remote sensing

MDPI

Article

Shift-Invariant Unsupervised Pansharpening Based on Diffusion Model

Jialei Xie [1,2,3], Luyan Ji [1,3,*], Jinzhou Ye [1,2,3], Jilei Liu [4], Qi Feng [4], Kejian Liu [4] and Yongchao Zhao [1,2,3]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; xiejialei21@mails.ucas.ac.cn (J.X.)
[2] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
[3] Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, Chinese Academy of Sciences, Beijing 100190, China
[4] Institute of Remote Sensing Application in Public Security, People's Public Security University of China, Beijing 100038, China
* Correspondence: jily@mail.ustc.edu.cn

Highlights

What are the main findings?

- A shift-invariant unsupervised diffusion-based pansharpening framework is proposed.
- A translation error estimation module and loss are introduced to handle spatial offsets.

What are the implication of the main findings?

- The work establishes a new framework for training unsupervised diffusion models in pansharpening.
- The study offers new insights into achieving shift-invariant unsupervised fusion.

Abstract

Pansharpening is a crucial topic in remote sensing, and numerous deep learning-based methods have recently been proposed to explore the potential of deep neural networks (DNNs). However, existing approaches are often sensitive to spatial translation errors between high-resolution panchromatic (HRPan) and low-resolution multispectral (LRMS) images, leading to noticeable artifacts in the fused results. To address this issue, we propose an unsupervised pansharpening method that is robust to translation misalignment between HRPan and LRMS inputs. The proposed framework integrates a shift-invariant module to estimate subpixel spatial offsets and a diffusion-based generative model to progressively enhance spatial and spectral details. Moreover, a multi-scale detail injection module is designed to guide the diffusion process with fine-grained structural information. In addition, a carefully formulated loss function is established to preserve the fidelity of fusion results and facilitate the estimation of translation errors. Experiments conducted on the GaoFen-2, GaoFen-1, and WorldView-2 datasets demonstrate that the proposed method achieves superior fusion quality compared with state-of-the-art approaches and effectively suppresses artifacts caused by translation errors.

Keywords: unsupervised pansharpening; diffusion model; image fusion

Check for updates

Academic Editor: Salah Bourennane

Received: 19 October 2025
Revised: 16 December 2025
Accepted: 17 December 2025
Published: 22 December 2025

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Remote sensing applications, ranging from fine-scale land-cover mapping to urban monitoring, often require imagery that simultaneously exhibits rich spatial textures and

accurate spectral information [1–3]. However, most existing satellites typically acquire high-resolution panchromatic (HRPan) and low-resolution multispectral (LRMS) signals separately, leading to a fundamental resolution imbalance between spatial and spectral modalities. Pansharpening has emerged as an effective strategy to bridge this gap by reconstructing high-resolution multispectral (HRMS) images from the two complementary sources. Despite the long-standing interest in this task, the rapid evolution of imaging sensors, data volume, and real-world deployment demands has highlighted practical challenges that remain insufficiently addressed by conventional fusion paradigms.

Existing pansharpening approaches can be broadly categorized into two classes: traditional methods and deep learning-based (DL-based) methods. Traditional methods are typically divided into component substitution-based (CS-based) methods [4–6], multiresolution analysis-based (MRA-based) methods [7–9], and variational optimization-based (VO-based) methods [10–12]. CS-based methods can effectively preserve spatial details but often suffer from spectral distortion [13]. In contrast, MRA-based methods generally achieve better spectral fidelity, while their ability to retain fine spatial details is limited [14]. Although VO-based methods can overcome these limitations, they are typically complex and computationally expensive [7,14].

DL-based methods have achieved remarkable progress by leveraging the strong capacity of neural networks, which can be divided into two classes: supervised and unsupervised. Supervised methods typically train models on reduced-resolution images, yet their generalization to full-resolution scenarios remains inconsistent due to scale discrepancies and distribution shifts [15]. Differently, unsupervised methods train models on full-resolution datasets, to address the scale generalization difficulties in supervised methods [16–18].

A large number of DL-based unsupervised pansharpening methods have been proposed in recent years. For instance, Lin et al. [19] utilized convolutional neural networks (CNNs) to estimate the blur kernel of LRMS, effectively preserving the spectral fidelity. Zhou et al. [18] developed a cycle-consistent generative adversarial network (UCGAN) to bridge the gap between reduced and full-resolution images. Recently, diffusion models (DMs) have attracted extensive attention owing to their strong generative ability across various vision tasks. Compared with GANs, DMs not only produce more accurate results but also provide a more stable training process without the risk of mode collapse [20]. For example, Rui et al. [21] proposed an unsupervised hyperspectral pansharpening method based on a pretrained DM. However, existing unsupervised pansharpening methods based on DMs mostly depend on pretrained DMs [22,23], while the direct training of native DMs for unsupervised pansharpening from scratch remains underexplored and lacks theoretical justification.

In addition, most methods, whether supervised or unsupervised, assume perfect alignment between HRPan and LRMS. However, in practical satellite imaging, HRPan and LRMS are rarely perfectly aligned due to differences in sensor geometry, acquisition time, or preprocessing, inevitably leading to shift (translation) errors between the two images [24]. Even subpixel misalignment causes spatial details from HRPan to be injected into incorrect LRMS locations, producing ghosting edges, duplicated contours, and spectral-spatial inconsistencies in the fused images, as shown in Figure 1. Therefore, a shift-invariant pansharpening method that is robust to translation offsets between LRMS and HRPan is essential for reliable fusion, as it enables the model to compensate for unknown spatial misalignments and effectively prevent the propagation of fusion artifacts.

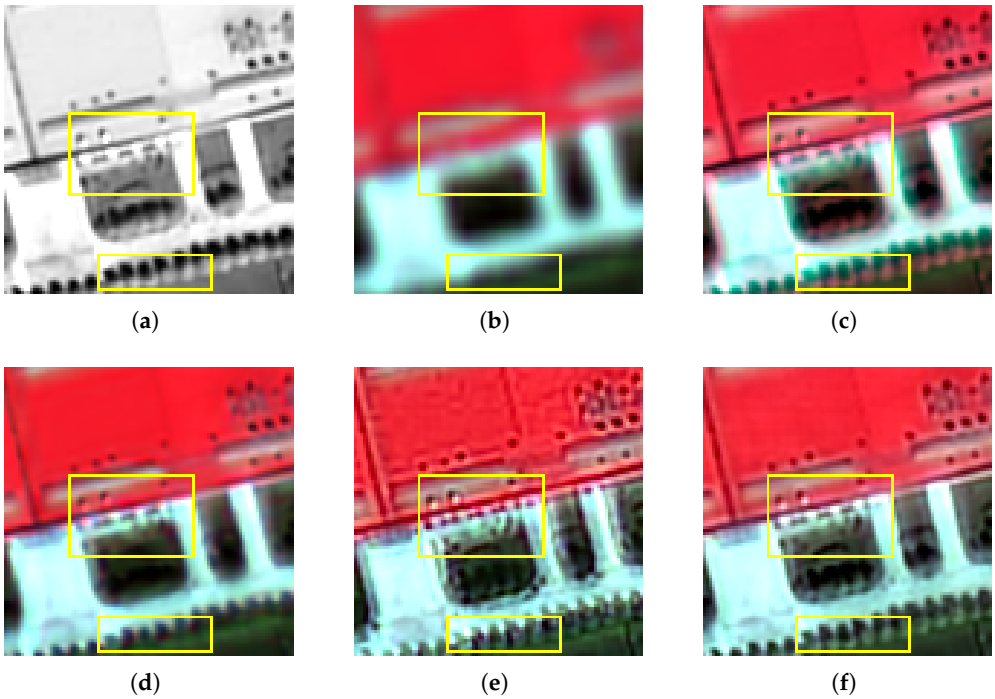


Figure 1. Fused results on the GF2 dataset. As marked by the rectangles,duplicated contours and spectral distortion appear in GSA and DDIF. In contrast, these artifacts in fusion results are eliminated in SIUPan (ours). (**a**) HRPan; (**b**) LRMS; (**c**) GSA; (**d**) DDIF; (**e**) λ-PNN; (**f**) SIUPan.

Several studies have recognized this issue and proposed corresponding solutions [16,25–28]. For example, Lee et al. [25] introduced a feature alignment module to generate an adjusted LRMS that better matches the HRPan input. However, this adjusted LRMS is only used as input to the fusion network, while the spectral loss computation still relies on multiple spatially shifted LRMS versions. Similarly, Ciotola et al. [16] estimated the offset between HRPan and LRMS by measuring their correlation across several spatial shifts before training. During loss computation, the LRMS is then shifted according to the estimated offset. Although these strategies improve fusion results compared with those ignoring translation errors, they still suffer from several limitations: (1) redundant computation, as multiple LRMS shifts are required; (2) the need to predefine a limited range of possible displacements; and (3) the offset-adjusted LRMS is not directly involved in loss optimization, limiting the effectiveness of these correction strategies.

To address the above limitations, first, we present a framework for training DMs in the context of unsupervised pansharpening and provide a mathematical proof to support their effectiveness. To better guide the diffusion process, we design a Multiscale Detail Injection module (MSDI) that progressively injects fine-grained spatial and spectral information during generation. Second, to achieve robustness against translation errors, we develop a Translation Error Estimation Module (TEM) together with a dedicated translation estimation loss, which adaptively estimates subpixel offsets between HRPan and LRMS. This design effectively enhances the spatial consistency and suppresses misalignment-induced fusion artifacts. Based on these designs, we propose Shift-Invariant Unsupervised Pansharpening based on Diffusion Model (SIUPan). The main contributions of our work are summarized as follows:

1. We propose a training framework for diffusion models in unsupervised pansharpening that directly trains a native DM in an unsupervised setting. We design an MSDI to guide the diffusion process by injecting hierarchical spatial and spectral details, ensuring a balanced improvement in both spatial and spectral fidelity.

2. We develop a TEM to adaptively estimate subpixel translation offsets between HRPan and LRMS, enabling shift-invariant fusion without redundant computations or predefined displacement ranges.
3. A new translation error estimation loss function is developed to guide the TEM estimate the translation error. A tailored loss function is constructed to jointly maintain spectral fidelity and spatial sharpness.

In the rest of this article, we describe the proposed method in Section 2. In Section 3, we conduct extensive experiments to evaluate the performance of our method. Finally, a brief conclusion is presented in Section 4.

2. Methodology

2.1. Overall Network Architecture

As shown in Figure 2, SIUPan consists of two primary modules: (1) a TEM for estimating the displacements and (2) a diffusion-based fusion module for image fusion. Let $\mathbf{I}_{\text{pan}} \in \mathbb{R}^{W\times H\times 1}$, $\mathbf{I}_{\text{lrms}} \in \mathbb{R}^{w\times h\times C}$, and $\mathbf{I}_{\text{hrms}} \in \mathbb{R}^{W\times H\times C}$ represent the corresponding HRPan, LRMS, and HRMS, respectively, where $W(w)$ and $H(h)$ are the width and height of HRPan (LRMS); C is the number of bands of LRMS. LRMS is upsampled to obtain the upsampled LRMS (MS), denoted as $\mathbf{I}_{\text{ms}} \in \mathbb{R}^{W\times H\times C}$. An 11th-order polynomial interpolator [29,30] is used to upsample LRMS. In SIUPan, $\mathbf{I}_{\text{pan}}$ and $\mathbf{I}_{\text{ms}}$ are firstly fed into the TEM to obtain the displacements $(\Delta m^{\theta}, \Delta n^{\theta})$. Next, we resample $\mathbf{I}_{\text{ms}}$ to obtain the shifted image $\mathbf{I}^{\theta}_{\text{Rms}}$. In the fusion module, we first define an initial estimate of HRMS, denoted as $\hat{\mathbf{I}}_{\text{hrms}} = \mathbf{I}^{C}_{\text{pan}}$. Here, $\mathbf{I}^{C}_{\text{pan}}$ represents $\mathbf{I}_{\text{pan}}$ stacked C times along band axis. The input residual image $\hat{\mathbf{I}}_{\text{res}} = \hat{\mathbf{I}}_{\text{hrms}} - \mathbf{I}^{\theta}_{\text{Rms}}$ is then added to Gaussian noise ϵ_t and fed into the diffusion model. The diffusion model with MSDI generates the residual image $\mathbf{I}^{\theta}_{\text{res}}$, which is added to $\mathbf{I}^{\theta}_{\text{Rms}}$ to obtain the fused image $\mathbf{I}^{\theta}_{\text{hrms}} = \mathbf{I}^{\theta}_{\text{Rms}} + \mathbf{I}^{\theta}_{\text{res}}$. Some notations used in the following discussion are listed in Table 1.

