms from physical properties and algorithmic optimization, among which the CS- and MRA-based approaches are the most representative.

The CS-based approaches optimize (1) from the perspective of CS, and the schematic can be seen in Fig. 2(a). They obtain HRMS by replacing the spatial structure components of IMS ($\mathbf{IMS_L}$) with the spatial detail information of PAN and injecting gain functions with the following equation:

$$\widehat{\mathbf{MS}} = \widetilde{\mathbf{MS}} + G_\theta(\mathbf{P} - \mathbf{I_L}) \tag{2}$$

where $\mathbf{I_L}$ is generated through the linear combination of the LRMS spectral bands and $G_\theta(\cdot)$ is the injection gain function.

The MRA-based approaches still follow the same idea of injecting high spatial structure information from PAN to IMS, but the method and source of extracting information are different. This category of approaches obtains high-frequency information ($\mathbf{PAN_H}$) directly through PAN, as in the following equation:

$$\widehat{\mathbf{MS}} = \widetilde{\mathbf{MS}} + G_\theta(\mathbf{P} - \mathbf{P_L}) \tag{3}$$

where $\mathbf{P_L}$ is generated by employing low-pass filtering on PAN.

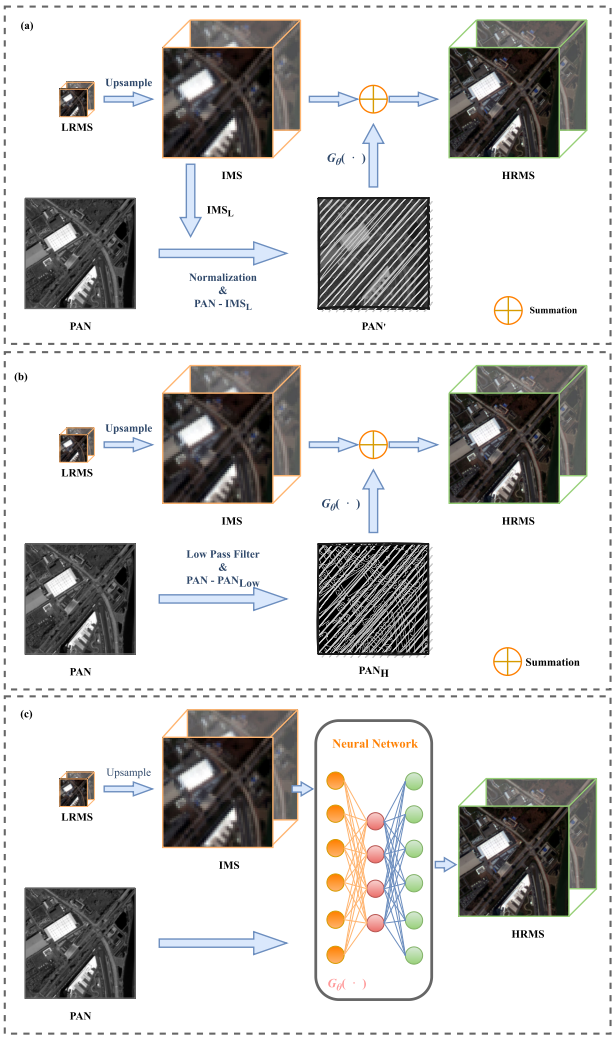

Fig. 2. Schematic of (a) CS-based approach, (b) MRA-based approach, and (c) DL-based approach in supervised fashion.

Unlike the two approaches described above, since the backpropagation algorithms [45] of DL-based approaches can effectively fit arbitrary functions, they often do not involve physically significant assumptions on pansharpening, allowing this family of approaches to solving for the mapping functions directly based on (1).

It is worth noting that both traditional and DL-based methods have their individual characteristics. The traditional methods essentially bridge the gap between IMS and HRMS by means of high spatial structure injection, while the DL-based methods directly model the relationships between the input LRMS and PAN and output HRMS due to the powerful function fitting capability of neural networks. However, both traditional and DL-based approaches try to extract the key features by directly performing complex transformations on PAN and LRMS and to find the better mapping function.

B. Denoising Diffusion Probabilistic Models

DDPMs are a kind of discrete latent variable model that is prevalent in many generative tasks, such as text-to-image generation [46], image-to-image translation [47], and image

editing [48]. It is parameterized by a finite T timesteps Markov chain and uses variational inference to gradually transform an isotropic Gaussian noise $x_T \sim \mathcal{N}(0, I)$ into a sample x_0 from the target distribution, which can be written as

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t) \tag{4}$$

where $x_1, \ldots, x_{T-1}$ are latent states of the same dimension as the prior data distribution $x_0 \sim q(x_0)$ and the starting state $p(x_T) = \mathcal{N}(x_T; 0, I)$.

DDPMs contain two stages: the forward diffusion process and the reverse denoised process.

1) Forward (Diffusion) Process: The forward diffusion process begins with the prior data distribution x_0. Then, an approximate standard normal distribution $x_T \sim \mathcal{N}(0, I)$ is obtained by continuously adding Gaussian noise to x_0 through the Markov chain process. The Gaussian transition in the forward diffusion process of any adjacent latent states on Markov chains is formulated as follows:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t} x_{t-1}, \sqrt{\beta_t} I\right) \tag{5}$$

where β_t represents the variance of the added Gaussian noise in the transition process from x_{t-1} to x_t and all the variance schedule $\beta_1, \ldots, \beta_T \in [0, 1)$.

The forward diffusion process is given by the approximate posterior $q(x_{1:T} \mid x_0)$ as the following equation:

$$q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q(x_t \mid x_{t-1}). \tag{6}$$

Substituting (5) into (6), after reparameterization [49], the data distribution $q(x_t)$ of the latent state x_t at any arbitrary timestep t can be derived based on x_0 and β_t. The following equation gives this derivation:

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}} x_0, \sqrt{1-\bar{\alpha}} I\right) \tag{7}$$

$$\bar{\alpha}_t = \prod_{i=1}^{t} (1 - \beta_i). \tag{8}$$

2) Reverse (Denoised) Process: The purpose of DDPM is to recreate a sample in the specific data distribution $q(x_0)$ from sampling the Gaussian noise x_T, which requires the reverse denoised process to learn the parameterized Gaussian transition $q(x_{t-1} \mid x_t)$. Note that if β_t is small enough, $q(x_{t-1} \mid x_t)$ will also be Gaussian [50]. However, it is hard to estimate $q(x_{t-1} \mid x_t)$, we have to use a model p_θ to approximate these conditional probabilities by fitting the mean and variance. The Gaussian transition in the reverse denoised process of any adjacent latent states on Markov chains is formulated as follows:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{9}$$

where $t \in [1, T]$, μ_θ and Σ_θ are the mean and variance of $p_\theta(x_{t-1} \mid x_t)$, respectively, and set $\Sigma_\theta(x_t, t) = \sigma_t^2 I$.

Substituting (5) and (7) into the conditional probability $q(x_{t-1} \mid x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I)$ and using Bayes' rule, the mean $\tilde{\mu}_t(x_t, x_0)$ and variance $\tilde{\beta}_t$ can be parameterized as follows:

$$\tilde{\mu}_t(x_t, x_0) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_t \right)$$

$$\alpha_t = 1 - \beta_t \tag{10}$$

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t. \tag{11}$$

3) Optimization Objective: The optimization objective of the DDPM is to recreate a sampling distribution $\tilde{x}_0$ that is as close as possible to the prior data distribution x_0, which can be achieved by minimizing the negative log likelihood (NLL) and optimized by using the variational lower bound

$$\begin{aligned} -\log p_\theta(x_0) \leq & -\log p_\theta(x_0) \\ & + D_{\mathrm{KL}}(q(x_{1:T} \mid x_0) \parallel p_\theta(x_{1:T} \mid x_0)) \\ = & \, \mathbb{E}_q \left[\log \frac{q(x_{1:T} \mid x_0)}{p_\theta(x_{0:T})} \right] \\ = & \, \mathbb{E}_q \left[-\log p(x_T) - \sum_{t \geq 1} \log \frac{p_\theta(x_{t-1} \mid x_t)}{q(x_t \mid x_{t-1})} \right] \\ = & \, \mathbb{E}_q [\underbrace{D_{\mathrm{KL}}(q(x_T \mid x_0) \parallel p(x_T))}_{\mathcal{L}_T} \\ & + \sum_{t>1} \underbrace{D_{\mathrm{KL}}(q(x_{t-1} \mid x_t, x_0) \parallel p_\theta(x_{t-1} \mid x_t))}_{\mathcal{L}_{t-1}} \\ & \underbrace{-\log p_\theta(x_0 \mid x_1)]}_{\mathcal{L}_0} \end{aligned} \tag{12}$$

where $D_{\mathrm{KL}}(\cdot \parallel \cdot)$ means the Kullback–Leibler divergence [51] and $\mathcal{L}_T$ and $\mathcal{L}_0$ are fixed values after the data distribution x_0 and the noise scheme β are determined. The parameterized $\mathcal{L}_{t-1}$ is given as the following equation after substituting the mean and variance of $q(x_{t-1} \mid x_t, x_0)$ and $p_\theta(x_{t-1} \mid x_t)$:

$$\mathcal{L}_{t-1} = \mathbb{E}_{x_0, \epsilon} \left[\frac{1}{2\sigma_t^2} \left\| \tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t) \right\|^2 \right]. \tag{13}$$

Thus, the optimization objective is simplified to make the predicted distribution $p_\theta(x_{t-1} \mid x_t)$ as close as possible to the posterior distribution $q(x_{t-1} \mid x_t, x_0)$ for any timestep $t > 1$.

III. PanDiff

In this section, we describe the details of PanDiff, including the design, the process, and the MIM.

A. Design of PanDiff

Although DDPM is a highly capable generative model, to be applicable to pansharpening, it is confronted with two major issues: 1) how to destroy HRMS with rich spatial and spectral information into approximate Gaussian noise x_T in limited timesteps and 2) how to guide the random Gaussian noise x_T to simulate the process of HRMS reconstruction, which inherently involves substantial uncertainty. We try to solve the above two problems by proposing PanDiff from a novel pansharpening fusion framework, as shown in Fig. 3.

In order to solve the above two problems, some adjustments are applied in PanDiff.

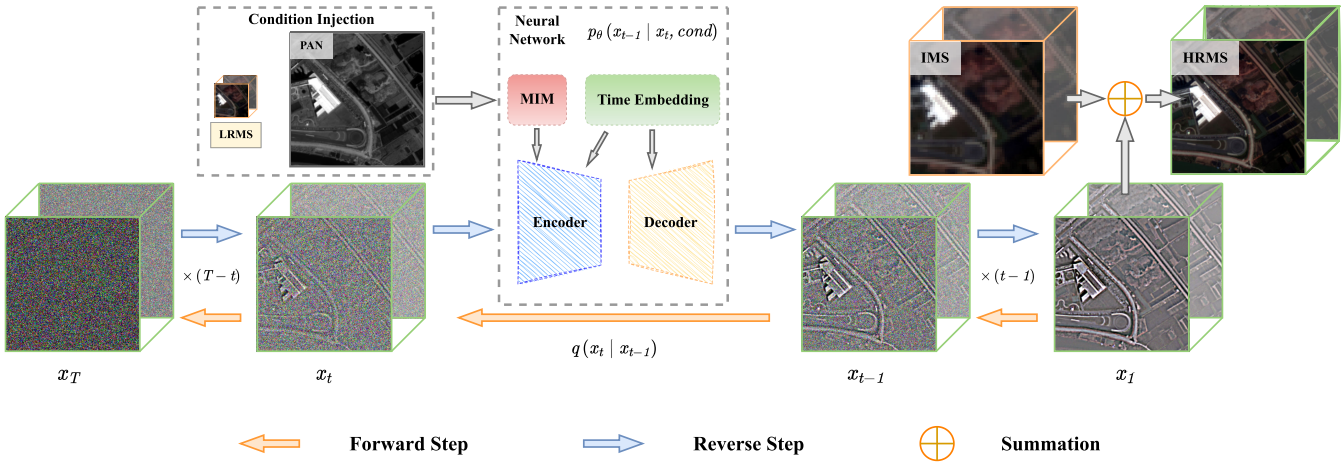

Fig. 3. Overall framework of the proposed PanDiff.