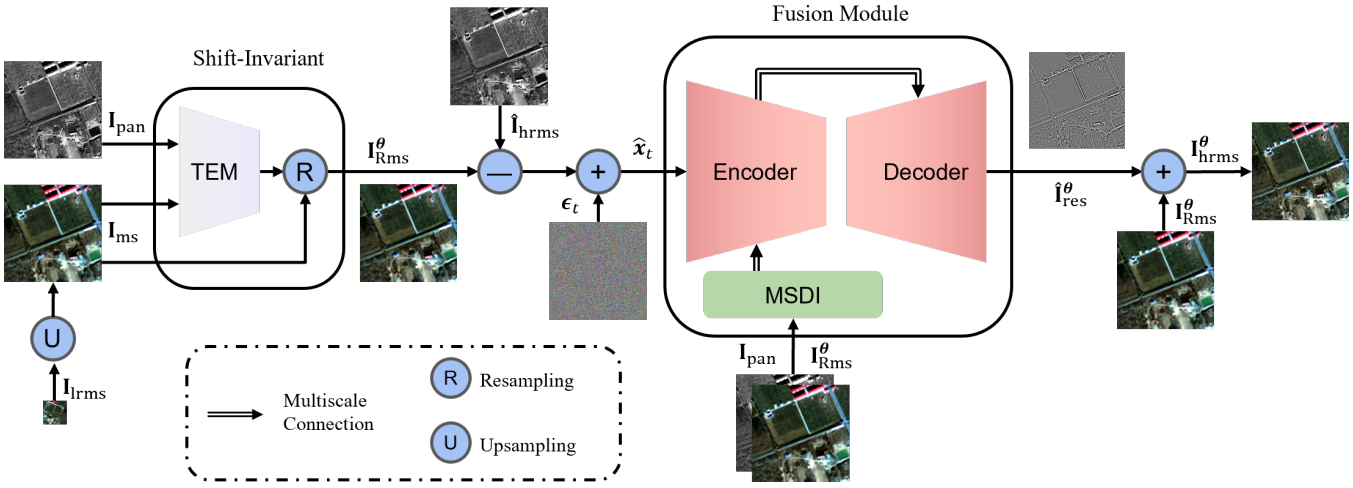


Figure 2. Overall flowchart of the proposed method. The $\mathbf{I}_{\text{pan}}$ and $\mathbf{I}_{\text{ms}}$ are first fed into TEM to obtain the displacements between them. The displacements are utilized to resample $\mathbf{I}_{\text{ms}}$ to obtain $\mathbf{I}^{\theta}_{\text{Rms}}$, which is employed for fusion and loss calculation. $\hat{\mathbf{I}}^{\theta}_{\text{res}}$ is generated by the diffusion model with MSDI. Fusion result, $\mathbf{I}^{\theta}_{\text{hrms}}$, is obtained by adding $\hat{\mathbf{I}}^{\theta}_{\text{res}}$ and $\mathbf{I}^{\theta}_{\text{Rms}}$ together.

2.2. Diffusion Model

Similar to most diffusion models [20,31–34], we utilize a U-Net as the backbone of our denoising network. Here, the SR3 [32] architecture is employed. We modify both the forward and reverse processes of diffusion to adapt it for unsupervised pansharpening.

Additionally, MSDI is designed to inject spatial details and spectral information into U-Net for better fusion results.

Table 1. The frequently used notations.

Notation	Size	Description
$\mathbf{I}_{\text{lrms}}$	$w \times h \times C$	Multi-spectral image
$\mathbf{I}_{\text{pan}}$	$W \times H \times 1$	Panchromatic image
$\mathbf{I}_{\text{ms}}$	$W \times H \times C$	Upsampled multi-spectral image
$\mathbf{I}_{\text{Rms}}$	$W \times H \times C$	Perfectly registered $\mathbf{I}_{\text{ms}}$
$\mathbf{I}^{\theta}_{\text{Rms}}$	$W \times H \times C$	Translation corrected $\mathbf{I}_{\text{ms}}$
$\mathbf{I}_{\text{hrms}}$	$W \times H \times C$	Fused multi-spectral image we desire
$\hat{\mathbf{I}}_{\text{hrms}}$	$W \times H \times C$	Estimate of $\mathbf{I}_{\text{hrms}}$
$\hat{\mathbf{I}}^{\theta}_{\text{hrms}}$	$W \times H \times C$	The output image of SIUPan
$\mathbf{I}_{\text{res}}$	$W \times H \times C$	$\mathbf{I}_{\text{hrms}} - \mathbf{I}_{\text{Rms}}$
$\hat{\mathbf{I}}_{\text{res}}$	$W \times H \times C$	$\hat{\mathbf{I}}_{\text{hrms}} - \mathbf{I}^{\theta}_{\text{Rms}}$
$\hat{\mathbf{I}}^{\theta}_{\text{res}}$	$W \times H \times C$	$\hat{\mathbf{I}}^{\theta}_{\text{hrms}} - \mathbf{I}^{\theta}_{\text{Rms}}$

2.2.1. Unsupervised Diffusion Model

In Denoising Diffusion Probabilistic Models (DDPM) [35], the forward process is defined on samples drawn from the true data distribution. However, in unsupervised pansharpening, the true data

$$x_0 = \mathbf{I}_{\text{res}} = \mathbf{I}_{\text{hrms}} - \mathbf{I}_{\text{Rms}} \tag{1}$$

is unavailable. To address this challenge, we adopt a proxy estimate

$$\hat{x}_0 = \hat{\mathbf{I}}_{\text{res}} = \hat{\mathbf{I}}_{\text{hrms}} - \mathbf{I}^{\theta}_{\text{Rms}}, \tag{2}$$

where $\hat{\mathbf{I}}_{\text{hrms}}$ is an initial HRMS estimate, which is set as the stacked HRPan along the band axis in our work. The relationship between x_0 and $\hat{x}_0$ can be described as

$$x_0 = \hat{x}_0 + z, \tag{3}$$

where z is the difference between x_0 and $\hat{x}_0$. We assume that z follows a Gaussian distribution

$$p(z) = \mathcal{N}(z; 0, \delta^2 \mathbf{I}), \tag{4}$$

where $\mathcal{N}$ means the Gaussian distribution, and δ^2 is the variance. Thus, $\hat{x}_0$ is a Gaussian-perturbed version of x_0. This assumption allows us to define a valid diffusion forward process without access to the true data distribution.

The forward process can be expressed as

$$p(\hat{x}_t | \hat{x}_{t-1}) = \mathcal{N}(\hat{x}_t; \sqrt{1-\beta_t}\hat{x}_{t-1}, \beta_t \mathbf{I}), \tag{5}$$

where $t \in [1:T]$, T is the number of timesteps, and β_t is a pre-defined variance. In forward process, we can sample $\hat{x}_t$ from $\hat{x}_0$ in a closed form:

$$p(\hat{x}_t | \hat{x}_0) = \mathcal{N}(\hat{x}_t; \sqrt{\overline{\alpha}_t}\hat{x}_0, (1-\overline{\alpha}_t)\mathbf{I}), \tag{6}$$

where $\overline{\alpha}_t = \prod_{i=0}^{t} \alpha_i$ and $\alpha_i = 1 - \beta_i$. Substituting (3) into (6), we obtain:

$$\hat{x}_t = \sqrt{\overline{\alpha}_t} x_0 + \eta_t, \quad \eta_t \sim \mathcal{N}(0, (1 - \overline{\alpha}_t + \delta^2 \overline{\alpha}_t)\mathbf{I}), \tag{7}$$

indicating that using $\hat{x}_0$ simply modifies the variance of the diffusion kernel but preserves the trajectory structure. The pseudocode of the forward process is given in Algorithm 1.

Algorithm 1 The Forward Process.

Input: the number of timestep, T, the estimate of x_0, $\hat{x}_0$
Output: the noised image, $\hat{x}_T$
1: **for** $t = 1 : T$ **do**
2: select β_t
3: sample $\hat{x}_t$ according to (6)
4: **end for**

In the reverse process, our network predicts x_0 directly rather than noise. Thus, we can sample $\hat{x}_t$ by the conditional distribution $p(\hat{x}_t|x_0)$, which also follows a Gaussian distribution

$$p(\hat{x}_t|x_0) = \mathcal{N}(\hat{x}_t; \sqrt{\overline{\alpha}_t}x_0, (1 - \overline{\alpha}_t + \delta^2\overline{\alpha}_t)\mathbf{I}). \tag{8}$$

Subsequently, we iteratively update x_0 using the sampled $\hat{x}_t$ until the sampling process is finished. The pseudocode of the reverse process is given in Algorithm 2. The forward and reverse processes of diffusion model in our work are illustrated in Figure 3.

Algorithm 2 The Reverse Process.

Input: the number of timestep, T, Gaussian noise image, $\hat{x}_T$, the network, f^θ
Output: sampled image, x_0
1: **for** $t = T : 1$ **do**
2: $x_0 = f^\theta(\hat{x}_t, t)$
3: sample $\hat{x}_{t-1}$ according to (8)
4: **end for**
5: $t = t - 1$
6: $x_0 = f^\theta(\hat{x}_0, t)$

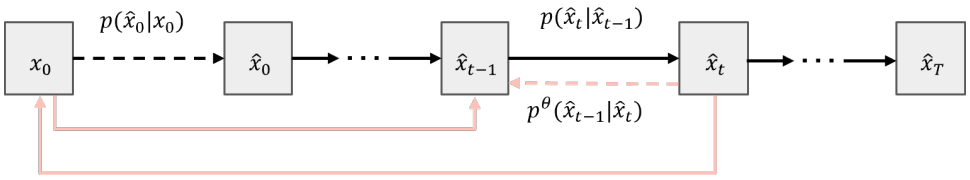


Figure 3. The forward process and reverse process of diffusion model in our work. In the forward process (indicated by the black solid arrows), $\hat{x}_0$ is used to sample $\hat{x}_t$. In the reverse process (indicated by the pink solid arrows), our network predicts x_0 directly, and the generated x_0 is used to sample $\hat{x}_{t-1}$.

Since our goal is to obtain x_0, the training objective is to maximize the log likelihood of $p(x_0)$, i.e., $\ln p(x_0)$,

$$\max \ln p(x_0) = \max \ln \int p(x_0, \hat{x}_{0:T}) d\hat{x}_{0:T}. \tag{9}$$

However, we cannot maximize it directly because of the latent variables ($\hat{x}_{0:T}$) in it. Therefore, we use its Evidence Lower Bound (ELBO), which can be written as

$$\ln p(x_0) \propto - \|x_0 - x_0^\theta\|_2^2, \tag{10}$$

where x_0^θ is the output of our network. The detailed proof of (10) is provided in Appendix A. Let $\Delta e = \|x_0 - x_0^\theta\|_2^2$; then, the objective function in (9) can be converted to

$$\min \Delta e = \|x_0 - x_0^\theta\|_2^2. \tag{11}$$

Clearly, there is no $\mathbf{I}_{\text{res}}$ ($\mathbf{I}_{\text{hrms}}$) available in unsupervised pansharpening. Fortunately, we have two transformations of $\mathbf{I}_{\text{hrms}}$ ($\mathbf{I}_{\text{pan}}$ and $\mathbf{I}^{\theta}_{\text{Rms}}$) that can be used for training,

$$\mathbf{I}_{\text{pan}} = \mathcal{T}_{\text{pan}}(\mathbf{I}_{\text{hrms}}), \quad \mathbf{I}^{\theta}_{\text{Rms}} = \mathcal{T}_{\text{Rms}}(\mathbf{I}_{\text{hrms}}), \tag{12}$$

where $\mathcal{T}_{\text{pan}}$ ($\mathcal{T}_{\text{Rms}}$) represent the transformation function. Moreover, we modify the objective function (11) into

$$\min \Delta e_{\text{pan}} + \Delta e_{\text{Rms}}, \tag{13}$$

where $\Delta e_{\text{pan}} = \mathrm{f}_{\text{pan}}(\mathcal{T}_{\text{pan}}(\mathbf{I}^{\theta}_{\text{Rms}} + \mathbf{I}^{\theta}_{\text{res}}) - \mathbf{I}_{\text{pan}})$, and $\Delta e_{\text{Rms}} = \mathrm{f}_{\text{Rms}}(\mathcal{T}_{\text{Rms}}(\mathbf{I}^{\theta}_{\text{Rms}} + \mathbf{I}^{\theta}_{\text{res}}) - \mathbf{I}^{\theta}_{\text{Rms}})$. f_{Rms} and f_{Rms} are the spatial and spectral loss functions, respectively, which will be given in the following.