1) Difference Map: The use of DM of HRMS and IMS as learning objects for the proposed fusion framework is mainly motivated by two considerations. On the one hand, this strategy can effectively alleviate the difficulty of the work that involves converting the HRMS into a Gaussian noise and reconstructing it by reversion in a limited number of timesteps. On the other hand, the fusion objective of PanDiff is more clearly defined, which undoubtedly leads to better performance of the model.

2) Condition Injection: Since DDPM has a large uncertainty in the reverse process of reconstructing the initial input x_0 from Gaussian noise, PAN and LRMS images are used as conditional injections to bootstrap (4), and this process is rewritten as follows:

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t, \text{cond}) \tag{14}$$

$$\text{cond} = \Phi(\mathbf{P}, \mathbf{MS}) \tag{15}$$

where $\Phi(\cdot)$ is the encoder branch for processing PAN and LRMS images as the injected condition. More specifically, we achieve condition injection by concatenating modal-calibrated PAN and MS features by MIM with x_t.

B. Process of PanDiff

PanDiff is an improved DDPM method designed for pansharpening, and we discuss it in greater detail in the following.

1) Forward (Diffusion) Process: In this process, we feed the DM as the input to PanDiff (x_0 in Fig. 3) and continuously add Gaussian noise to x_0 based on the Markov chain modeling approach to obtain an approximate standard Gaussian distribution (x_T in Fig. 3) and generate $t-1$ latent states $\{x_1, x_2, \ldots, x_{T-1}\}$. The process of adding noise is executed a total of T times and the noise variance β_t increases linearly as timestep t increases. Since the function of β_t versus t is artificially set, it is possible to directly calculate the latent states at each step by reparameterized (7), as indicated in the following equation:

$$x_t = \sqrt{\bar{\alpha}} x_0 + \sqrt{1-\bar{\alpha}}\epsilon$$
$$\epsilon \sim \mathcal{N}(0, I). \tag{16}$$

2) Reverse (Denoised) Process: The intention of this process is to learn the unknown reverse process $q(x_{t-1} \mid x_t)$ of T-step. We first sample a random noise (x_T in Fig. 3) from a standard Gaussian distribution and build a neural network ϵ_θ to model the transition from x_t to x_{t-1} given conditions by learning the data distribution [i.e., the mean $\mu_\theta(x_t, \text{cond}, t)$ and variance $\Sigma_\theta(x_t, \text{cond}, t)$] of $p_\theta(x_{t-1} \mid x_t, \text{cond})$. Timestep t is input into the network via time embedding in order to bolster the variance at each timestep. In addition, to guide the network's learning of this DM reconstruction process, the modal-calibrated PAN and MS features are treated as the injected conditions for each timestep.

3) Architecture of Network ϵ_θ: When DDPM is proposed, U-Net [52] with the attention module is utilized [53] and proven to be more adept at fitting the data distribution [54]. We follow the main structure of U-Net in DDPM and add an information intercalibration module to the injected conditions PAN and MS for better handling of the relationship between different modal data.

4) Time Embedding: In order to enhance the precision of predicting the distribution $p_\theta(x_{t-1} \mid x_t, \text{cond})$, it is crucial to increase the model ϵ_θ's sensitivity toward timesteps. To accomplish this, the technique of time embedding is introduced. The simplified optimization objective (18) explicitly indicates that $\sqrt{\bar{\alpha}_t}$ has the most direct impact on model prediction performance among all mappings of timesteps t, and sinusoidal position embeddings [33] are applied to $\sqrt{\bar{\alpha}_t}$

$$\text{TE}(t, i) = \text{Concat}\left(\left[\sin\left(\frac{\sqrt{\bar{\alpha}_t}}{10000^{\frac{i}{d-1}}}\right), \cos\left(\frac{\sqrt{\bar{\alpha}_t}}{10000^{\frac{i}{d-1}}}\right)\right]\right) \tag{17}$$

where d is the half dimension of the time embedding, $i = \{0, 1, \ldots, d-1\}$. $\text{TE}(t, i) \in \mathbb{R}^2$ and the time embedding feature $\text{TE}(t) \in \mathbb{R}^{2d}$ is obtained by concatenating $\{\text{TE}(t, i)\}_{i=0}^{d-1}$. $\text{TE}(t)$ is currently a constant value that is not trainable.

As a solution, two multilayer perceptron (MLP) layers are introduced to acquire the final trainable TE(t).

5) Model Training: Following the forward and reverse processes of PanDiff described above, the latent state transition is learned by the neural network for each step and the error is backpropagated to train the model. We give the optimization objective $\mathcal{L}_{\text{simple}}$ in the following equation after adding the condition and simplifying the parameterized $\mathcal{L}_{t-1}$ in (13):

$$\mathcal{L}_{\text{simple}}(\theta) = \mathbb{E}_{t,x_0,\epsilon}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \text{cond}, t\right)\right\|^2\right]. \quad (18)$$

The detailed training process of PanDiff can be found in Algorithm 1.

Algorithm 1 Training Algorithm for PanDiff

Input: Pansharpening dataset $\mathbf{D} = \{(\mathbf{P}_i, \mathbf{MS}_i, \mathbf{GT}_i)\}_{i=1}^N$.

repeat
  Sample $(\mathbf{P}_i, \mathbf{MS}_i, \mathbf{GT}_i) \sim \mathbf{D}$
  $t \sim \text{Uniform}(\{1, \ldots, T\})$
  $\epsilon \sim \mathcal{N}(0, I)$
  $\widetilde{\mathbf{MS}}_i = \text{Interpolate}(\mathbf{MS}_i)$
  $x_0 = \Delta\mathbf{MS}_i = \mathbf{GT}_i - \widetilde{\mathbf{MS}}_i$
  $cond = \Phi(\mathbf{P}_i, \mathbf{MS}_i)$
  Take gradient descent step on $\nabla_\theta\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, cond, t\right)\right\|^2$
until *converged*;

6) Model Sampling: After training U-Net ϵ_θ, the inference of x_{t-1} from x_t becomes possible according to (9). The distribution of $q(x_{t-1} \mid x_t)$ is predicted by successive T timesteps, and by sampling noise and reparameterization from the predicted distribution, the data distribution of the DM can finally be deduced. The details of the sampling process are shown in Algorithm 2.

Algorithm 2 Sampling Algorithm for PanDiff

Input: Pansharpening data $\mathbf{D}_i = (\mathbf{P}_i, \mathbf{MS}_i, \mathbf{GT}_i) \sim \mathbf{D}$, Neural Network $\boldsymbol{\epsilon_\theta}$.
Output: $\widehat{\mathbf{MS}}_i$

$x_T \sim \mathcal{N}(0, I)$
for $t \leftarrow T$ **to** 1 **do**
  $z \sim \mathcal{N}(0, I)$ if t>1, else $z = 0$
  $cond = \Phi(\mathbf{P}_i, \mathbf{MS}_i)$
  $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(x_t, cond, t)\right) + \sigma_t z$
end
$\widehat{\mathbf{MS}}_i = x_0 + \widetilde{\mathbf{MS}}_i$
return $\widehat{\mathbf{MS}}_i$

C. Modal Intercalibration Module

Significant modal differences (i.e., spectral and spatial differences) exist between PAN and LRMS images, allowing PAN and LRMS images to guide the modeling of $q(x_{t-1} \mid x_t)$ in neural networks by focusing on various aspects. Directly feeding PAN and LRMS into U-Net is not a sensible strategy. To explicitly leverage the modal variations between PAN and LRMS images to enhance the injection conditions, we propose an MIM considering both the channel and the spatial information intercalibration.

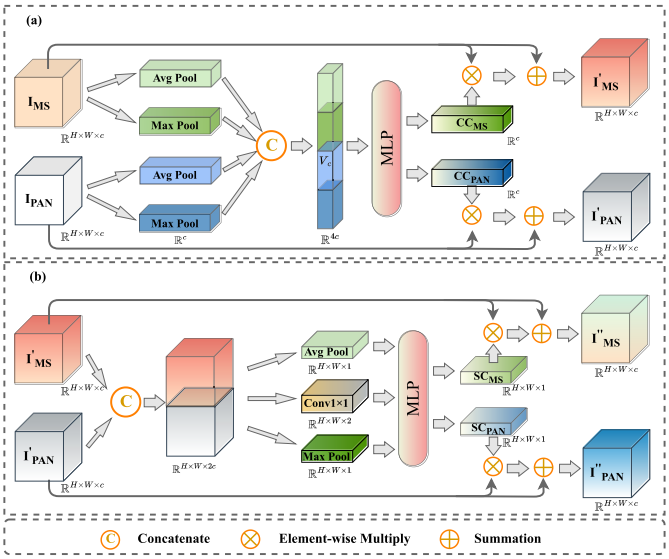

Fig. 4. Schematic of (a) MIM-Spectral and (b) MIM-Spatial.

First, PAN and IMS images are fed into the convolution module to obtain dimension-specific PAN features $I_{\text{PAN}} \in \mathbb{R}^{H\times W\times c}$ and IMS features $I_{\text{MS}} \in \mathbb{R}^{H\times W\times c}$. Following this, the spectrum information intercalibration module (MIM-Spectral) and the spatial information intercalibration module (MIM-Spatial) are applied to the features I_{PAN} and I_{MS}, respectively; MIM-Spectral and MIM-Spatial are introduced in detail next.