In fact, we can employ the sample solvers that are compatible with DDPM, as the structure of our DM is consistent with that of DDPM. The Denoising Diffusion Implicit Model sample solver [36] is used in our work.

2.2.2. MSDI

Many studies have shown that incorporating conditional information in DMs can enhance generation quality [20,31,37,38]. In pansharpening, we naturally have two prior conditions, HRPan and LRMS, which are leveraged to provide conditional information. HRPan contains rich spatial details but lacks spectral information, while LRMS provides abundant spectral information but has fewer distinct details. Therefore, to balance the spatial details and spectral information, we design the MSDI.

As shown in Figure 4, the MSDI is a multi-layer-structure module, where each layer consists of two cross-attention blocks and two downsampling blocks. Cross-attention blocks are employed to enhance the spectral information of HRPan features and improve the spatial details of LRMS features. Downsampling blocks are used to obtain feature maps at multiple scales.

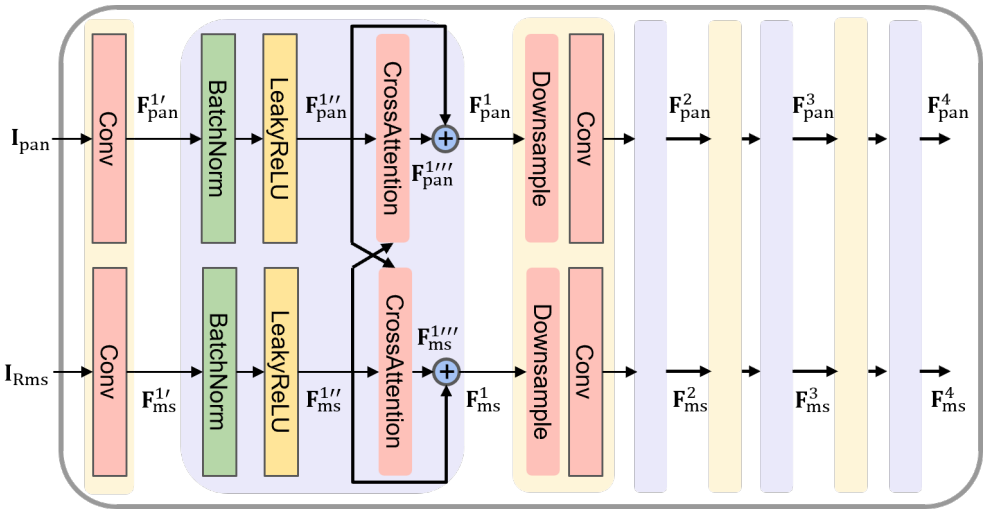


Figure 4. The architecture of the MSDI module. Each layer of the MSDI consists of four blocks: two cross-attention blocks and two downsample blocks. The cross-attention blocks are used to exploit the potential information from HRPan and MS, and the downsampling blocks are used to reduce the size of the feature map.

For the i-th layer of the MSDI (except the first layer), the inputs, $\mathbf{F}^{i-1}_{\text{pan}}$ and $\mathbf{F}^{i-1}_{\text{ms}}$, are first downsampled to reduce the size of the feature maps, respectively. Convolution layers are followed to control the depth of the feature map,

$$\mathbf{F}_{\text{pan}}^{i'} = \text{Conv}(\text{Downsample}(\mathbf{F}_{\text{pan}}^{i-1})),$$
$$\mathbf{F}_{\text{ms}}^{i'} = \text{Conv}(\text{Downsample}(\mathbf{F}_{\text{ms}}^{i-1})).$$

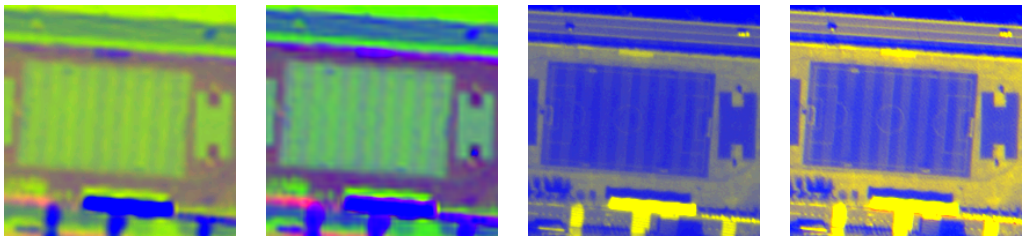
Sequentially, $\mathbf{F}_{\text{ms}}^{i''}$ and $\mathbf{F}_{\text{pan}}^{i''}$ are obtained after BatchNormalization and LeakyReLU. Two cross-attentions are employed to enhance the two feature maps, which are formulated as

$$\mathbf{F}_{\text{pan}}^{i'''} = \text{E}(\mathbf{F}_{\text{pan}}^{i''}, \mathbf{F}_{\text{ms}}^{i''}),$$
$$\mathbf{F}_{\text{ms}}^{i'''} = \text{E}(\mathbf{F}_{\text{ms}}^{i''}, \mathbf{F}_{\text{pan}}^{i''}),$$

where $\text{E}(\text{input}, \text{context})$ is the linear complexity cross-attention [39]. The output of the i-th layer is

$$\mathbf{F}_{\text{pan}}^{i} = \mathbf{F}_{\text{pan}}^{i''} + \mathbf{F}_{\text{pan}}^{i'''},$$
$$\mathbf{F}_{\text{ms}}^{i} = \mathbf{F}_{\text{ms}}^{i''} + \mathbf{F}_{\text{ms}}^{i'''}.$$

As shown in Figure 5, $\mathbf{F}_{\text{ms}}^{1'}$, the input feature, is blurred, and $\mathbf{F}_{\text{pan}}^{1'}$ has low contrast. After cross-attention blocks, the sharpness in $\mathbf{F}_{\text{ms}}^{1}$ is enhanced, and the contrast in $\mathbf{F}_{\text{pan}}^{1}$ is improved. Finally, $\mathbf{F}_{\text{ms}}^{i}$ and $\mathbf{F}_{\text{pan}}^{i}$ are injected into the encoder to guide image generation.



(**a**) (**b**) (**c**) (**d**)

Figure 5. False-color images of feature maps. (**a**) $\mathbf{F}_{\text{ms}}^{1'}$. (**b**) $\mathbf{F}_{\text{ms}}^{1}$. (**c**) $\mathbf{F}_{\text{pan}}^{1'}$. (**d**) $\mathbf{F}_{\text{pan}}^{1}$.

2.3. Translation Error Estimation Module

The translation errors between $\mathbf{I}_{\text{pan}}$ and $\mathbf{I}_{\text{ms}}$, can be described as

$$[m_{\text{pan}}\ n_{\text{pan}}]^{\text{T}} = \mathcal{T}_{\text{R}}([m_{\text{ms}}\ n_{\text{ms}}]^{\text{T}}) = \mathbf{A}[m_{\text{ms}}\ n_{\text{ms}}\ 1]^{\text{T}}, \tag{14}$$

where $[m_{\text{pan}}\ n_{\text{pan}}]$ and $[m_{\text{ms}}\ n_{\text{ms}}]$ are the coordinates in $\mathbf{I}_{\text{pan}}$ and $\mathbf{I}_{\text{ms}}$, respectively. $\mathbf{A}$ is an affine transformation matrix, which can be expressed as

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & \Delta m \\ 0 & 1 & \Delta n \end{bmatrix}. \tag{15}$$

However, Δm and Δn are unknown. Thus, the TEM is designed to estimate them.

As shown in Figure 6, the TEM is comprised of two structurally identical Feature Extractors and one Feature Integration Block. In the Feature Extractor, the inputs, $\mathbf{I}_{\text{pan}}$ and $\mathbf{I}_{\text{ms}}$, are mapped into feature domain via a convolutional layer. Considering our goal is to shift $\mathbf{I}_{\text{ms}}$ to match $\mathbf{I}_{\text{pan}}$, and they are heterogenous images (with significant style differences), we concatenate $\mathbf{I}_{\text{ms}}$ with $\mathbf{I}_{\text{pan}}$ at the input of MS Feature Extractor, which provides $\mathbf{I}_{\text{ms}}$ with style information from $\mathbf{I}_{\text{pan}}$. The Feature Integration Block is composed of a MidBlock, three DownBlocks, a convolutional layer, an activation function, and a global average pooling. The MidBlock, containing two convolutional layers and two activation functions,

integrates the two extracted features. Inspired by [40], three cascaded DownBlocks are employed to extract the difference between $\mathbf{I}_{\text{pan}}$ and $\mathbf{I}_{\text{ms}}$ in the multiscale feature domain. Each DownBlock applies average pooling with a 3×3 kernel and stride 2 to downsample the feature maps. After a global average pooling, an estimated displacement $[\Delta m^{\theta}\ \Delta n^{\theta}]$ is obtained.

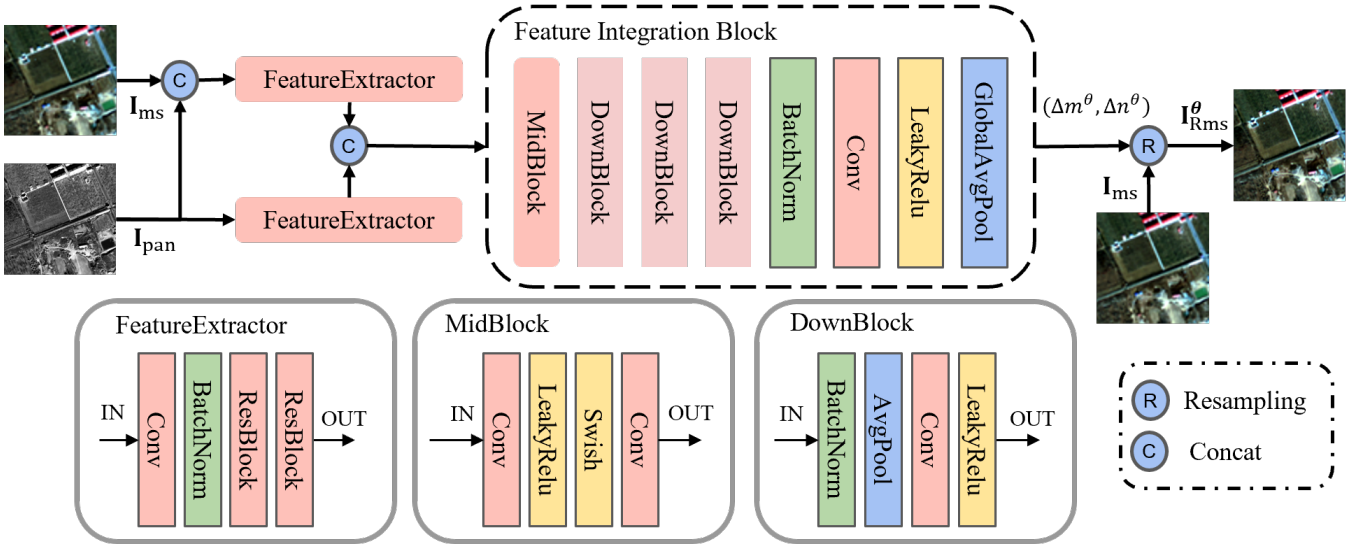


Figure 6. The architecture of TEM. Two feature extractors map $\mathbf{I}_{\text{pan}}$ and $\mathbf{I}_{\text{ms}}$ into the feature domain. A feature integration block, consisting of a MidBlock, three DownBlocks, a convolutional layer, an activation function, and a global average pooling, extracts their differences and outputs the displacement vector $[\Delta m^{\theta}, \Delta n^{\theta}]$.