1) MIM-Spectral: The diagram of MIM-Spectral is shown in Fig. 4(a). To calibrate the spectral differences between PAN and LRMS images, we obtain sufficient spectral information vectors $v_p \in \mathbb{R}^c$ as a basis for intercalibration by averaging global pooling and max global pooling via the channel dimension and concatenate them to form $V_c \in \mathbb{R}^{4c}$. Next, an MLP and a sigmoid function $\sigma(\cdot)$ are used to obtain channel calibration weights for PAN and MS features, named CC_{PAN} and CC_{MS}. Finally, the PAN and MS features are multiplied elementwise with CC_{PAN} and CC_{MS} in the channel dimension, respectively, and are added with the PAN and MS features themselves to obtain the final spectral intercalibrated information I'_{PAN} and I'_{MS}. The process of MIM-Spectral is shown in the following equation:

$$\begin{aligned}\text{CC}_{\text{MS}}, \text{CC}_{\text{PAN}} &= \sigma(\text{MLP}(V_C))\\ I'_{\text{MS}} &= I_{\text{MS}} \times \text{CC}_{\text{MS}} + I_{\text{MS}}\\ I'_{\text{PAN}} &= I_{\text{PAN}} \times \text{CC}_{\text{PAN}} + I_{\text{PAN}}.\end{aligned} \quad (19)$$

2) MIM-Spatial: The illustration of MIM-Spatial can be found in Fig. 4(b). MIM-Spatial concatenates the input spectral intercalibrated information I'_{PAN} and I'_{MS} and performs average global pooling and maximum global pooling in the channel

dimension to obtain $v_p \in \mathbb{R}^{h\times w\times 1}$. Compared with MIM-Spectral, MIM-Spatial has to deal with more complicated modeling information, so we generate a small amount of key information $v_{1\times 1} \in \mathbb{R}^{H\times W\times 2}$ by 1×1 convolution. Then, we concatenate v_p and $v_{1\times 1}$ as $V_S \in \mathbb{R}^{H\times W\times 4}$ and input them into an MLP and obtain spatial correction weights $\mathrm{SC_{PAN}}$ and $\mathrm{SC_{MS}}$ by a sigmoid function

$$\begin{aligned} \mathrm{SC_{MS}}, \mathrm{SC_{PAN}} &= \sigma(\mathrm{MLP}(V_S)) \\ I''_{\mathrm{MS}} &= I'_{\mathrm{MS}} \times \mathrm{SC_{MS}} + I'_{\mathrm{MS}} \\ I''_{\mathrm{PAN}} &= I'_{\mathrm{PAN}} \times \mathrm{SC_{PAN}} + I'_{\mathrm{PAN}}. \end{aligned} \tag{20}$$

IV. Datasets and Experimental Details

A. Datasets

The PanCollection dataset [55] containing data from three satellites (GaoFen-2, QuickBird, and WorldView-3) is utilized to evaluate our PanDiff with other state-of-the-art methods fairly and comprehensively. The three types of data cover different geographical locations with various bands and spatial resolutions. Specifically, the spatial resolutions of PAN images in GaoFen-2 and QuickBird are 0.8 m and 0.6 m, respectively, and their corresponding MS images are 3.2 and 2.4 m with four bands, including red, green, blue, and near infrared. Compared with the two data mentioned above, WorldView-3 has a higher spatial resolution (PAN: 0.3 m and MS: 1.2 m) and provides four more bands, including coastal, yellow, red edge, and near infrared-2. Moreover, the data from WorldView-2 with eight bands are utilized for generation tests. The more specific information of these four data is presented in Table I.

B. Benchmark

The benchmark consists of two representative CS-based methods (the Brovey transform with haze correction (BT-H) [56] and the band-dependent spatial detail with physical constraints (BDSD-PC) [57]), two representative MRA-based methods (the generalized Laplacian pyramid with MTF-matched filters with an FS regression-based injection model (MTF-GLP-FS) [16], and the generalized Laplacian pyramid with MTF-matched filters and a high-pass modulation injection model with a preliminary regression-based spectral matching phase (MTF-GLP-HPM-R) [17]) and seven state-of-the-art DL-based CNN methods (PNN [27], PanNet [28], DRPNN [29], the multiscale and multidepth convolutional neural network (MSDCNN) [58], the detail injection-based CNN (DiCNN) [59], the explicit spectral-to-spatial convolution (SSconv) [60], and the triple-double convolutional neural network (TDNet) [61]), and two state-of-the-art GAN methods (PSGAN [39] and MDSSC-GAN [40]). The implementation of the above methods heavily relies on DLPan-Toolbox [55], [62].[1]

C. Evaluation Metrics

To quantitatively assess the rationality and superiority of the proposed method, we introduce several reduced- and full-resolution evaluation metrics in our experiment. For the reduced resolution, the evaluation metrics include the peak signal-to-noise ratio (PSNR), the structural similarity index (SSIM) [63], the spectral angle mapper (SAM) [64], the erreur relative globale adimensionnelle de synthèse (ERGAS) [65], and the spatial correlation coefficient (SCC) [66]. These evaluation metrics possess good discriminative power for different features of the fused images. SSIM and SCC more effectively measure the spatial similarity of the results, while SAM focuses on discriminating spectral differences. PSNR and ERGAS evaluate the model performance considering both spectral and spatial differences.

For the full-resolution experiments, we employ the no-reference evaluation metric quality without reference (QNR) [67] to evaluate the performance of the model on real images, as no reference image is available. QNR comprises evaluation metrics D_λ and D_S, which are used to assess spectral and spatial distortions, respectively. However, it has been pointed out that the assumptions in the calculation of the spectral distortion index are flawed [68]. Thus, we also utilize another hybrid QNR (HQNR) [69] to circumvent this issue and prevent potential misjudgment of the experimental results.

D. Data Preprocessing and Augmentation

The specific strategy for data preprocessing and augmentation in this experiment is given as follows.

1) *Preprocessing (Data Normalization):* Since both the input and output of DDPM need to be approximated as standard Gaussian distributions, the normalized data need to be obtained by the calculation of the following equations:

$$d' = 2 \times \frac{d}{2^\gamma} - 1 \tag{21}$$

where γ is the radiometric resolution of the data, and the output d' will be normalized to $[-1, 1]$.

2) *Augmentation (Random Flipping):* A good pansharpening algorithm should be insensitive to the rotation of images, so we strengthen the algorithm by random flipping (horizontal and vertical) and rotation (90°, 180°, and 270°).

E. Implementation Details

For all experiments, we establish a virtual Anaconda environment with Python 3.7 and PyTorch 1.8.2 as the standard. The specific graph computation platform contains CUDA 11.1, CUDNN 8.0.4, and TensorRT 7.2.3.4 on two NVIDIA RTX 3090 GPUs.

The specific parameter configuration is given as follows. All experiments of PanDiff use a batch size of 384. The models are iterated 320 000 times with weights and evaluated every 5000 iterations for model performance. Adam with weight decay (AdamW) [70] is set as the standard optimizer. We adopt a MultiStep learning rate (LR) scheduler and the LR is initially set to 1×10^{-4}. The milestones and gamma of the LR scheduler are [96, 000, 192, 000, 288, 000, 304, 000] and

[1]The source code of these comparison methods can be found at the website: https://github.com/liangjiandeng/DLPan-Toolbox

TABLE I
INFORMATION OF SATELLITE DATA (GAOFEN-2, QUICKBIRD, WORLDVIEW-3, AND WORLDVIEW-2)

Satellite		GaoFen-2	QuickBird	WorldView-3	WorldView-2
Band		4	4	8	8
Spatial Resolution (m)	PAN	0.8	0.6	0.3	0.46
	MS	3.2	2.4	1.2	1.84
Radiometric Resolution (bit)		10	11	11	11
Spatial Resolution Ratio		4	4	4	4
Train / Val		19809 / 2201	17139 / 1905	9714 / 1080	- / 20
Image Size	PAN	64 × 64 × 1	64 × 64 × 1	64 × 64 × 1	512 × 512 × 1
	MS	16 × 16 × 4	16 × 16 × 4	16 × 16 × 8	128 × 128 × 8
Location		Guangzhou, China	Indianapolis, USA	Rio, Brazil Tripoli, Lebanon	Washington, D.C., USA

Note that the '-' in the table indicates that no data from WorldView-2 is used for training.

TABLE II
TRAINING PARAMETERS FOR COMPARISON METHODS. 0.1/200 IN LR SCHEDULER MEANS THAT THE LR IS MULTIPLIED BY 0.1 EVERY 200 EPOCHS

Epoch	Batch Size	Optimizer	Initial LR	LR Scheduler	Loss
500	384	AdamW	3×10^{-4}	0.1/200	L_2

0.5. In addition, for the hyperparameters of PanDiff, the total timesteps T is set to 2000, and the variance β_t of added noise in the forward process is set from 1×10^{-2} to 1×10^{-6}.

The training parameters for the comparison methods are listed in Table II.

V. EXPERIMENTAL RESULTS

In our experiments, the method performance on both reduced resolution and full resolution is conducted. It is worth noting that all of the figures are shown in true color (red, green, and blue), where the specific ground objects are marked by blue and green boxes and are enlarged at the bottom. The bold results in the tables indicate the best. In addition, we performed ablation experiments to verify the effectiveness of the DM and MIM. Finally, the effect of the hyperparameter total timesteps T on the model performance and the model runtime efficiency are also fully discussed.

A. Reduced-Resolution Experiments

In this section, we present the qualitative and quantitative results of the proposed method and comparison methods on the GaoFen-2, QuickBird, and WorldView-3 datasets. The size of the IMS, PAN, and GT images used in the figures in this section is 256 × 256, to facilitate visual presentation.

As shown in Fig. 5, all of these methods exhibit excellent fusion performance. From a subjective visual perspective, the differences between the fusion images are subtle. However, the fusion performance between traditional methods and DL-based methods can still be clearly compared on the error map. Based on the error maps corresponding to each image, it is evident that the DL-based methods have greater fusion accuracy than the traditional ones, as indicated by the lower brightness of their error maps. Among DL-based methods, the color of the error map for PanDiff is primarily purple and blue, indicating that the difference between PanDiff and GT is the smallest, and the fusion performance is the best, followed by MDSSC-GAN and PSGAN. In addition, due to the tendency of smoothing features of CNN, the high-frequency edge information of the fusion result generates larger errors, which are specifically manifested as high brightness of the terrain boundaries in the error map. It can also be seen from the enlarged blue and green boxes of the fusion image that PanDiff has the closest spatial texture information and spectral tone to GT.

In Table III, all the evaluation metrics of the DL-based method significantly outperform the traditional methods, consistent with the subjective analysis mentioned above. PanDiff excels in spatial information enhancement and spectral fidelity, with its minimum SAM value and maximum SSIM and SCC values. Furthermore, the fusion effect of MDSSC-GAN and PSGAN is also satisfactory, with relatively good evaluation metrics. In contrast, PNN and PanNet methods have poor spatial learning ability and spectral preservation within the DL methods. In traditional fusion methods, the CS-based methods exhibit severe spectral distortion (their SAM values are the highest among these methods), and the MRA-based methods have poor spatial information (their SCC values are the lowest among these methods).

As can be seen in Fig. 6, the fusion results of QuickBird case resemble those of GaoFen2, that is, PanDiff, PSGAN, and MDSSC-GAN exhibit superior fusion performance, as observed from the error map. The difference between traditional fusion methods and DL-based fusion methods can also be clearly compared from the error map. However, in the DL-based method, PNN, DiCNN, and TDNet have poor fusion performance, with high brightness values in their error maps, even comparable to traditional fusion methods. Moreover, the clarity of the boats in the blue box can adequately demonstrate that the fusion results of the DL-based methods contain richer spatial information compared to those of the traditional methods. The predominance of PanDiff in retaining spectral information is visible in the building roof color in the green box, which has the least visual difference between the color distribution of GT and that of PanDiff. In contrast, the spectral learning ability of PanNet and TDNet is low, and their fused images have comparatively darker tones than traditional fusion methods.