Our objective is to obtain $\mathbf{I}_{\text{Rms}}$, which has the same coordinates as $\mathbf{I}_{\text{pan}}$. Then, according to (14), we can get the estimated coordinates of $\mathbf{I}_{\text{Rms}}$ as

$$[m^{\theta}_{\text{Rms}}\ n^{\theta}_{\text{Rms}}]^{\text{T}} = \begin{bmatrix} 1 & 0 & \Delta m^{\theta} \\ 0 & 1 & \Delta n^{\theta} \end{bmatrix} [m_{\text{ms}}\ n_{\text{ms}}\ 1]^{\text{T}}. \tag{16}$$

Then, the estimated $\mathbf{I}_{\text{Rms}}$ (noted as $\mathbf{I}^{\theta}_{\text{Rms}}$) can be resampled by

$$\mathbf{I}^{\theta}_{\text{Rms}}(m^{\theta}_{\text{Rms}}, n^{\theta}_{\text{Rms}}, c) = \sum_{i}^{H} \sum_{j}^{W} \mathbf{I}_{\text{ms}}(i, j, c) k(m_{\text{ms}} - i) k(n_{\text{ms}} - j), \tag{17}$$

where $k(\cdot)$ is the generic sampling kernel, and c is the band number. Here, the bilinear sampling kernel is used. Moreover, a translation error estimation loss is designed to achieve more accurate estimation of the spatial offsets. Since the true displacements $[\Delta m, \Delta n]$ are unknown, it is infeasible to directly minimize the differences between $[\Delta m, \Delta n]$ and $[\Delta m^{\theta}, \Delta n^{\theta}]$. Instead, we construct the loss function based on $\mathbf{I}_{\text{pan}}$ and $\mathbf{I}^{\theta}_{\text{Rms}}$, as detailed in Section 2.4.1.

2.4. Loss Function

In SIUPan, the loss consists of two terms: the translation error estimation loss and the fusion loss, which is formulated as

$$L = w_1 L_{\text{R}} + L_{\text{F}}, \tag{18}$$

where L_{R} is the translation error estimation loss, L_{F} is the fusion loss, and w_1 is the trade-off parameter.

2.4.1. Translation Error Estimation Loss

The objective of TEM is to make $[\Delta m^{\theta} \ \ \Delta n^{\theta}]$ as close as possible to $[\Delta m \ \ \Delta n]$. However, since the GT offsets are unavailable, we redefine the objective as evaluating whether $\mathbf{I}^{\theta}_{\text{Rms}}$ is properly aligned with $\mathbf{I}_{\text{pan}}$. According to [41–44], if two images are perfectly matched, their phase difference is zero in frequency domain. Thus, we formulate the translation error estimation loss function as the phase difference between $\mathbf{I}_{\text{pan}}$ and $\mathbf{I}^{\theta}_{\text{Rms}}$

$$L_{\text{R}} = \text{f}_{\text{R}}(\mathbf{I}_{\text{pan}}, \mathbf{I}^{\theta}_{\text{Rms}}), \tag{19}$$

where f_{R} is the translation error estimation loss function. Assuming that two images $\mathbf{X} \in \mathbb{R}^{H \times W}$ and $\mathbf{Y} \in \mathbb{R}^{H \times W}$ have displacements $[\Delta m' \ \ \Delta n']$, their phase difference in frequency domain can be estimated using normalized cross-power spectrum (NCPS), noted as $\mathbf{Q}(u, v)$

$$\mathbf{Q}(u, v) = \frac{\widetilde{\mathbf{X}}(u, v)\widetilde{\mathbf{Y}}(u, v)}{|\widetilde{\mathbf{X}}(u, v)\widetilde{\mathbf{Y}}(u, v)|}, \tag{20}$$

where $\widetilde{\mathbf{X}}(u, v)$ and $\widetilde{\mathbf{Y}}(u, v)$ are Fourier transforms of $\mathbf{X}(m, n)$ and $\mathbf{Y}(m, n)$, respectively. However, NCPS is sensitive to noise [43]. To improve its robustness, Zhu et al. [43] proposed an autocorrelated version, termed ANCPS, which can be formulated as

$$\begin{aligned} \mathbf{R}(\mu, \nu) &= \mathbf{Q}(u, v)\mathbf{Q}^{*}(u - \mu, v - \nu) \\ &= e^{j2\pi(u\Delta m'/H + v\Delta n'/W)}, \end{aligned} \tag{21}$$

where $\mathbf{R}(\mu, \nu)$ is ANCPS, and $\mathbf{Q}^{*}$ is the complex conjugate of $\mathbf{Q}$. From (21), we can find the displacements are embedded in the phase of ANCPS. Thus, we can use the ANCPS of $\mathbf{I}_{\text{pan}}$ and $\mathbf{I}^{\theta}_{\text{Rms}}$ to estimate their displacements. For simplification, we also denote the displacements and ANCPS of $\mathbf{I}_{\text{pan}}$ and $\mathbf{I}^{\theta}_{\text{Rms}}$ as $[\Delta m' \ \Delta n']$ and $\mathbf{R}(\mu, \nu)$, respectively. To ensure proper alignment in both directions, we must compute $\Delta m'$ and $\Delta n'$ separately. Thus, we have

$$\begin{aligned} p_x(\mu, \nu) &= \mathbf{R}(\mu, \nu)\mathbf{R}^{*}(\mu - 1, \nu) = e^{j2\pi\Delta m'/H}, \\ p_y(\mu, \nu) &= \mathbf{R}(\mu, \nu)\mathbf{R}^{*}(\mu, \nu - 1) = e^{j2\pi\Delta n'/W}, \end{aligned} \tag{22}$$

where $\mathbf{R}^{*}$ is the complex conjugate of $\mathbf{R}$. From (22), we can obtain $\Delta m'$ and $\Delta n'$ by p_x and p_y, respectively. To enhance the robustness, we use r pixels to calculate their average phase. Thus, the translation error estimation loss is formulated as

$$L_{\text{R}} = \frac{1}{r}\sum_{i}^{r} |\text{angle}(p_x(\mu_i, \nu_i))| + |\text{angle}(p_y(\mu_i, \nu_i))|, \tag{23}$$

where angle means the extracting phase.

2.4.2. Fusion Loss

SIUPan incorporates two categories of fusion loss functions: the spatial loss function (i.e., f_{pan} in Section 2.2) and the spectral loss function (i.e., f_{Rms} in Section 2.2). Specifically, the spatial loss consists of one term (L_{spa}), while the spectral loss includes two terms (L_{spe1} and L_{spe2}). Therefore, the fusion loss is expressed as

$$L_{\text{F}} = w_2 L_{\text{spa}} + w_3 L_{\text{spe1}} + w_4 L_{\text{spe2}}, \tag{24}$$

where w_2, w_3, and w_4 are trade-off parameters.

L_{spa} is designed to preserve the spatial details. According to refs. [28,45], the correlation coefficient between $\mathbf{I}_{pan}$ and $\mathbf{I}^{\theta}_{hrms}$ is used to measure the spatial distortion. The correlation coefficient between two images $\mathbf{X}$ and $\mathbf{Y}$ is defined as

$$\mathrm{CORR}(\mathbf{X},\mathbf{Y}) = \frac{\sigma_{\mathbf{XY}}}{\sigma_{\mathbf{X}}\sigma_{\mathbf{Y}}}, \tag{25}$$

where $\sigma_{\mathbf{X}}$, $\sigma_{\mathbf{Y}}$, and $\sigma_{\mathbf{XY}}$ are the variances and covariance of $\mathbf{X}$ and $\mathbf{Y}$. It is obviously unreasonable to assume that each band of $\mathbf{I}^{\theta}_{hrms}$ would have a high correlation with $\mathbf{I}_{pan}$. Therefore, the correlation coefficient between degraded $\mathbf{I}_{pan}$ and $\mathbf{I}^{\theta}_{Rms}$ is taken as a threshold mask. Moreover, a $s \times s$ window is used to calculate the local correlation coefficient. The spatial loss is formulated as

$$L_{spa} = \frac{1}{CS}\sum_{c}^{C}\sum_{s}^{S}((1-\rho^{s}(c))u(\rho^{s}_{m}-\rho^{s}(c))), \tag{26}$$

where $\rho^{s}(c) = \mathrm{CORR}(\mathbf{I}^{s}_{pan}, \mathbf{I}^{\theta s}_{hrms}(c))$, and $\rho^{s}_{m}(c) = \mathrm{CORR}(\mathrm{G}(\mathbf{I}^{s}_{pan}), \mathbf{I}^{\theta s}_{Rms}(c))$. C is the number of bands, and S is the number of windows. $\mathrm{G}(\cdot)$ represents the Gaussian degrading operation, and $u(\cdot)$ is the unit step function.

We design two spectral loss terms L_{spe1} and L_{spe2} to preserve the spectral information. Since there is a lack of GT to evaluate the spectral distortion, following Wald's protocol [46], we degrade $\mathbf{I}^{\theta}_{hrms}$ first and then compute the spectral losses between the degraded $\mathbf{I}^{\theta}_{hrms}$ and $\mathbf{I}^{\theta}_{Rms}$.

Due to the fluctuations in the spectrum, a small change in low-value bands can lead to significant spectral distortions, whereas the situation is reversed in high-value bands. To address this issue, we use the relative difference, which is defined as

$$L_{spe1} = \frac{1}{C}\sum_{c=1}^{C}\frac{|\mathrm{D}(\mathrm{G}(\mathbf{I}^{\theta}_{hrms}(c))) - \mathrm{D}(\mathbf{I}^{\theta}_{Rms}(c))|}{\mathrm{MEAN}(\mathrm{D}(\mathbf{I}^{\theta}_{Rms}(c)))}, \tag{27}$$

where $\mathrm{D}(\cdot)$ means downsampling.

L_{spe2} is designed to finetune the shape of spectral curve, ensuring the pixelwise and local spectral similarity, which contains two components. The first one is the Spectral Angle Mapper (SAM) [47,48], which calculates the angle between two spectral vectors, $\mathbf{x}$ and $\mathbf{y}$,

$$\mathrm{SAM}(\mathbf{x},\mathbf{y}) = \arccos\left(\frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\cdot\|\mathbf{y}\|}\right). \tag{28}$$

The second one is the structural Similarity Index Measure (SSIM) [49], which is an effective measure of similarity between two images, which is defined as

$$\mathrm{SSIM}(\mathbf{X},\mathbf{Y}) = \frac{(2\mu_{\mathbf{X}}\mu_{\mathbf{Y}}+C_1)(2\sigma_{\mathbf{XY}}+C_2)}{(\mu^2_{\mathbf{X}}+\mu^2_{\mathbf{Y}}+C_1)(\sigma^2_{\mathbf{X}}+\sigma^2_{\mathbf{Y}}+C_2)}, \tag{29}$$

where $\mu_{\mathbf{X}}$ and $\mu_{\mathbf{Y}}$ are the averages of $\mathbf{X}$ and $\mathbf{Y}$, respectively. C_1 and C_2 are constants to avoid instability. SSIM not only accelerates the convergence but also ensures the local smoothness of the fused results.

Therefore, L_{spe2} is defined as

$$L_{spe2} = \mathrm{SAM}(\mathrm{D}(\mathbf{I}^{\theta}_{Rms}), \mathrm{D}(\mathrm{G}(\mathbf{I}^{\theta}_{hrms}))) + 1 - \mathrm{SSIM}(\mathbf{I}^{\theta}_{Rms}, \mathrm{G}(\mathbf{I}^{\theta}_{hrms})). \tag{30}$$

3. Experiments

3.1. Data and Training Details

We conduct experiments on three satellite datasets with varying degrees of translation errors. GF2 and part of the WV2 images are sourced from [50]. Although our method is unsupervised, additional experiments are performed on reduced-resolution datasets, degraded according to Wald's protocol [46] for evaluation purposes. Training samples are generated by cropping 16 × 16 patches from LRMS and 64 × 64 patches from HRPan. The details of the datasets are listed in Table 2. It should be noted that the displacements listed in the table represent preliminary estimates obtained using the method described in [43].

Table 2. Details of the datasets. We build full resolution and reduced resolution datasets on GF2, WV2, and GF1 images.