TABLE III
QUANTITATIVE METRICS FOR ALL THE COMPARISON METHODS ON THE REDUCED-RESOLUTION GAOFEN-2 DATASET

Methods	$PSNR\uparrow(\pm std)$	$SSIM\uparrow(\pm std)$	$SAM\downarrow(\pm std)$	$ERGAS\downarrow(\pm std)$	$SCC\uparrow(\pm std)$
BT-H [56]	35.77±1.35	0.9260±0.0034	1.8485±0.0772	1.7452±0.0677	0.8633±0.0184
BDSD-PC [57]	35.27±1.49	0.9206±0.0029	1.9638±0.0740	1.9046±0.0760	0.8634±0.0202
MTF-GLP-FS [16]	35.94±1.34	0.9206±0.0020	1.7617±0.0512	1.7342±0.0529	0.8581±0.0234
MTF-GLP-HPM-R [17]	35.96±1.32	0.9215±0.0021	1.7530±0.0514	1.7301±0.0496	0.8598±0.0242
PNN [27]	39.81±0.86	0.9668±0.0033	1.1385±0.0718	1.0474±0.0622	0.9479±0.0059
PanNet [28]	39.68±0.83	0.9663±0.0033	1.2009±0.0736	1.0656±0.0628	0.9480±0.0059
DRPNN [29]	40.48±0.86	0.9711±0.0029	1.0873±0.0681	0.9714±0.0581	0.9548±0.0053
MSDCNN [58]	40.46±0.90	0.9705±0.0029	1.0741±0.0688	0.9789±0.0596	0.9537±0.0054
DiCNN [59]	39.81±0.91	0.9676±0.0033	1.1277±0.0696	1.0559±0.0630	0.9491±0.0058
SSconv [60]	40.90±0.91	0.9726±0.0027	1.0193±0.0656	0.9339±0.0579	0.9571±0.0050
TDNet [61]	39.72±0.87	0.9668±0.0033	1.2147±0.0764	1.0649±0.0633	0.9491±0.0058
PSGAN [39]	41.77±0.81	0.9795±0.0021	0.9443±0.0564	0.8279±0.0412	0.9684±0.0045
MDSSC-GAN [40]	42.55±0.92	0.9818±0.0018	0.8295±0.0530	0.7623±0.0447	0.9722±0.0032
PanDiff	**43.40±0.64**	**0.9837±0.0013**	**0.7735±0.0367**	**0.6875±0.0307**	**0.9771±0.0022**

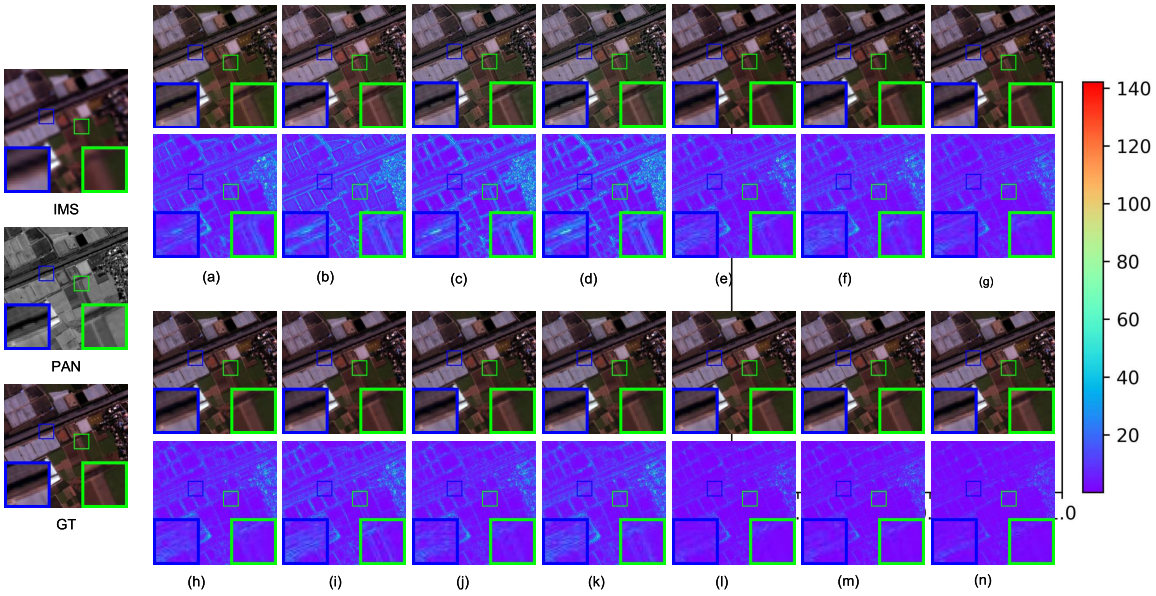

Fig. 5. Visual comparisons on a reduced-resolution GaoFen-2 case. Lines 1 and 3 are the predicted HRMS for each method, and lines 2 and 4 are the error map of the predicted HRMS versus GT for each one. (a) BH-T. (b) BDSD-PC. (c) MTF-GLP-FS. (d) MTF-GLP-HPM-R. (e) PNN. (f) PanNet. (g) DRPNN. (h) MSDCNN. (i) DiCNN. (j) SSconv. (k) TDNet. (l) PSGAN. (m) MDSSC-GAN. (n) PanDiff.

In Table IV, it can be clearly contrasted that PanDiff has superior evaluation metrics, especially its global evaluation metrics far surpassing that of other methods. PSGAN and MDSSC-GAN, which are also based on generative models, have excellent evaluation metrics results as well. Unlike the GaoFen-2 fusion results, the DiCNN and TDNet spatial fusion metrics are unsatisfactory and even inferior to PNN and PanNet among the DL-based methods. Similar to the GaoFen-2 fusion results, PNN, PanNet, and TDNet perform poorly in QuickBird fusion with respect to spectral learning. Surprisingly, among the traditional methods, the spectral evaluation metric (SAM) of BT-H, which belongs to the CS-based method, is better than that of MRA-based methods, corresponding to subjective vision.

As observed in Fig. 7, in the fusion process of WorldView-3 imagery, the brightness values from the error map indicate that both PanDiff and MDSSC-GAN display exceptional fusion performance. The fusion results, particularly the clear and magnified rooftop textures, reveal that these two methods excel in enhancing spatial information. Moreover, their fusion results consistently maintain a high degree of spectral fidelity. Similarly, the intensity of the error map indicates that the DL-based method fusion effect is superior to the traditional one, with the exception of SSconv. SSconv is the most inferior in the DL-based method, and its feature color in the blue box of the error map is the brightest, which is comparable to the traditional method. In addition, the amplified texture details of the ground objects reveal that the spatial information learning capability of DL fusion methods is significantly superior to that of traditional fusion techniques.

In Table V, we can see that the fusion performance of MDSSC-GAN is comparable to that of the PanDiff network, and it surpasses PanDiff in some evaluation metrics. However, the variance of MDSSC-GAN's evaluation metrics is larger than that of PanDiff, indicating that the stability of the network model is inferior. This can also be observed in the subjective visual presentation, where the spectral fidelity of the fused images presented in this article does not appear to be closer to GT than that of PanDiff. Furthermore, while PSGAN demonstrates commendable spectral fidelity, it is also

TABLE IV
QUANTITATIVE METRICS FOR ALL THE COMPARISON METHODS ON THE REDUCED-RESOLUTION QUICKBIRD DATASET

Methods	$PSNR\uparrow(\pm std)$	$SSIM\uparrow(\pm std)$	$SAM\downarrow(\pm std)$	$ERGAS\downarrow(\pm std)$	$SCC\uparrow(\pm std)$
BT-H [56]	35.61±0.90	0.8940±0.0058	6.3700±0.3679	7.0396±1.3166	0.8109±0.1796
BDSD-PC [57]	35.95±0.60	0.8957±0.0045	6.7232±0.3074	6.3606±0.0745	0.8979±0.0032
MTF-GLP-FS [16]	36.12±0.64	0.8967±0.0055	6.4589±0.3403	6.2240±0.0841	0.8975±0.0035
MTF-GLP-HPM-R [17]	36.12±0.65	0.8985±0.0059	6.4546±0.3811	6.7141±1.0089	0.8650±0.0070
PNN [27]	37.69±0.82	0.9289±0.0057	5.1577±0.2658	5.3945±0.3255	0.9411±0.0105
PanNet [28]	37.92±0.85	0.9321±0.0069	5.0604±0.2593	5.2545±0.3729	0.9511±0.0090
DRPNN [29]	39.31±0.73	0.9494±0.0051	4.5977±0.2165	4.4963±0.3218	0.9661±0.0066
MSDCNN [58]	38.62±0.78	0.9410±0.0054	4.8659±0.2411	4.8606±0.3183	0.9560±0.0083
DiCNN [59]	37.55±0.88	0.9266±0.0060	5.1528±0.2750	5.4922±0.3181	0.9370±0.0104
SSconv [60]	38.85±0.78	0.9433±0.0059	4.7277±0.2329	4.7828±0.3686	0.9627±0.0076
TDNet [61]	37.50±0.85	0.9266±0.0065	5.2085±0.2707	5.5289±0.3581	0.9451±0.0096
PSGAN [39]	40.07±0.75	0.9565±0.0047	4.2570±0.2025	4.1415±0.3126	0.9718±0.0058
MDSSC-GAN [40]	40.01±0.76	0.9557±0.0049	**4.2450±0.1998**	4.1772±0.3145	0.9710±0.0058
PanDiff	**41.70±0.73**	**0.9569±0.0073**	4.3193±0.2553	**3.7824±0.3526**	**0.9725±0.0067**

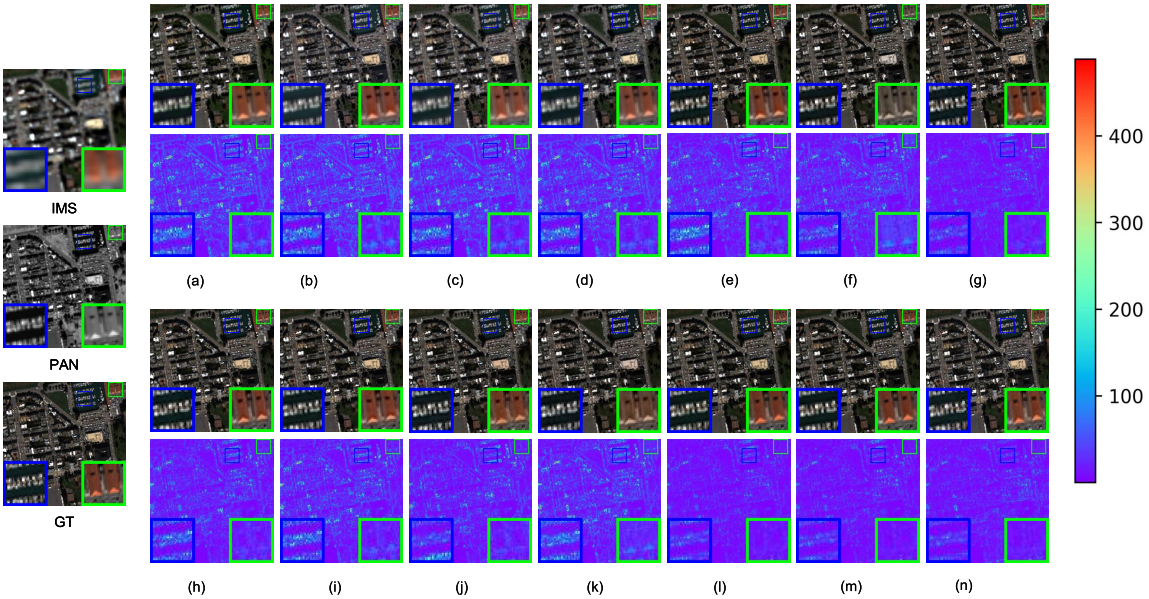

Fig. 6. Visual comparisons on a reduced-resolution QuickBird case. Lines 1 and 3 are the predicted HRMS for each method, and lines 2 and 4 are the error map of the predicted HRMS vs. GT for each one. (a) BH-T. (b) BDSD-PC. (c) MTF-GLP-FS. (d) MTF-GLP-HPM-R. (e) PNN. (f) PanNet. (g) DRPNN. (h) MSDCNN. (i) DiCNN. (j) SSconv. (k) TDNet. (l) PSGAN. (m) MDSSC-GAN. (n) PanDiff.

confronted with the challenge of higher variance. In light of these findings, we deduce that the PanDiff method not only offers superior fusion performance but also maintains a higher level of stability. In contrast, TDNet and SSconv spectral are the most distorted, even comparable to the spectral learning effect of traditional fusion methods, and their spatial enhancement effect is also not desirable among the DL-based methods. Similar to the QuickBird fusion effect, as for the traditional methods, the spectral fidelity of the BT-H fused effect is higher, while the spatial information of the MTF-GLP-FS fused effect is richer, which differs from the individual theoretical advantages of the CS- and MRA-based methods.

It is worth noting that the standard deviation (std) values of all evaluation metrics of PanDiff are small in these three sets of reduced-resolution experiments, which are in the middle-to-upper level of these comparison methods, indicating that PanDiff has strong fusion stability for various categories of ground objects in different types of remote sensing data.

B. Full-Resolution Experiments

The reduced-resolution experiments primarily reflect the performance of the method on simulated data, but its applicability to real data remains uncertain. To thoroughly demonstrate the efficacy of our method, we conduct evaluation experiments on full-resolution satellite images in this section. For ease of visual presentation, the size of the PAN and IMS images used in the figures in this section is 512 × 512.

Table VI and Fig. 8 present the quantitative and graphical results, respectively, both based on the GaoFen-2 dataset. In full-resolution experiments, the evaluation metric D_λ clearly highlights the spectral fidelity gap between traditional and DL-based methods, which is a positive indication of the prominent spectral learning potential of DL-based methods, except for DRPNN. The evaluation metric D_S also contributes to the evidence that the DL-based method recovers spatial features at full resolution more successfully. Our PanDiff outperforms other methods in both spectral and spatial evaluation metrics. In addition, the fusion results of SSconv and PSGAN

TABLE V

QUANTITATIVE METRICS FOR ALL THE COMPARISON METHODS ON THE REDUCED-RESOLUTION WORLDVIEW-3 DATASET

Methods	$PSNR\uparrow(\pm std)$	$SSIM\uparrow(\pm std)$	$SAM\downarrow(\pm std)$	$ERGAS\downarrow(\pm std)$	$SCC\uparrow(\pm std)$
BT-H [56]	34.01±0.17	0.8919±0.0049	5.0507±0.2034	4.6116±0.5112	0.8396±0.0725
BDSD-PC [57]	34.16±0.50	0.8936±0.0070	5.4270±0.3254	4.2158±0.4114	0.8896±0.0121
MTF-GLP-FS [16]	34.34±0.46	0.8912±0.0061	5.2064±0.2666	4.1189±0.3519	0.8887±0.0111
MTF-GLP-HPM-R [17]	34.29±0.32	0.8908±0.0056	5.2746±0.2372	5.2877±1.6626	0.7303±0.2373
PNN [27]	35.33±0.97	0.9353±0.0106	4.7710±0.3126	3.2770±0.2629	0.9466±0.0120
PanNet [28]	35.73±1.02	0.9420±0.0109	4.5640±0.2895	3.1073±0.2657	0.9511±0.0114
DRPNN [29]	35.89±1.01	0.9420±0.0107	4.5424±0.3056	3.0348±0.2442	0.9536±0.0113
MSDCNN [58]	36.26±1.04	0.9452±0.0106	4.3532±0.2944	2.9265±0.2432	0.9549±0.0110
DiCNN [59]	36.13±1.08	0.9442±0.0106	4.3644±0.2970	3.0155±0.2580	0.9519±0.0111
SSconv [60]	34.85±1.04	0.9297±0.0113	4.9073±0.3127	3.4328±0.2902	0.9437±0.0116
TDNet [61]	34.80±0.99	0.9326±0.0110	5.0430±0.3077	3.4117±0.2781	0.9447±0.0120
PSGAN [39]	36.37±1.16	0.9465±0.0111	4.2046±0.2998	2.9449±0.2691	0.9537±0.0115
MDSSC-GAN [40]	**37.14±1.14**	**0.9520±0.0108**	**3.8891±0.2770**	**2.6950±0.2373**	0.9600±0.0108
PanDiff	36.58±0.80	0.9444±0.0096	4.2191±0.2346	2.8682±0.1659	**0.9602±0.0089**

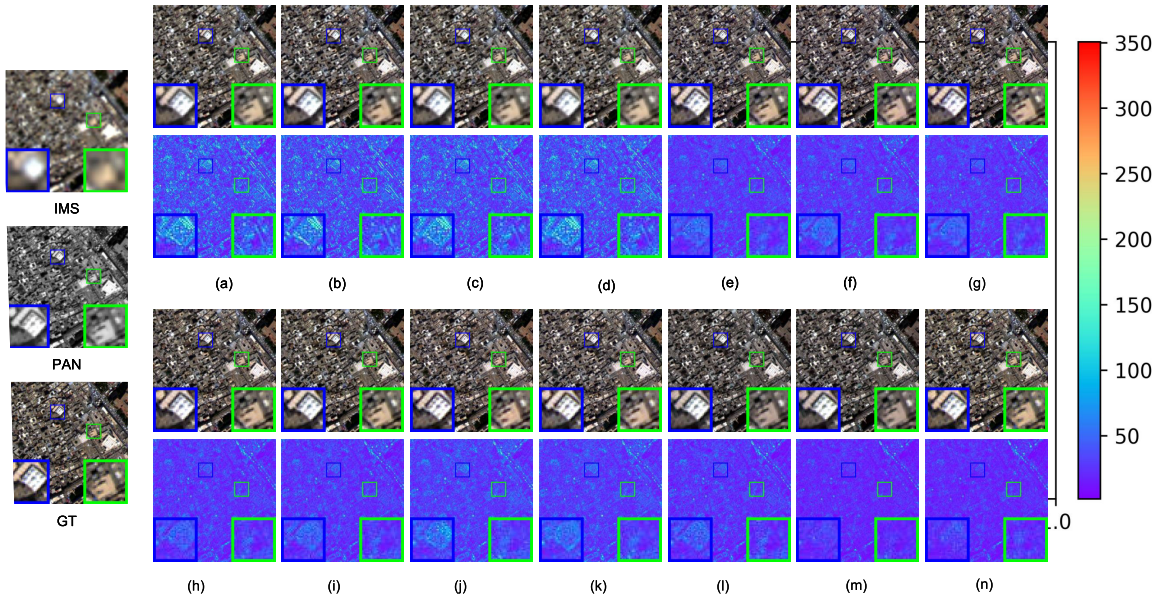

Fig. 7. Visual comparisons on a reduced-resolution WorldView-3 case. Lines 1 and 3 are the predicted HRMS for each method, and lines 2 and 4 are the error map of the predicted HRMS versus GT for each one. (a) BH-T. (b) BDSD-PC. (c) MTF-GLP-FS. (d) MTF-GLP-HPM-R. (e) PNN. (f) PanNet. (g) DRPNN. (h) MSDCNN. (i) DiCNN. (j) SSconv. (k) TDNet. (l) PSGAN. (m) MDSSC-GAN. (n) PanDiff.

TABLE VI

QUANTITATIVE METRICS FOR ALL THE COMPARISON METHODS ON THE FULL-RESOLUTION GAOFEN-2 DATASET

Methods	$D_\lambda\downarrow(\pm std)$	$D_S\downarrow(\pm std)$	$QNR\uparrow(\pm std)$	$HQNR\uparrow(\pm std)$
BT-H [56]	0.0891±0.0335	0.1712±0.0388	0.7399±0.0551	0.7558±0.0567
BDSD-PC [57]	0.0926±0.0292	0.1652±0.0362	0.7767±0.0539	0.7582±0.0509
MTF-GLP-FS [16]	0.0370±0.0138	0.1539±0.0351	0.7636±0.0542	0.8150±0.0404
MTF-GLP-HPM-R [17]	0.0364±0.0131	0.1531±0.0353	0.7650±0.0545	0.8163±0.0400
PNN [27]	0.0490±0.0693	0.1263±0.0338	0.8256±0.0194	0.8236±0.0564
PanNet [28]	0.0353±0.0105	0.1035±0.0258	0.8494±0.0409	0.8649±0.0271
DRPNN [29]	0.0374±0.0148	0.1115±0.0321	0.8265±0.0519	0.8555±0.0390
MSDCNN [58]	0.0298±0.0118	0.0869±0.0194	0.8729±0.0323	0.8859±0.0202
DiCNN [59]	0.0329±0.0098	0.0921±0.0248	0.8580±0.0394	0.8781±0.0273
SSconv [60]	0.0228±0.0084	0.0478±0.0156	0.9232±0.0274	0.9304±0.0146
TDNet [61]	0.0301±0.0096	0.0839±0.0202	0.8786±0.0324	0.8885±0.0201
PSGAN [39]	0.0224±0.0080	0.0698±0.0186	0.9012±0.0296	0.9093±0.0177
MDSSC-GAN [40]	0.0280±0.0083	0.0904±0.0228	0.8664±0.0361	0.8842±0.0237
PanDiff	**0.0223±0.0103**	**0.0323±0.0131**	**0.9396±0.0250**	**0.9461±0.0125**

are also comparable at full resolution. However, surprisingly, DRPNN, which performs well at reduced resolution, does not seem to effectively transfer the satisfactory performance to full resolution. In addition, the optimal QNR and HQNR values can comprehensively reflect that our proposed PanDiff has superior fusion performance among all comparison methods. The green vegetation and the brown ground in the bottom boxes of Fig. 8 can also be a good discriminator of the differences in the information retention capacity of these methods. Specifically, the fusion results of CS-based methods