Sensor	Image Details			Resolution		Image Crop	
	Bands	Bit	Displacement			Train	Test
GF2	4	10	[1.8 1.9]	Reduced	PAN:3.2m MS:12.8m Number	64 × 64 × 1 16 × 16 × 4 16,000	512 × 512 × 1 128 × 128 × 4 30
				Full	PAN:0.8m MS:3.2m Number	64 × 64 × 1 16 × 16 × 4 16,000	512 × 512 × 1 128 × 128 × 4 30
GF1	4	10	[0.6 0.8]	Reduced	PAN:8.0m MS:32.0m Number	64 × 64 × 1 16 × 16 × 4 15,616	512 × 512 × 1 128 × 128 × 4 27
				Full	PAN:2.0m MS:8.0m Number	64 × 64 × 1 16 × 16 × 4 15,616	512 × 512 × 1 128 × 128 × 4 27
WV2	4	11	[0.4 0.3]	Reduced	PAN:2.0m MS:8.0m Number	64 × 64 × 1 16 × 16 × 4 16,000	512 × 512 × 1 128 × 128 × 4 30
				Full	PAN:0.5m MS:2.0m Number	64 × 64 × 1 16 × 16 × 4 16,000	512 × 512 × 1 128 × 128 × 4 30

The proposed method is implemented in Pytorch and is trained on a single NVIDIA GeForce RTX 2080Ti GPU. We use the Adam optimizer, the initial learning rate is set to 0.0002, and the decay rate is 0.97. The iteration is 50000 in our experiment, and the batch size is 64. The trade-off parameters in the loss function are set as $w_1 = 100$, $w_2 = 4.5$, $w_3 = 4$, and $w_4 = 1$ for the WV2 dataset. The experiments for all methods are performed on the same datasets.

3.2. Experiment Design

The proposed method is compared with eleven pansharpening methods, including the following:

1. Traditional methods: adaptive Gram-Schmidt (GSA) [6], spatial fidelity with nonlocal regression (SFNLR) [51], and the nonconvex framelet sparse reconstruction method (NC-FSRM) [52]; these methods are conducted without training.
2. Supervised methods: diffusion model with disentangled modulations for sharpening multispectral and hyperspectral images (DDIF) [20], pansharpening GAN (PSGAN) [50], spatial–spectral dual back-project network (S2DBPN) [53] and dual-conditionally guided registration and fusion diffusion network (RFDifNet) [54]; these methods are trained on reduced resolution datasets and tested on both reduced and full resolution datasets. Among them, RFDifNet addresses the misalignment between HRPan and LRMS by explicitly estimating a deformation field to correct the spatial discrepancies between them.

3. Unsupervised methods: zero-shot semi-supervised method for pansharpening (ZSPan) [55], pansharpening via deep image prior (PSDip) [56], λ-CNN-based pansharpening (λ-PNN) [16], unsupervised cycle-consistent generative adversarial network (UCGAN) [18]; these methods are trained and tested on reduced and full resolution datasets, respectively. Among them, λ-PNN handles the translation misalignment between HRPan and LRMS by applying multiple shifted versions of the inputs to approximate the spatial offsets between the two images.

To evaluate the performance of different methods, seven widely used metrics are used for quantitative evaluation. Three of these are no-reference metrics used for full resolution assessment, including the spectral distortion index (D_λ), spatial distortion index (D_s), and quality with no-reference (QNR) [4,57]. Four of them are reference metrics used for reduced resolution assessment, including the erreur relative globale adimensionnelle de synthèse (ERGAS) [46], SSIM, SAM, and spatial correlation coefficient (SCC) [58]. Since translation errors exist between HRPan and LRMS, it is essential to account for their impact when computing evaluation metrics. Therefore, we compute the metrics using both the original (unaligned) and translation-corrected LRMS–HRPan pairs as reference images to provide a more comprehensive assessment.

3.3. Comparison on the GF2 Dataset

3.3.1. Visual Results

We test the performance of our method and the baseline methods on the GF2 dataset. Figures 7 and 8 display the fusion results at full and reduced resolutions. As shown in Figure 7, noticeable displacements in the GF2 dataset adversely affect all methods that lack translation match, particularly traditional and unsupervised ones, leading to duplicated edges in the fusion results. While λ-PNN mitigates some artifacts through iterative shift-based alignment, it fails to fully eliminate them. RFDifNet, which estimates a deformation field to register HRPan and LRMS, also struggles with the large misalignment and introduces spatial distortions. In contrast, our method effectively suppresses artifacts by employing TEM, which accurately compensates for the spatial translation between HRPan and MS inputs. Notably, the impact of translation errors becomes less significant on the reduced-resolution dataset (Figure 8), as displacements diminish during image downsampling. Nevertheless, among traditional and unsupervised approaches, our method remains highly competitive, preserving the fine structural details of buildings with higher fidelity.

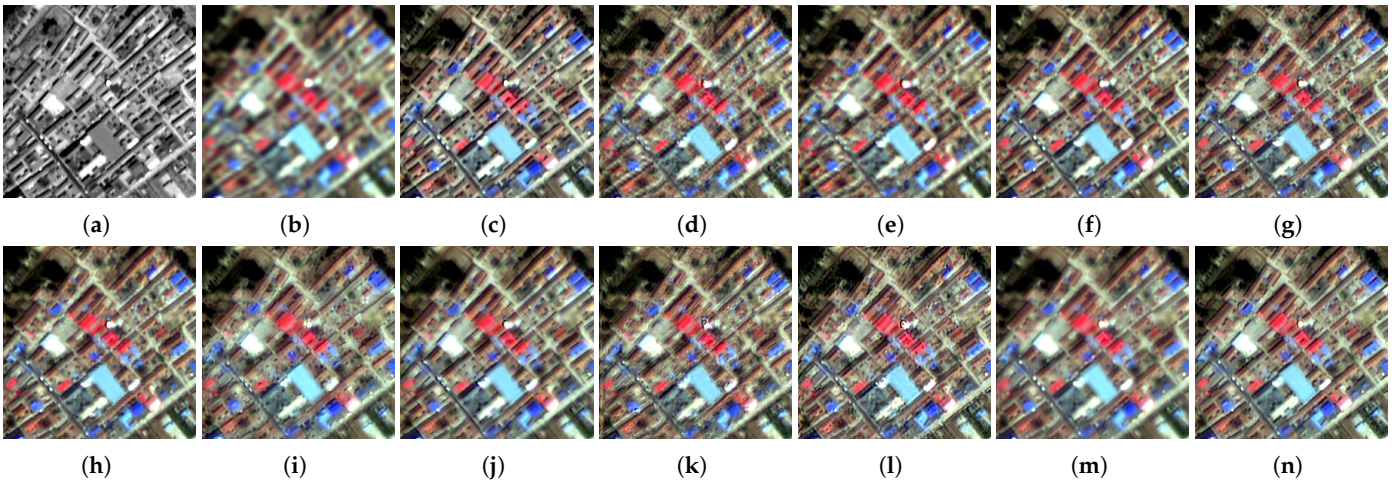


Figure 7. Fused results on the GF2 full resolution dataset. Visualized in RGB. (**a**) HRPan. (**b**) MS. Traditional: (**c**) GSA, (**d**) SFNLR, (**e**) NC-FSRM. Supervised: (**f**) DDIF, (**g**) S2DBPN, (**h**) PSGAN, (**i**) RFDifNet. Unsupervised: (**j**) ZSPan, (**k**) PSDip, (**l**) λ-PNN, (**m**) UCGAN. (**n**) SIUPan.

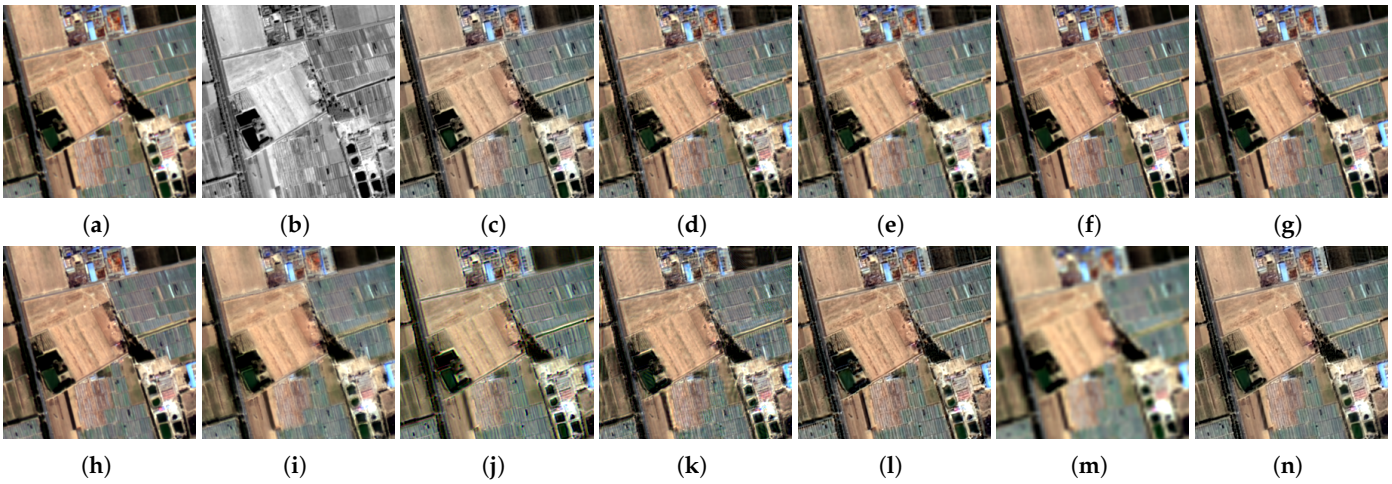


Figure 8. Fused results on the GF2 reduced resolution dataset. Visualized in RGB. (**a**) GT. (**b**) HRPan. Traditional: (**c**) GSA, (**d**) SFNLR, (**e**) NC-FSRM. Supervised: (**f**) DDIF, (**g**) S2DBPN, (**h**) PSGAN. (**i**) RFDifNet. Unsupervised: (**j**) ZSPan, (**k**) PSDip, (**l**) λ-PNN, (**m**) UCGAN, (**n**) SIUPan.

3.3.2. Quantitative Results

We provide quantitative comparisons on the GF2 dataset. Tables 3 and 4 show the quantitative results based on aligned and original image pairs, respectively. We emphasize that Table 3 more accurately reflects the fusion quality, as it minimizes the influence of translation misalignments. As shown in Table 3, most methods yield satisfactory results in terms of D_λ, with our method outperforming all other unsupervised methods. Benefiting from MSDI, our approach achieves a better balance between spectral preservation and spatial fidelity. Consequently, ours attains the best performance not only on D_s but also on QNR. For reference metrics, supervised methods naturally perform better due to the availability of GT, while our approach remains highly competitive among unsupervised and traditional methods.

Table 3. Quantitative results on the GF2 dataset based on translation-corrected LRMS and HRPan. The best result in each group is in bold font.

Type	Model	Full-Resolution Dataset			Reduced-Resolution Dataset			
		D_λ	D_s	QNR	ERGAS	SSIM	SCC	SAM
Traditional	GSA	0.0897	**0.0884**	**0.8348**	**1.5926**	**0.9206**	**0.9666**	**0.0288**
	SFNLR	**0.0577**	0.1961	0.7577	2.6620	0.8631	0.9064	0.0393
	NC-FSRM	0.0629	0.2051	0.7454	2.4518	0.8838	0.9171	0.0372
Supervised	DDIF	0.0631	0.1194	0.8258	**1.2568**	**0.9569**	**0.9770**	**0.0184**
	S2DBPN	0.0573	**0.1174**	**0.8327**	1.2840	0.9550	0.9765	0.0205
	PSGAN	0.0598	0.1306	0.8183	1.3197	0.9552	0.9756	0.0187
	RFDifNet	**0.0477**	0.2580	0.7066	1.9688	0.8760	0.9392	0.0264
Unsupervised	ZSPan	0.1309	0.0992	0.7826	2.8301	0.8516	0.9095	0.0500
	PSDip	0.0737	0.2525	0.6923	3.7595	0.7951	0.8641	0.0442
	λ-PNN	0.0687	0.1081	0.8296	2.7382	0.8546	0.9029	0.0369
	UCGAN	0.0600	0.1267	0.8213	2.5537	0.8323	0.9073	0.0344
	SIUPan	**0.0557**	**0.0521**	**0.8954**	**1.4554**	**0.9238**	**0.9673**	**0.0238**

When comparing Tables 3 and 4, we observe that some methods (e.g., DDIF, ZSPan) achieve better performance when evaluated on the original unregistered pairs because they are trained and inferred on unregistered data, causing their outputs to preserve the same geometric misalignment. In contrast, our method performs worse under this setting because it explicitly aligns HRPan and LRMS, producing fused results that are more

consistent with the registered references. This also indicates that evaluation metrics become unreliable when misalignment exists between HRPan and LRMS.