TABLE VII
QUANTITATIVE METRICS FOR ALL THE COMPARISON METHODS ON THE FULL-RESOLUTION QUICKBIRD DATASET

Methods	$D_\lambda \downarrow (\pm std)$	$D_S \downarrow (\pm std)$	$QNR \uparrow (\pm std)$	$HQNR \uparrow (\pm std)$
BT-H [56]	0.2788±0.1323	0.1835±0.0861	0.7601±0.0937	0.5965±0.1535
BDSD-PC [57]	0.2245±0.0588	0.1789±0.1051	0.7806±0.1227	0.6420±0.1222
MTF-GLP-FS [16]	0.0674±0.0295	0.1708±0.0783	0.7634±0.0899	0.7751±0.0903
MTF-GLP-HPM-R [17]	0.0719±0.0353	0.1558±0.0800	0.7841±0.0928	0.7851±0.0922
PNN [27]	0.0992±0.0480	0.1205±0.0981	0.8129±0.1330	0.7961±0.1223
PanNet [28]	0.1475±0.0796	0.1224±0.0956	0.7993±0.1424	0.7545±0.1391
DRPNN [29]	0.0934±0.0463	0.0933±0.0644	0.8386±0.1164	0.8244±0.0927
MSDCNN [58]	0.0841±0.0406	0.1004±0.0862	0.8284±0.1270	0.8270±0.1094
DiCNN [59]	0.1085±0.0330	0.1381±0.0984	0.8049±0.1311	0.7711±0.1114
SSconv [60]	0.1180±0.0711	0.1036±0.0804	0.8112±0.1305	0.7955±0.1244
TDNet [61]	0.2210±0.1042	0.1531±0.1055	0.7700±0.1491	0.6688±0.1553
PSGAN [39]	**0.0623±0.0383**	0.0769±0.0529	0.8651±0.0973	0.8672±0.0785
MDSSC-GAN [40]	0.1133±0.0905	**0.0642±0.0393**	0.8709±0.0887	0.8325±0.1092
PanDiff	0.0706±0.0379	0.0657±0.0480	**0.8855±0.0877**	**0.8697±0.0745**

TABLE VIII
QUANTITATIVE METRICS FOR ALL THE COMPARISON METHODS ON THE FULL-RESOLUTION WORLDVIEW-3 DATASET

Methods	$D_\lambda \downarrow (\pm std)$	$D_S \downarrow (\pm std)$	$QNR \uparrow (\pm std)$	$HQNR \uparrow (\pm std)$
BT-H [56]	0.1851±0.1848	0.1493±0.0937	0.7568±0.1678	0.7080±0.2157
BDSD-PC [57]	0.1505±0.1204	0.1464±0.1159	0.7901±0.1713	0.7375±0.1860
MTF-GLP-FS [16]	**0.0631±0.0579**	0.1390±0.1191	0.7833±0.1781	0.8126±0.1524
MTF-GLP-HPM-R [17]	0.0635±0.0575	0.1370±0.1179	0.7870±0.1752	0.8139±0.1507
PNN [27]	0.1251±0.1302	0.0659±0.0241	0.8396±0.1019	0.8193±0.1352
PanNet [28]	0.1862±0.1886	0.0721±0.0263	0.8292±0.1023	0.7579±0.1849
DRPNN [29]	0.1157±0.1163	0.0903±0.0712	0.8086±0.1498	0.8107±0.1528
MSDCNN [58]	0.1105±0.1119	0.0761±0.0466	0.8341±0.1297	0.8258±0.1350
DiCNN [59]	0.1160±0.1086	0.0667±0.0210	0.8374±0.1184	0.8262±0.1121
SSconv [60]	0.2021±0.2147	0.0925±0.0592	0.8021±0.1386	0.7350±0.2268
TDNet [61]	0.2116±0.2183	0.1043±0.0950	0.7961±0.1627	0.7222±0.2401
PSGAN [39]	0.1185±0.1244	0.0965±0.0779	0.8143±0.1608	0.8047±0.1686
MDSSC-GAN [40]	0.0998±0.0983	0.0883±0.0780	0.8274±0.1613	0.8268±0.1469
PanDiff	0.0982±0.1096	**0.0537±0.0467**	**0.9091±0.0742**	**0.8571±0.1336**

have obvious spectrum disparities with IMS, and DRPNN performs badly in both spectral and spatial aspects among the DL-based methods.

The full-resolution experiments conducted on QuickBird are shown in Fig. 9 and Table VII. PanDiff outperforms all competing methods in terms of QNR and HQNR evaluation metrics. However, the D_λ values of PSGAN and MTF-GLP-FS both slightly exceed that of PanDiff, while the D_S value of MDSSC-GAN is slightly lower. Compared with the traditional methods, the spatial enhancement effects of DL-based methods remain exceptional. The superior spatial learning capabilities of PanDiff, MDSSC-GAN, and DRPNN can be discerned from the clarity of magnified buildings in Fig. 9. According to the QNR and HQNR, the fusion result of TDNet is the poorest among DL-based methods, which contrasts with the GaoFen-2 fusion.

The results of the full-resolution experiments for the WorldView-3 dataset are shown in Table VIII and Fig. 10. PanDiff maintains its typical advantages over comparative methods. Nevertheless, the D_λ values of the two MRA-based methods demonstrate remarkable advantages, showcasing high spectral fidelity in the fusion of WorldView-3's real resolution data. It is worth noting that PNN has exhibited superior spatial information learning capability and fusion performance, with its evaluation metrics even surpassing many complex networks, which is in stark contrast to the two experiments mentioned above. We are also able to observe similar results as on the QuickBird data, i.e., the traditional methods are able to remain competitive with the DL-based methods in terms of D_λ but have a significant disadvantage in terms of D_S.

The results of the three full-resolution experiments in this section provide convincing evidence for the usefulness of our proposed PanDiff in real data.

C. Generalization Test

In addition to the traditional fusion methods, it is essential to evaluate the generalization capacity of the DL models discussed above. Therefore, we use WorldView-2 images to perform cross-sensor, cross-resolution generalization experiments on the model trained with WorldView-3 data. Fig. 11 shows the results of the experiment testing generalization.

The results of the generalization test indicate that PanDiff, MDSSC-GAN, PSGAN, TDNet, SSconv, and PanNet are rich in spatial information, while the other fusion methods have poor spatial details with blurred feature textures. PanDiff and PSGAN have good spectral retention, and the color tone of their fusion results is closer to that of IMS. In contrast, other fusion methods suffer from spectral distortion, where the fused images of PanNet and SSconv are lighter in color than IMS, whereas that of PNN, MSDCNN, MDSSC-GAN, DiCNN, and TDNet methods are darker. In summary, PanDiff shows high robustness with excellent spectral retention and spatial enhancement capabilities. PNN and MSDCNN have less stable networks due to their simple network structure. Moreover, the overly complex network structure of TDNet leads to training

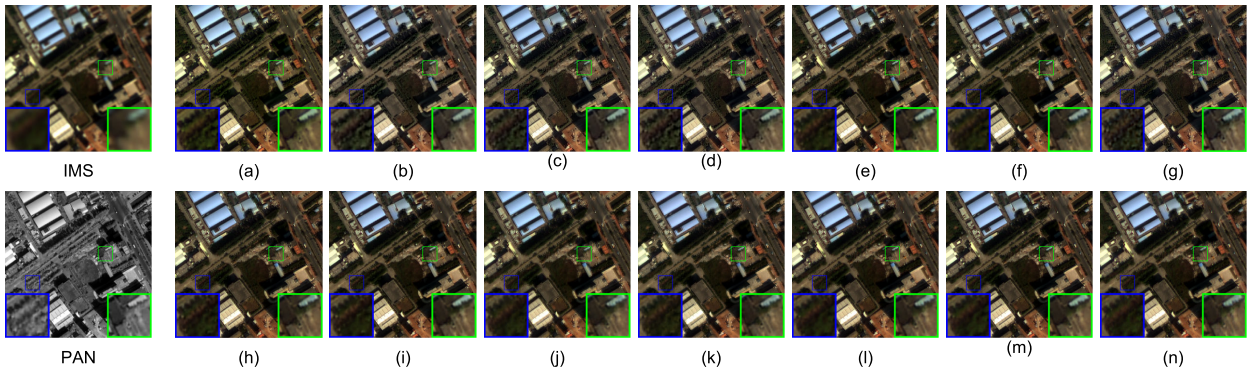

Fig. 8. Visual comparisons on a full-resolution GaoFen-2 case. (a) BH-T. (b) BDSD-PC. (c) MTF-GLP-FS. (d) MTF-GLP-HPM-R. (e) PNN. (f) PanNet. (g) DRPNN. (h) MSDCNN. (i) DiCNN. (j) SSconv. (k) TDNet. (l) PSGAN. (m) MDSSC-GAN. (n) PanDiff.

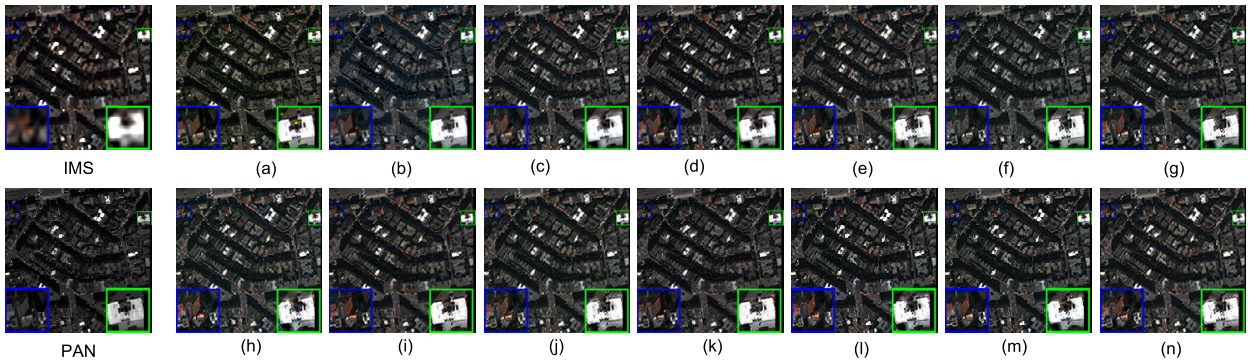

Fig. 9. Visual comparisons on a full-resolution QuickBird case. (a) BH-T. (b) BDSD-PC. (c) MTF-GLP-FS. (d) MTF-GLP-HPM-R. (e) PNN. (f) PanNet. (g) DRPNN. (h) MSDCNN. (i) DiCNN. (j) SSconv. (k) TDNet. (l) PSGAN. (m) MDSSC-GAN. (n) PanDiff.

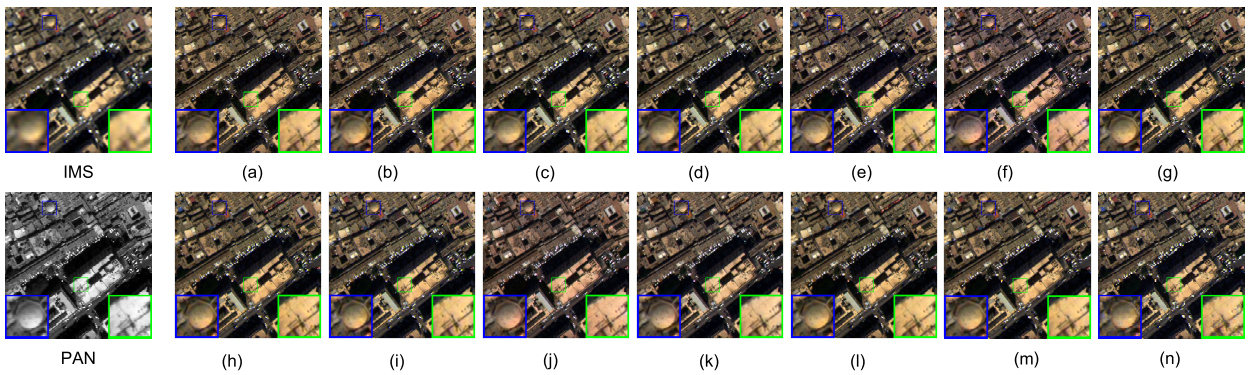

Fig. 10. Visual comparisons on a full-resolution WorldView-3 case. (a) BH-T. (b) BDSD-PC. (c) MTF-GLP-FS. (d) MTF-GLP-HPM-R. (e) PNN. (f) PanNet. (g) DRPNN. (h) MSDCNN. (i) DiCNN. (j) SSconv. (k) TDNet. (l) PSGAN. (m) MDSSC-GAN. (n) PanDiff.

overfitting, which makes TDNet exhibit a poor degree of spectral retention when processing other images.