Table 4. Quantitative results on the GF2 dataset based on original LRMS and HRPan. The best result in each group is in bold font.

Type	Model	Full-Resolution Dataset			Reduced-Resolution Dataset			
		D_λ	D_s	QNR	ERGAS	SSIM	SCC	SAM
Traditional	GSA	0.0872	**0.0641**	**0.8589**	**1.9471**	**0.8954**	**0.9504**	**0.0324**
	SFNLR	**0.0532**	0.2210	0.7376	2.3560	0.8748	0.9274	0.0364
	NC-FSRM	0.0590	0.2300	0.7249	**2.1972**	**0.8885**	**0.9347**	0.0348
Supervised	DDIF	0.0585	**0.1293**	0.8204	**0.6769**	**0.9817**	**0.9928**	**0.0140**
	S2DBPN	0.0530	0.1298	**0.8245**	0.8675	0.9726	0.9892	0.0180
	PSGAN	0.0556	0.1442	0.8089	0.7950	0.9780	0.9914	0.0152
	RFDifNet	**0.0435**	0.2840	0.6848	2.0089	0.8889	0.9416	0.0293
Unsupervised	ZSPan	0.1269	0.1058	0.7801	2.7796	0.8525	0.9132	0.0496
	PSDip	0.0697	0.2779	0.6714	3.3020	0.8325	0.9026	0.0409
	λ-PNN	0.0658	0.1247	0.8163	2.4448	0.8636	0.9211	0.0342
	UCGAN	0.0566	0.1493	0.8027	2.7057	0.8077	0.9006	0.0369
	SIUPan	**0.0520**	**0.0642**	**0.8868**	**1.7824**	**0.9015**	**0.9541**	**0.0282**

3.4. Comparison on the GF1 Dataset

3.4.1. Visual Results

We further evaluate the proposed method on the GF1 dataset. Figures 9 and 10 present the fusion results at full and reduced resolutions, respectively. Although the GF1 dataset contains only sub-pixel translation errors (typically <1 pixel), these small misalignments still induce noticeable artifacts in the fusion results of methods that do not explicitly handle translation errors (Figure 9). Our method, along with RFDifNet and λ-PNN, effectively compensates for these errors and better preserves spectral integrity at land-cover boundaries. In contrast, supervised methods exhibit limited capability in preserving fine spatial details at full resolution, leading to noticeable boundary blurring (Figure 9). However, our method achieves an effective balance between spatial and spectral fidelity across both resolutions, outperforming other methods in preserving fine structures.

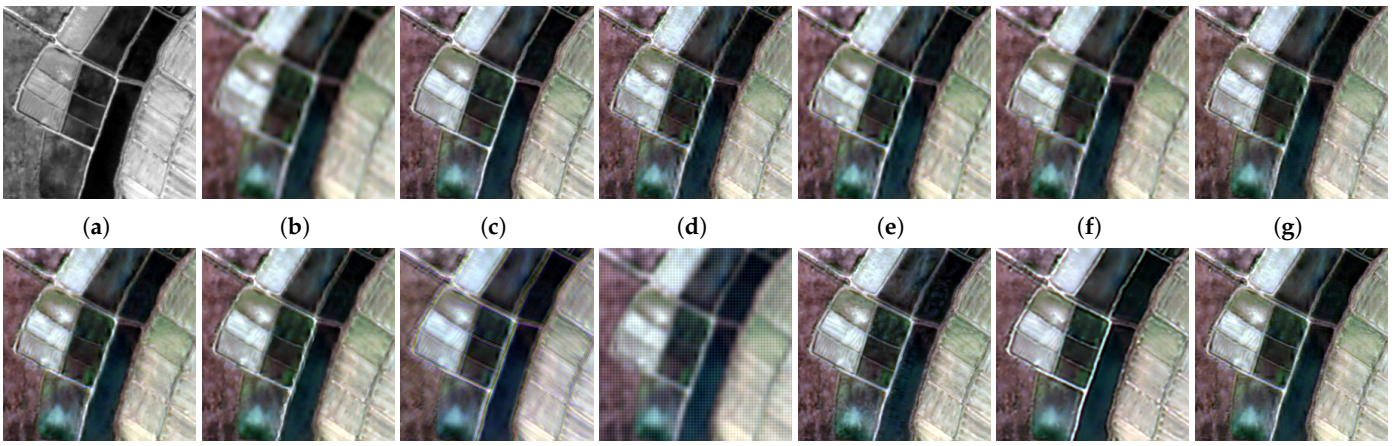


Figure 9. Fused results on the GF1 full resolution dataset. Visualized in RGB. (**a**) HRPan. (**b**) MS. Traditional: (**c**) GSA, (**d**) SFNLR, (**e**) NC-FSRM. Supervised: (**f**) DDIF, (**g**) S2DBPN, (**h**) PSGAN, (**i**) RFDifNet. Unsupervised: (**j**) ZSPan, (**k**) PSDip, (**l**) λ-PNN, (**m**) UCGAN. (**n**) SIUPan.

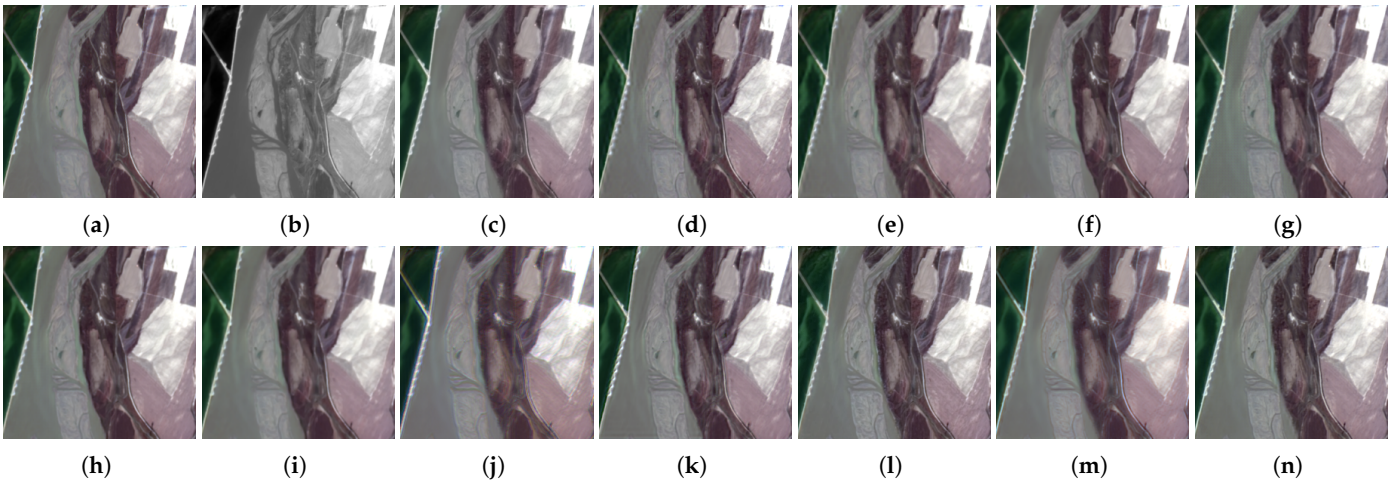


Figure 10. Fused results on the GF1 reduced resolution dataset. Visualized in RGB. (**a**) GT. (**b**) HRPan. Traditional: (**c**) GSA, (**d**) SFNLR, (**e**) NC-FSRM. Supervised: (**f**) DDIF, (**g**) S2DBPN, (**h**) PSGAN, (**i**) RFDifNet. Unsupervised: (**j**) ZSPan, (**k**) PSDip, (**l**) λ-PNN, (**m**) UCGAN, (**n**) SIUPan.

3.4.2. Quantitative Results

Tables 5 and 6 present the quantitative results based on the translation-corrected and original images, respectively. As shown in Table 5, supervised methods perform strongly on D_λ, with PSGAN achieving the best overall results. This is because the GF1 dataset is dominated by large homogeneous land-cover regions (e.g., farmland), where the spectral difference between full- and reduced-resolution images is relatively small [59]. Among unsupervised methods, UCGAN achieves the best D_λ, while our method ranks second. However, UCGAN introduces obvious spectral distortions along linear structures such as roads (Figure 9m). Our method achieves the best D_s and QNR, demonstrating its strong spatial reconstruction capability and practical potential.

Table 5. Quantitative results on the GF1 dataset based on translation-corrected LRMS and HRPan. The best result in each group is in bold font.

Type	Model	Full-Resolution Dataset			Reduced-Resolution Dataset			
		D_λ	D_s	QNR	ERGAS	SSIM	SCC	SAM
Traditional	GSA	0.0675	0.0672	0.8717	**1.3197**	**0.9208**	**0.9627**	**0.0291**
	SFNLR	**0.0425**	**0.0668**	**0.8936**	1.7499	0.9028	0.9401	0.0344
	NC-FSRM	0.0445	0.0698	0.8891	1.6488	**0.9125**	0.9435	0.0331
Supervised	DDIF	0.0442	0.0880	0.8720	**0.7454**	**0.9729**	**0.9871**	**0.0160**
	S2DBPN	0.0421	**0.0629**	0.8979	0.8622	0.9642	0.9831	0.0190
	PSGAN	**0.0393**	0.0841	0.8800	0.8508	0.9667	0.9843	0.0184
	RFDifNet	**0.0416**	0.0631	**0.8980**	1.0630	0.9547	0.9760	0.0216
Unsupervised	ZSPan	0.0551	0.0723	0.8777	1.6870	0.8811	0.9295	0.0394
	PSDip	0.0498	0.0837	0.8707	2.0766	0.8707	0.9148	0.0397
	λ-PNN	0.0616	0.0669	0.8766	1.4342	**0.9162**	0.9542	0.0283
	UCGAN	**0.0402**	0.0687	0.8945	1.8216	0.8817	0.9357	0.0395
	SIUPan	0.0419	**0.0520**	**0.9086**	**1.3738**	0.9125	**0.9609**	**0.0283**

When comparing Tables 5 and 6, we observe that the no-reference metrics change only slightly, whereas the reference-based metrics exhibit more noticeable variations. For the no-reference metrics, the differences between the results computed using translation-corrected and original pairs are much smaller on GF1 than on GF2. This further supports our explanation that such improvements on unregistered data mainly stem from training

and generating results on unregistered inputs. By comparing the behaviors of both the no-reference and reference metrics across the two tables, we can also find that the no-reference metrics are less sensitive to sub-pixel translation errors.

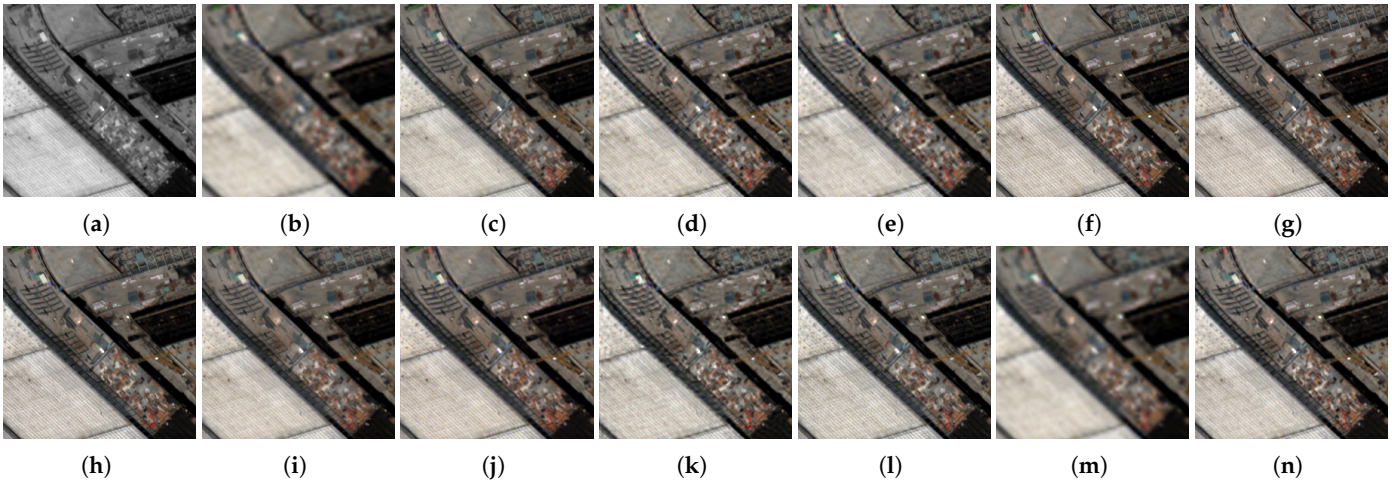
Table 6. Quantitative results on the GF1 dataset based on original LRMS and HRPan. The best result in each group is in bold font.