D. Ablation Experiments of PanDiff

To demonstrate the effectiveness of the differential map-based design in overcoming the problem of high uncertainty of HRMS generated by DDPM in the field of pansharpening and to show the usefulness of the MIM, we present the results of ablation experiments on the GaoFen-2 dataset in this section.

1) Effectiveness of DM: As shown in Table IX, the introduction of differential maps can lead to a significant enhancement of model performance. Examining the outcomes of reduced- and full-resolution studies separately reveals some interesting insights. In the reduced-resolution experiments, the results are as expected; since more data information must be reconstructed, not using the DM decreases performance by 0.93, 0.0028, 0.0922, and 0.0589 for PSNR, SSIM, SAM, and ERGAS, respectively, but the drop is not excessive. However, omitting the DM significantly degrades the model's ability to retain spectral information for full-resolution images, although the difference in spatial detail retention is negligible.

2) Effectiveness of MIM: We demonstrate the validity of MIM-Spectral and MIM-Spatial based on the validation of the DM. MIM-Spectral and MIM-Spatial can improve PanDiff's spectral and spatial fidelity, respectively. As shown in

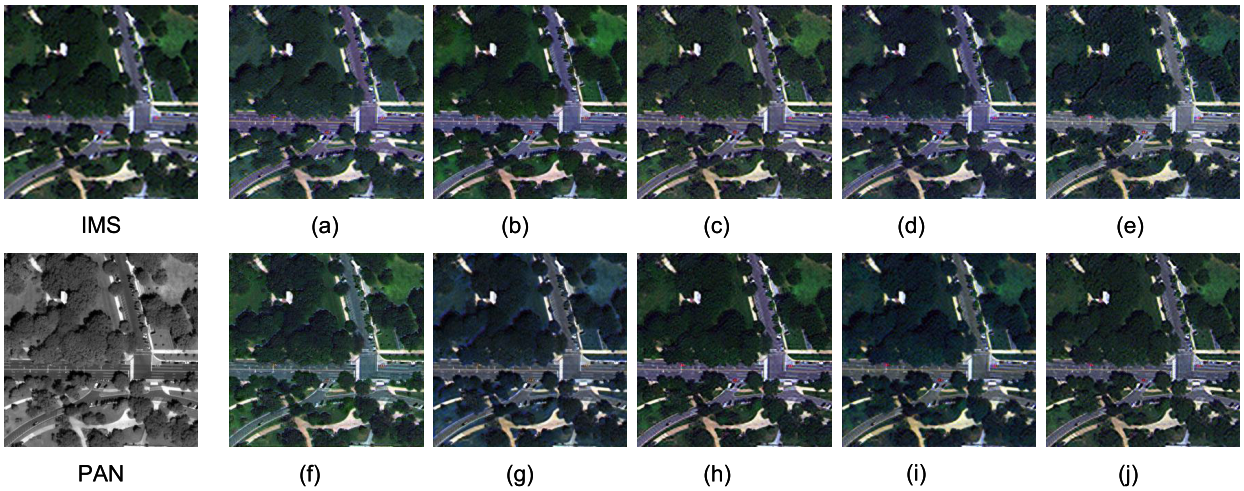

Fig. 11. Visual comparisons on a full-resolution WorldView-2 case for generalization test. (a) BH-T. (b) BDSD-PC. (c) MTF-GLP-FS. (d) MTF-GLP-HPM-R. (e) PNN. (f) PanNet. (g) DRPNN. (h) MSDCNN. (i) DiCNN. (j) SSconv. (k) TDNet. (l) PSGAN. (m) MDSSC-GAN. (n) PanDiff.

TABLE IX
QUANTITATIVE METRICS FOR PANDIFF ABLATION EXPERIMENTS ON THE GAOFEN-2 DATASET

DM	$MIM_{Spectral}$	$MIM_{Spatial}$	$PSNR\uparrow(\pm std)$	$SSIM\uparrow(\pm std)$	$SAM\downarrow(\pm std)$	$ERGAS\downarrow(\pm std)$	$D_\lambda\downarrow(\pm std)$	$D_S\downarrow(\pm std)$	$QNR\uparrow(\pm std)$
✘	✘	✘	41.88±1.13	0.9762±0.0024	0.9625±0.0681	0.8766±0.0547	0.2787±0.0824	0.0368±0.0191	0.8789±0.0451
✘	✔	✔	42.47±0.92	0.9809±0.0022	0.8657±0.0579	0.7464±0.0483	0.2574±0.0785	0.0325±0.0177	0.8859±0.0437
✔	✘	✔	42.85±0.76	0.9829±0.0016	0.8433±0.0438	0.7173±0.0409	0.0304±0.0119	0.0336±0.0136	0.9269±0.0317
✔	✔	✘	42.78±0.81	0.9795±0.0036	0.8133±0.0342	0.7175±0.0422	0.0248±0.0107	0.0394±0.0157	0.9305±0.0301
✔	✔	✔	**43.40±0.64**	**0.9837±0.0013**	**0.7735±0.0367**	**0.6875±0.0307**	**0.0223±0.0103**	**0.0323±0.0131**	**0.9396±0.0250**

TABLE X
QUANTITATIVE METRICS FOR PANDIFF WITH DIFFERENT TOTAL TIMESTEPS T ON THE GAOFEN-2 DATASET

Methods	T	$PSNR\uparrow(\pm std)$	$SSIM\uparrow(\pm std)$	$SAM\downarrow(\pm std)$	$ERGAS\downarrow(\pm std)$	$D_\lambda\downarrow(\pm std)$	$D_S\downarrow(\pm std)$	$QNR\uparrow(\pm std)$
PanDiff	10	42.03±1.01	0.9801±0.0030	0.9217±0.0647	0.8041±0.0531	0.0388±0.0163	0.0488±0.0160	0.9007±0.0496
PanDiff	100	42.96±0.92	0.9814±0.0024	0.8103±0.0522	0.7467±0.0484	0.0313±0.0147	0.0453±0.0151	0.9094±0.0431
PanDiff	500	43.37±0.79	**0.9840±0.0017**	**0.7679±0.0471**	**0.6835±0.0409**	0.0251±0.0104	0.0414±0.0134	0.9258±0.0265
PanDiff	1000	43.40±0.69	0.9838±0.0014	0.7711±0.0411	0.6860±0.0342	0.0240±0.0103	0.0363±0.0131	0.9344±0.0256
PanDiff	2000	43.40±0.64	0.9837±0.0013	0.7735±0.0367	0.6875±0.0307	0.0223±0.0103	0.0323±0.0131	0.9396±0.0250
PanDiff	3000	**43.41±0.64**	0.9837±0.0013	0.7738±0.0365	0.6877±0.0302	**0.0221±0.0102**	**0.0318±0.0130**	**0.9403±0.0248**

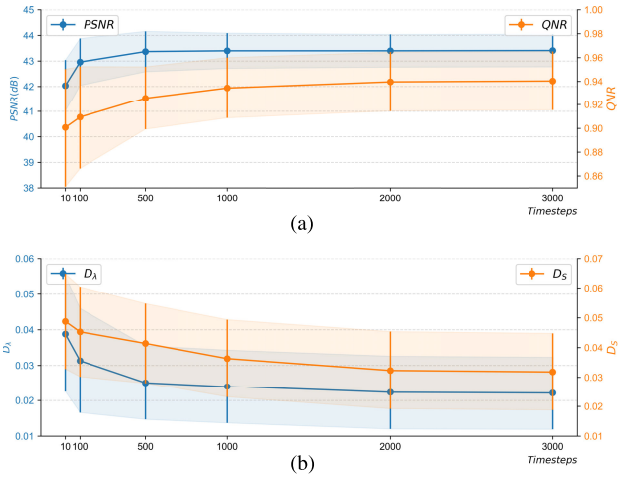

Fig. 12. Experiment results of different timesteps T. (a) PSNR and QNR versus timesteps. (b) D_λ and D_S versus T.

Table IX, PanDiff without MIM-Spectral has a substantial reduction in the spectral metrics SAM and D_λ, by 0.0698 and 0.0081, respectively; PanDiff without MIM-Spatial has a reduction in the spatial structure metrics SSIM and D_S, by 0.0042 and 0.0071, respectively.

In summary, both the design of DM and the MIM proposed in this article make significant contributions to the excellent performance of PanDiff.

E. How Timesteps Affect PanDiff

The total number T of timesteps is one of the key factors affecting the performance and operating efficiency of PanDiff. Table X shows the results of PanDiff at reduced- and full-resolution under different T values. When T decreases from 2000 to 500, the mean value of PanDiff's performance at reduced resolution is relatively stable, but the std becomes larger, which means that the stability of PanDiff performance is related to T. However, in the full-resolution experiment, both the mean value and the variance of PanDiff's performance become worse to a certain extent. Moreover, when T continues to drop to 100, the performance of PanDiff decreases significantly both at reduced resolution and full resolution, which may be related to the fact that 100 timesteps of noise addition in the forward process are not enough to turn DM into noise

TABLE XI
COMPARISON OF PARAMETERS, INFERENCE SPEED, AND PSNR (ON GAOFEN-2) OF THE COMPETING PANSHARPENING METHODS. THE DIMENSION OF TESTED IMAGE IS $2 \times 4 \times 1024 \times 1024$

Methods	Params (MB)	Inference Time (ms)	$PSNR$ (dB)
BT-H [56]	-	496.19	35.77
BDSD-PC [57]	-	1,456.67	35.27
MTF-GLP-FS [16]	-	904.03	35.94
MTF-GLP-HPM-R [17]	-	851.54	35.96
PNN [27]	0.31	165.62	39.81
PanNet [28]	0.30	44.32	39.68
DRPNN [29]	1.60	180.16	40.48
MSDCNN [58]	0.72	425.63	40.46
DiCNN [59]	0.16	22.29	39.81
SSconv [60]	5.20	306.08	40.90
TDNet [61]	1.86	472.63	39.72
PSGAN [39]	8.70	138.43	41.77
MDSSC-GAN [40]	59.18	284.86	42.55
PanDiff ($T = 10$)	45.33	666.07	42.05
PanDiff ($T = 100$)	45.33	6,655.20	42.96
PanDiff ($T = 200$)	45.33	127,865.14	43.40

conforming to an approximate Gaussian distribution. However, it is not true that the larger T, the better the experimental results. When T increases from 2000 to 3000, there is only a very weak improvement in the spatial metrics, but the spectral metrics at the reduced resolution become even slightly worse, and in short, the overall improvement in model performance is not significant.