Type	Model	Full-Resolution Dataset			Reduced-Resolution Dataset			
		D_λ	D_s	QNR	ERGAS	SSIM	SCC	SAM
Traditional	GSA	0.0667	0.0692	0.8706	**1.4269**	**0.9133**	**0.9558**	**0.0312**
	SFNLR	**0.0423**	**0.0661**	**0.8945**	1.9427	0.8891	0.9267	0.0376
	NC-FSRM	0.0442	0.0691	0.8900	1.8574	0.8967	0.9295	0.0364
Supervised	DDIF	0.0440	0.0873	0.8728	**0.6684**	**0.9743**	**0.9882**	**0.0154**
	S2DBPN	0.0418	0.0624	0.8985	0.8545	0.9615	0.9824	0.0197
	PSGAN	**0.0392**	0.0832	0.8811	0.7770	0.9680	0.9852	0.0180
	RFDifNet	0.0413	**0.0619**	**0.8994**	1.0244	0.9540	0.9769	0.0219
Unsupervised	ZSPan	0.0761	0.0549	0.8736	1.7844	0.8737	0.9221	0.0411
	PSDip	0.0495	0.0824	0.8721	2.3010	0.8511	0.8970	0.0431
	λ-PNN	0.0613	0.0681	0.8757	1.6296	0.9048	0.9435	0.0318
	UCGAN	**0.0401**	0.0681	0.8952	1.9599	0.8668	0.9274	0.0418
	SIUPan	0.0415	**0.0526**	**0.9083**	**1.4971**	**0.9063**	**0.9542**	**0.0312**

3.5. Comparison on the WV2 Dataset

3.5.1. Visual Results

Figures 11 and 12 illustrate the fusion results on the WV2 dataset at full and reduced resolutions, respectively. As the displacements between HRPan and MS are minimal (<0.5 pixels in average), no noticeable artifacts appear in the fusion results across all methods. Among unsupervised methods, UCGAN preserves spectral fidelity well but exhibits evident spatial distortion, due to the difficulty in balancing spectral and spatial fidelity. On the reduced resolution dataset, supervised methods still show a significant advantage. Compared with unsupervised and traditional methods, our method achieves better results at both full and reduced resolution.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**) (**g**)

(**h**) (**i**) (**j**) (**k**) (**l**) (**m**) (**n**)

Figure 11. Fused results on the WV2 full resolution dataset. Visualized in RGB. (**a**) HRPan. (**b**) MS. Traditional: (**c**) GSA, (**d**) SFNLR, (**e**) NC-FSRM. Supervised: (**f**) DDIF, (**g**) S2DBPN, (**h**) PSGAN, (**i**) RFDifNet. Unsupervised: (**j**) ZSPan, (**k**) PSDip, (**l**) λ-PNN, (**m**) UCGAN, (**n**) SIUPan.

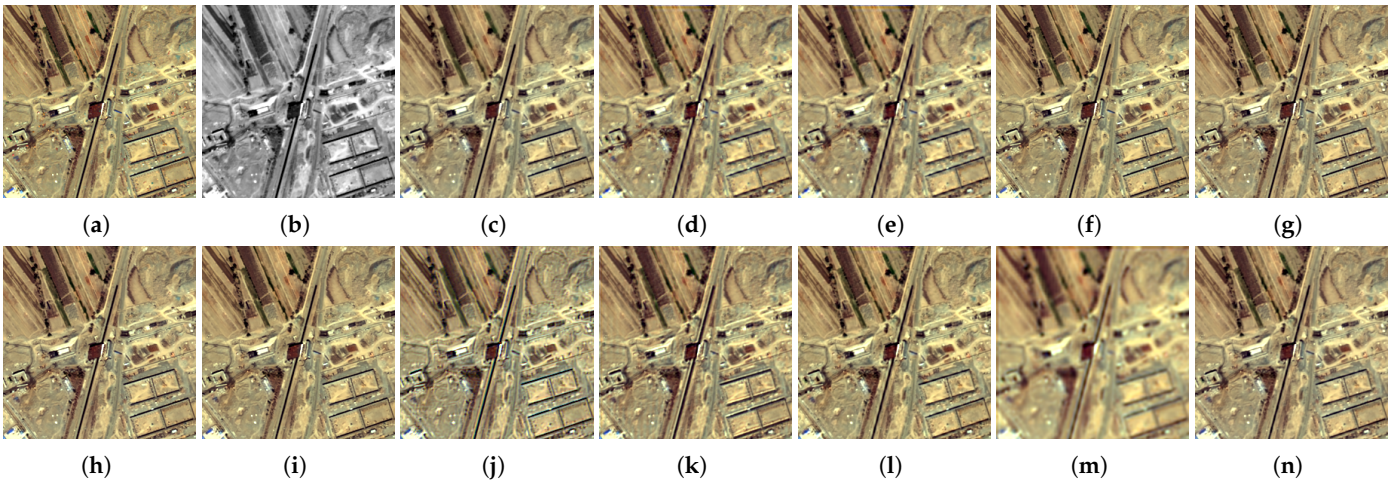


Figure 12. Fused results on the WV2 reduced resolution dataset. Visualized in RGB. (**a**) GT. (**b**) HRPan. Traditional: (**c**) GSA, (**d**) SFNLR, (**e**) NC-FSRM. Supervised: (**f**) DDIF, (**g**) S2DBPN, (**h**) PSGAN, (**i**) RFDifNet. Unsupervised: (**j**) ZSPan, (**k**) PSDip, (**l**) λ-PNN, (**m**) UCGAN, (**n**) SIUPan.

3.5.2. Quantitative Results

Tables 7 and 8 present the quantitative results based on the translation-corrected and original images, respectively. As shown in Table 7, on the full-resolution dataset, our method achieves the best performance in both D_s and QNR and ranks second in D_λ. Although RFDifNet attains the best D_λ, its D_s is considerably worse, indicating spatial distortions. On the reduced-resolution dataset, supervised methods maintain superiority. Among unsupervised methods, our method achieves the best performance across all metrics, effectively preserving both spatial and spectral information, which also demonstrates the powerful generation capability of diffusion models.

Comparing Tables 7 and 8, we observe that differences still exist even under small displacements (<0.5 pixel), indicating that even slight misalignment can influence the evaluation of the fusion quality.

Table 7. Quantitative results on the WV2 dataset based on translation-corrected LRMS and HRPan. The best result in each group is in bold font.

Type	Model	Full-Resolution Dataset			Reduced-Resolution Dataset			
		D_λ	D_s	QNR	ERGAS	SSIM	SCC	SAM
Traditional	GSA	**0.0492**	0.0603	0.8946	**1.5177**	**0.9532**	**0.9723**	**0.0375**
	SFNLR	0.0525	0.0553	0.8955	2.1754	0.9306	0.9469	0.0401
	NC-FSRM	0.0514	**0.0531**	**0.8987**	2.5857	0.9075	0.9253	0.0460
Supervised	DDIF	0.0472	0.0601	0.8967	**1.0864**	**0.9772**	**0.9863**	**0.0212**
	S2DBPN	0.0485	**0.0540**	**0.9029**	1.1842	0.9729	0.9831	0.0254
	PSGAN	0.0497	0.0591	0.8965	1.0987	0.9762	0.9858	0.0224
	RFDifNet	**0.0440**	0.0582	0.9005	1.1228	0.9749	0.9840	0.0227
Unsupervised	ZSPan	0.0886	0.0808	0.8400	2.8627	0.9274	0.9295	0.0448
	PSDip	0.0515	0.0782	0.8748	2.7615	0.9004	0.9149	0.0482
	λ-PNN	0.0618	0.0601	0.8833	2.1212	0.9413	0.9531	0.0391
	UCGAN	0.0509	0.0945	0.8610	2.8051	0.8678	0.9109	0.0471
	SIUPan	**0.0472**	**0.0511**	**0.9050**	**1.6833**	**0.9533**	**0.9684**	**0.0358**

Table 8. Quantitative results on the WV2 dataset based on original LRMS and HRPan. The best result in each group is in bold font.

Type	Model	Full-Resolution Dataset			Reduced-Resolution Dataset			
		D_λ	D_s	QNR	ERGAS	SSIM	SCC	SAM
Traditional	GSA	**0.0487**	0.0688	0.8870	**1.8513**	**0.9362**	**0.9606**	**0.0423**
	SFNLR	0.0510	0.0522	0.9000	2.5727	0.9123	0.9277	0.0454
	NC-FSRM	0.0498	**0.0515**	**0.9017**	2.5857	0.9075	0.9253	0.0460
Supervised	DDIF	0.0471	0.0590	0.8983	**0.9485**	**0.9802**	**0.9883**	**0.0221**
	S2DBPN	0.0486	**0.0562**	0.9015	1.1909	0.9709	0.9827	0.0273
	PSGAN	0.0497	0.0588	0.8976	1.0521	0.9766	0.9859	0.0239
	RFDifNet	**0.0439**	0.0562	**0.9028**	1.0787	0.9750	0.9842	0.0240
Unsupervised	ZSPan	0.0872	0.0786	0.8430	3.1150	0.9094	0.9137	0.0487
	PSDip	0.0504	0.0737	0.8801	3.2033	0.8725	0.8869	0.0539
	λ-PNN	0.0624	0.0618	0.8820	2.5124	0.9240	0.9346	0.0447
	UCGAN	0.0489	0.0867	0.8699	3.1174	0.8437	0.8931	0.0516
	SIUPan	**0.0473**	**0.0526**	**0.9040**	**2.0620**	**0.9366**	**0.9535**	**0.0412**

3.6. Ablation

3.6.1. TEM

In the proposed method, TEM is an essential module to match HRPan and MS; thus, we conduct ablation experiments on the GF2 and WV2 datasets. To demonstrate the capability of TEM, we perform GSA using $\mathbf{I}_{\text{pan}}$ and $\mathbf{I}^{\theta}_{\text{Rms}}$, as well as $\mathbf{I}_{\text{pan}}$ and $\mathbf{I}_{\text{ms}}$, respectively, and the quantitative results are tabulated in Table 9. For the GF2 dataset, the quality of the GSA fusion results improves significantly after translation correction. In contrast, for the WV2 dataset, only slight improvements are observed. These results indicate that TEM successfully maps HRPan and MS into a shared feature space, effectively reducing translation errors.

Table 9. Quantitative assessment of translation correction on GF2 and WV2 datasets. GSA is used to fuse HRPan and MS.

Dataset	MS	GSA Fusion Results		
		D_λ	D_s	QNR
GF2	$\mathbf{I}_{\text{ms}}$	0.0897	0.0884	0.8348
	$\mathbf{I}^{\theta}_{\text{Rms}}$	0.0800	0.0748	0.8553
WV2	$\mathbf{I}_{\text{ms}}$	0.0492	0.0603	0.8946
	$\mathbf{I}^{\theta}_{\text{Rms}}$	0.0485	0.0595	0.8960

To verify the validity of TEM in SIUPan, we remove it from the proposed network and then conduct comparative experiments. As TEM is excluded, $\mathbf{I}^{\theta}_{\text{Rms}}$ is unavailable, and thus L_{R} cannot be computed, requiring its omission. Figure 13 shows the fused images before and after removing TEM, while Table 10 provides the corresponding quantitative evaluation.



(**a**) (**b**) (**c**) (**d**)

Figure 13. Results of TEM ablation study. Visualized in RGB. (**a**) With TEM on the GF2 dataset. (**b**) Without TEM on the GF2 dataset. (**c**) With TEM on the WV2 dataset. (**d**) Without TEM on the WV2 dataset.

Table 10. Quantitative assessment of TEM ablation study on GF2 and WV2 datasets.