F. Runtime Efficiency

This section discusses the runtime efficiency of the model from two perspectives: the number of parameters and inference speed. Table XI presents the basic information on the parameter count and inference speed of PanDiff compared to other methods. PanDiff exhibits a larger parameter count than most competing techniques, and due to the requirement of T timesteps sampling during inference, its speed is dramatically reduced. This results in PanDiff often taking 50–100 times longer for inference than other methods. This drawback is expected to be addressed in future research by employing interval sampling techniques to accelerate the algorithm's inference speed.

VI. CONCLUSION

In this article, a novel DDPM-based pansharpening method called PanDiff is proposed, which is the first application of DDPM to pansharpening. Our method presents two fresh perspectives with a successful case based on DDPM given in this article for pansharpening: 1) without directly learning the spatial and spectral information of HRMS images, transferring the learning objective of the pansharpening fusion network to learning the data distribution of the DM can yield outstanding results; 2) PAN and LRMS images are not required to be the object of feature extraction; rather, they are employed as injected conditions to guide the neural network modeling HRMS reconstruction procedure. Moreover, for further enhancing the guidance effect of PAN and LRMS images, the MIM is proposed. The opening benchmark dataset, including GaoFen-2, QuickBird, and WorldView-3 images, are used to evaluate the fusion quality of our PanDiff for reduced- and full-resolution fusion tests and generalization tests. Through the qualitative and quantitative comparison, it can be concluded that PanDiff excels in spatial information enhancement and spectral information fidelity with high robustness. The reasonable construction of PanDiff is well confirmed in the ablation experiments.

In the future, how to accelerate the PanDiff sampling of DM and the more efficient method to handle PAN and LRMS as injection conditions will be further studied.

REFERENCES

[1] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

[2] Q. Meng, M. Zhao, L. Zhang, W. Shi, C. Su, and L. Bruzzone, "Multilayer feature fusion network with spatial attention and gated mechanism for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[3] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Exp. Syst. Appl.*, vol. 169, May 2021, Art. no. 114417.

[4] W. Shi, Q. Meng, L. Zhang, M. Zhao, C. Su, and T. Jancsó, "DSANet: A deep supervision-based simple attention network for efficient semantic segmentation in remote sensing imagery," *Remote Sens.*, vol. 14, no. 21, p. 5399, Oct. 2022.

[5] B. Huang, B. Zhao, and Y. Song, "Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery," *Remote Sens. Environ.*, vol. 214, pp. 73–86, Sep. 2018.

[6] W. Li, L. Meng, J. Wang, C. He, G. Xia, and D. Lin, "3D building reconstruction from monocular remote sensing images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 12528–12537.

[7] J. Wang et al., "Developing a method to extract building 3D information from GF-7 data," *Remote Sens.*, vol. 13, no. 22, p. 4532, Nov. 2021.

[8] H. West, N. Quinn, and M. Horswell, "Remote sensing for drought monitoring & impact assessment: Progress, past challenges and future opportunities," *Remote Sens. Environ.*, vol. 232, Oct. 2019, Art. no. 111291.

[9] H. M. Pham, Y. Yamaguchi, and T. Q. Bui, "A case study on the relation between city planning and urban growth using remote sensing and spatial metrics," *Landscape Urban Planning*, vol. 100, no. 3, pp. 223–230, Apr. 2011.

[10] P. S. Chavez, S. C. Sides, and J. A. Anderson, "Comparison of three different methods to merge multiresolution and multispectral data: LANDSAT TM and SPOT panchromatic: ABSTRACT," *AAPG Bull.*, vol. 74, pp. 295–303, Mar. 1990.

[11] H. R. Shahdoosti and H. Ghassemian, "Combining the spectral PCA and spatial PCA fusion methods by an optimal filter," *Inf. Fusion*, vol. 27, pp. 150–160, Jan. 2016.

[12] 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.

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

[14] 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.

[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] G. Vivone, R. Restaino, and J. Chanussot, "Full scale regression-based injection coefficients for panchromatic sharpening," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3418–3431, Jul. 2018.

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

[18] J. Nunez, X. Otazu, O. Fors, A. Prades, V. Pala, and R. Arbiol, "Multiresolution-based image fusion with additive wavelet decomposition," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 3, pp. 1204–1211, May 1999.

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

[20] H. Shen, X. Meng, and L. Zhang, "An integrated framework for the spatio-temporal–spectral fusion of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7135–7148, Dec. 2016.

[21] 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.

[22] D. Fasbender, J. Radoux, and P. Bogaert, "Bayesian data fusion for adaptable image pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 6, pp. 1847–1857, Jun. 2008.

[23] T. Wang, F. Fang, F. Li, and G. Zhang, "High-quality Bayesian pansharpening," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 227–239, Jan. 2019.

[24] S. Li and B. Yang, "A new pan-sharpening method using a compressed sensing technique," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 2, pp. 738–746, Feb. 2011.

[25] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*, vol. 3361, no. 10, 1995, pp. 1–8.

[26] W. Huang, L. Xiao, Z. Wei, H. Liu, and S. Tang, "A new pan-sharpening method with deep neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 5, pp. 1037–1041, May 2015.

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

[28] 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.

[29] Y. Wei, Q. Yuan, H. Shen, and L. Zhang, "Boosting the accuracy of multispectral image pansharpening by learning a deep residual network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1795–1799, Oct. 2017.

[30] X. Liu, Q. Liu, and Y. Wang, "Remote sensing image fusion based on two-stream fusion network," *Inf. Fusion*, vol. 55, pp. 1–15, Mar. 2020.

[31] Q. Guo, S. Li, and A. Li, "An efficient dual spatial–spectral fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5412913.

[32] Z.-Q. J. Xu, Y. Zhang, T. Luo, Y. Xiao, and Z. Ma, "Frequency principle: Fourier analysis sheds light on deep neural networks," 2019, *arXiv:1901.06523*.

[33] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 5998–6008.

[34] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Represent.*, 2021, pp. 1–22.

[35] X. Meng, N. Wang, F. Shao, and S. Li, "Vision transformer for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5409011.

[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] S. Li, Q. Guo, and A. Li, "Pan-sharpening based on CNN+ pyramid transformer by using no-reference loss," *Remote Sens.*, vol. 14, no. 3, p. 624, Jan. 2022.

[38] I. J. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Montreal, QC, Canada, 2014, pp. 2672–2680.

[39] 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.

[40] A. Gastineau, J. 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.

[41] 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.

[42] 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.

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

[44] M. Zhou, J. Huang, X. Fu, F. Zhao, and D. Hong, "Effective pansharpening by multiscale invertible neural network and heterogeneous task distilling," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5411614.

[45] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.

[46] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 18187–18197.

[47] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "ILVR: Conditioning method for denoising diffusion probabilistic models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 14347–14356.

[48] C. Meng et al., "SDEdit: Guided image synthesis and editing with stochastic differential equations," in *Proc. 10th Int. Conf. Learn. Represent.*, 2022, pp. 1–33.

[49] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. 2nd Int. Conf. Learn. Represent.*, Y. Bengio and Y. LeCun, Eds. Banff, AB, Canada, 2014, pp. 1–14.

[50] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. 32nd Int. Conf. Mach. Learn.*, Lille, France, 2015, pp. 2256–2265.

[51] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[52] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Munich, Germany: Springer, 2015, pp. 234–241.

[53] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.

[54] A. Jolicoeur-Martineau, R. Piche-Taillefer, I. Mitliagkas, and R. T. Des Combes, "Adversarial score matching and improved sampling for image generation," in *Proc. 9th Int. Conf. Learn. Represent.*, 2021, pp. 1–29.

[55] L. 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.

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

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

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

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

[60] Y. Wang, L.-J. Deng, T.-J. Zhang, and X. Wu, "SSconv: Explicit spectral-to-spatial convolution for pansharpening," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4472–4480.

[61] T. Zhang, L. Deng, T. Huang, J. Chanussot, and G. Vivone, "A triple-double convolutional neural network for panchromatic sharpening," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 9, 2022, doi: 10.1109/TNNLS.2022.3155655.

[62] 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.

[63] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[64] R. H. Yuhas, A. F. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm," in *Proc. Summaries 3rd Annu. JPL Airborne Geosci. Workshop*, vol. 1, 1992, pp. 147–149.

[65] L. Wald, "Quality of high resolution synthesised images: Is there a simple criterion?" in *Proc. 3rd Conf. Fusion Earth Data, Merging Point Meas., Raster Maps Remotely Sensed Images*, 2000, pp. 99–103.

[66] J. Zhou, D. L. Civco, and J. A. Silander, "A wavelet transform method to merge Landsat TM and SPOT panchromatic data," *Int. J. Remote Sens.*, vol. 19, no. 4, pp. 743–757, Jan. 1998.
[67] G. Vivone et al., "A critical comparison among pansharpening algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2565–2586, May 2015.
[68] A. Arienzo, G. Vivone, A. Garzelli, L. Alparone, and J. Chanussot, "Full-resolution quality assessment of pansharpening: Theoretical and hands-on approaches," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 3, pp. 168–201, Sep. 2022.
[69] B. Aiazzi, L. Alparone, S. Baronti, R. Carla, A. Garzelli, and L. Santurri, "Full scale assessment of pansharpening methods and data products,' in *Proc. SPIE*, vol. 9244, Oct. 2014, Art. no. 924402.
[70] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. 7th Int. Conf. Learn. Represent.*, New Orleans, LA, USA, May 2019, pp. 1–19.

Qingyan Meng received the Ph.D. degree in ecology from Zhejiang University, Hangzhou, China, in 1999.

He is currently a Professor with the National Engineering Laboratory of Remote Sensing Satellite Applications, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. He has presided over 40 national projects, such as the China National Science and Technology Support Project, the China National Natural Science Foundation Project, the EU's Seventh Framework Project, and the National Major Science and Technology Special Project. He has conducted international cooperation research with more than ten countries. He has published more than 170 academic articles, published three books, and authorized and accepted 49 invention patents. His research interests include remote sensing of urban environment, remote sensing image analysis, and earthquake infrared remote sensing.

Dr. Meng has been selected as the South Sea Master Program of Hainan Province in 2020. He obtained 19 awards. His book titled *Urban Green Space Remote Sensing* was awarded the Hsue-Shen Tsien's Gold Award in Urbanism in 2020 and selected as the Top Ten Remote Sensing Events in China in 2020. He held an international training course titled Remote Sensing Information Processing and Urban Application for APSCO member states in 2020.

Wenxu Shi received the B.S. degree in surveying and mapping engineering from Southwest Jiaotong University, Chengdu, China, in 2020. He is currently pursuing the Ph.D. degree in cartography and GIS with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

His research interests include remote sensing, image processing, deep learning-based scene parsing for urban scenes, and building 3-D reconstruction.

Sijia Li received the B.S. degree in remote sensing science and technology from Southwest Jiaotong University, Chengdu, China, in 2020. She is currently pursuing the M.Sc. degree in signal and information processing with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

Her research interests include image processing, image fusion, and deep learning.

Linlin Zhang received the B.E. degree from Central South University, Changsha, China, in 2015, and the Ph.D. degree from the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China, in 2020.

From 2019 to 2020, she was a Visiting Ph.D. Student with Monash University, Clayton, VIC, Australia. She is currently an Assistant Professor with the National Engineering Laboratory of Remote Sensing Satellite Applications, Aerospace Information Research Institute, Chinese Academy of Sciences. Her research interests include thermal infrared remote sensing and application of quantitative remote sensing.