Dataset	TEM	Full-Resolution Datasets			Reduced-Resolution Dataset			
		D_λ	D_s	QNR	ERGAS	SSIM	SCC	SAM
GF2	w/o	0.0607	0.1041	0.8409	2.3472	0.8749	0.9298	0.0337
	w/	0.0557	0.0521	0.8954	1.4554	0.9238	0.9673	0.0238
WV2	w/o	0.0481	0.0524	0.9027	1.8962	0.9482	0.9625	0.0386
	w/	0.0472	0.0511	0.9050	1.6833	0.9533	0.9684	0.0358

As shown in Figure 13a,b, for the GF2 dataset, removing the TEM leads to pronounced spatial distortions and noticeable artifacts, severely degrading the fusion quality. For the WV2 dataset, where translation errors are minimal, no apparent visual differences are observed in Figure 13c,d; however, the quantitative metrics still show a consistent decline (Table 10). These findings demonstrate that even sub-pixel translation discrepancies can negatively affect the pansharpening performance, highlighting the necessity of the TEM for achieving robust and high-fidelity fusion in SIUPan.

3.6.2. $\hat{X}_0$

In SIUPan, we replace x_0 with $\hat{x}_0$. We assume that the difference between x_0 and $\hat{x}_0$ follows a zero-mean Gaussian distribution. This assumption is relatively mild, implying that our method should be insensitive to the choice of $\hat{x}_0$. To validate this, we conduct experiments using four different estimates of $\hat{x}_0$: (1) $\mathbf{I}_{\text{ones}}$, (2) $\mathbf{I}^{\text{C}}_{\text{pan}}$, (3) $\mathbf{I}_{\text{GSA}}$, and (4) GT. Here, $\mathbf{I}_{\text{ones}}$ denotes an all-ones matrix, while $\mathbf{I}_{\text{GSA}}$ represents the fusion result obtained by the GSA method. Additionally, since GT is only available on reduced resolution datasets, we conducted experiments for the fourth case only at reduced resolution. The results are shown in Table 11.

Table 11. Quantitative assessment of $\hat{x}_0$ ablation study on the WV2 dataset.

$\hat{x}_0$	Full-Resolution Dataset			Reduced-Resolution Dataset			
	D_λ	D_s	QNR	ERGAS	SSIM	SCC	SAM
$\mathbf{I}_{\text{ones}}$	0.0491	0.0545	0.9006	1.7288	0.9516	0.9669	0.0364
$\mathbf{I}^{\text{C}}_{\text{pan}}$	0.0472	0.0511	0.9050	1.6833	0.9533	0.9684	0.0358
$\mathbf{I}_{\text{GSA}}$	0.0475	0.0519	0.9041	1.6856	0.9531	0.9687	0.0356
GT	-	-	-	1.6889	0.9530	0.9682	0.0357

As tabulated in Table 11, $\mathbf{I}_{\text{ones}}$ results in a relatively poor performance, because $\mathbf{I}_{\text{ones}} - \mathbf{I}_{\text{hrms}}$ deviates from the zero-mean assumption in (4). In contrast, the other three estimates produce comparable results, demonstrating that the proposed method is largely insensitive to the choice of $\hat{x}_0$. Therefore, considering both the accuracy and efficiency, we use $\mathbf{I}^{\text{C}}_{\text{pan}}$ as $\hat{x}_0$ in our method.

3.6.3. MSDI

MSDI is designed to inject spatial and spectral information into the encoder when generating HRMS in our work, enabling it to achieve strong performance even on small-scale datasets. To evaluate the effectiveness of the MSDI, we conduct ablation experiments on the WV2 dataset. The MSDI employs cross-attention at multiscale to fully exploit the information contained in HRPan and MS, balancing the spatial and spectral details in the fusion results. As shown in Table 12, removing the MSDI leads to a noticeable performance drop on the WV2 dataset. In particular, D_s degrades substantially, indicating that without the MSDI, the fused images suffer from spatial distortions. These results verify that the

MSDI plays a crucial role in enhancing the spatial consistency and overall fusion quality of SIUPan.

Table 12. Quantitative assessment of MSDI ablation study on the WV2 dataset.

MSDI	Full-Resolution Dataset			Reduced-Resolution Dataset			
	D_λ	D_s	QNR	ERGAS	SSIM	SCC	SAM
w/o	0.0470	0.1154	0.8440	2.7202	0.8667	0.9152	0.0396
w/	0.0472	0.0511	0.9050	1.6833	0.9533	0.9684	0.0358

3.6.4. Loss Terms

There are four loss terms in the proposed method, three of which are fusion losses and one of which is registration loss. To validate the necessity of each component, we conduct ablation experiments, and the quantitative results are summarized in Table 13. As shown in Table 13, removing L_{spa} leads to severely degraded fusion results, as it is the only loss directly constraining the spatial fidelity. The absence of L_{spe1} causes the model to fail in achieving the best results at full resolution. The loss term L_{spe2} further refines the spectral curve and helps balance the fusion quality across multiple scales; its removal results in moderate degradation in all metrics. With L_{R}, the TEM provides more accurate translation error estimation between HRPan and MS, leading to improved performance on D_s. According to the experiment results, each term of the loss function ensures that the model can achieve good results at both full and reduced resolutions.

Table 13. Quantitative assessment of different loss functions on the WV2 dataset.

Loss Term				Full-Resolution Dataset		Reduced-Resolution Dataset				
L_{R}	L_{spa}	L_{spe1}	L_{spe2}	D_λ	D_s	QNR	ERGAS	SSIM	SCC	SAM
×	✓	✓	✓	0.0459	0.0536	0.9044	1.7838	0.9455	0.9654	0.0355
✓	×	✓	✓	0.1367	0.6200	0.3307	51.8171	0.0429	0.1461	0.0845
✓	✓	×	✓	0.0542	0.0629	0.8871	1.5635	0.9548	0.9701	0.0362
✓	✓	✓	×	0.0508	0.0546	0.8984	1.7283	0.9520	0.9670	0.0360
✓	✓	✓	✓	0.0472	0.0511	0.9050	1.6833	0.9533	0.9684	0.0358

4. Conclusions

In this article, we propose a shift-invariant unsupervised pansharpening network. In the proposed method, a translation error estimation module, i.e., TEM, is employed to estimate the displacements between HRPan and MS. Moreover, a multiscale detail injection module is designed to balance the spatial and spectral fidelity. Extensive experiments on the GF2, GF1, and WV2 datasets demonstrate that the proposed method outperforms baseline pansharpening approaches. The ablation studies further confirm that each designed module and loss function plays an essential role in the overall performance of the model.

Compared to supervised methods, unsupervised methods can exploit the original information of HRPan and MS. Nevertheless, because GT is absent, its spectral and spatial losses are computed independently, highlighting the importance of accurately handling translation errors between HRPan and MS. Furthermore, our work primarily focuses on translation errors between HRPan and MS, while potential misalignments among individual MS bands also deserve further study.

Author Contributions: Conceptualization, J.X. and L.J.; methodology, J.X., L.J. and J.Y.; validation, J.X. and J.Y.; formal analysis, Q.F., J.X. and J.L.; investigation, J.X. and J.L.; resources, Y.Z., Q.F. and K.L.; data curation, J.L. and K.L.; writing—original draft preparation, J.X.; writing—review and editing,

L.J. and J.Y.; visualization, J.X.; funding acquisition, K.L. and Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article; further inquiries can be directed to the author.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

The log likelihood in (7) can be expressed as follows:

$$\begin{aligned}
\ln p(x_0) =& \ln \int p(x_0, \hat{x}_{0:T}) d\hat{x}_{0:T} \\
=& \ln \int \frac{p(x_0, \hat{x}_{0:T}) d\hat{x}_{0:T} q(\hat{x}_{0:T}|x_0)}{q(\hat{x}_{0:T}|x_0)} d\hat{x}_{0:T} \\
\geq& \mathbb{E}_q \left[\ln \frac{p^\theta(x_0, \hat{x}_{0:T})}{q(\hat{x}_{0:T}|x_0)} \right] \\
=& \mathbb{E}_q \left[\ln \frac{p(\hat{x}_T) p(x_0|\hat{x}_0) \prod_{t=1}^T p^\theta(\hat{x}_{t-1}|\hat{x}_t)}{q(\hat{x}_0|x_0) \prod_{t=1}^T q(\hat{x}_t|\hat{x}_{t-1})} \right] \\
=& \mathbb{E}_q \left[\ln \frac{p(\hat{x}_T) p^\theta(\hat{x}_0|\hat{x}_1) p(x_0|\hat{x}_0)}{q(\hat{x}_1|\hat{x}_0) q(\hat{x}_0|x_0)} \right] + \mathbb{E}_q \left[\ln \frac{\prod_{t=2}^T p^\theta(\hat{x}_{t-1}|\hat{x}_t)}{\prod_{t=2}^T q(\hat{x}_t|\hat{x}_{t-1}, x_0)} \right] \\
=& \mathbb{E}_q \left[\ln \frac{p(\hat{x}_T) p^\theta(\hat{x}_0|\hat{x}_1) p(x_0|\hat{x}_0)}{q(\hat{x}_1|\hat{x}_0) q(\hat{x}_0|x_0)} \right] + \mathbb{E}_q \left[\ln \frac{\prod_{t=2}^T p^\theta(\hat{x}_{t-1}|\hat{x}_t)}{\prod_{t=2}^T \frac{q(\hat{x}_{t-1}|\hat{x}_t, x_0) q(\hat{x}_t|x_0)}{q(\hat{x}_{t-1}|x_0)}} \right] \\
=& \mathbb{E}_q \left[\ln \frac{p(\hat{x}_T) p^\theta(\hat{x}_0|\hat{x}_1) p(x_0|\hat{x}_0) q(\hat{x}_1|x_0)}{q(\hat{x}_1|\hat{x}_0) q(\hat{x}_0|x_0) q(\hat{x}_T|x_0)} \right] + \mathbb{E}_q \left[\ln \prod_{t=1}^T \frac{p^\theta(\hat{x}_0|\hat{x}_1)}{q(\hat{x}_{t-1}|\hat{x}_t, x_0)} \right] \\
=& \mathbb{E}_q \left[\ln \frac{p^\theta(\hat{x}_0|\hat{x}_1) p(x_0|\hat{x}_0)}{q(\hat{x}_0|x_0)} \right] + \mathbb{E}_q \left[\ln \frac{p(\hat{x}_T)}{q(\hat{x}_T|x_0)} \right] + \mathbb{E}_q \left[\sum_{t=2}^T \ln \frac{p^\theta(\hat{x}_{t-1}|\hat{x}_t)}{q(\hat{x}_{t-1}|\hat{x}_t, x_0)} \right] + \mathbb{E}_q \left[\ln \frac{q(\hat{x}_1|x_0)}{q(\hat{x}_1|\hat{x}_0)} \right] \\
:=& L_{\text{ELBO}}.
\end{aligned} \tag{A1}$$

We notice that, once $\hat{x}_0$ is fixed, δ is independent of $\hat{x}_t$; thus, L_{ELBO} can be rewritten as

$$L_{\text{ELBO}} = \underbrace{\mathbb{E}_q \left[\ln p^\theta(\hat{x}_0|\hat{x}_1) \mathrm{C}_1(\delta) \right]}_{\text{reconstruction term}} - \underbrace{D_{KL}(q(\hat{x}_T|x_0)||p(\hat{x}_T))}_{\text{prior matching term}} - \sum_{t=2}^T \underbrace{\mathbb{E}_q \left[D_{KL}(q(\hat{x}_{t-1}|\hat{x}_t, x_0)||p^\theta(\hat{x}_{t-1}|\hat{x}_t)) \right]}_{\text{denoising matching term}} + \underbrace{\mathrm{C}_2(\delta)}_{\text{constant term}}, \tag{A2}$$

where $\mathrm{C}_1(\delta)$ and $\mathrm{C}_2(\delta)$ are terms that only depend on δ. Therefore, similar to DDPM, to maxmize L_{ELBO}, we only need to maximize the reconstruction term and denoising matching term.

For the reconstruction term, we need to maximize $\ln p^\theta(\hat{x}_0|\hat{x}_1)$. Expanding $\ln p^\theta(\hat{x}_0|\hat{x}_1)$, we can get

$$\begin{aligned}
\ln p^\theta(\hat{x}_0|\hat{x}_1) =& \ln \frac{1}{(2\pi)^n |\hat{\Sigma}|^{\frac{1}{2}} \delta} e^{-\frac{1}{2}(\hat{x}_0 - \mu_{t_1})^{\mathrm{T}} |\hat{\Sigma}|^{-1} (\hat{x}_0 - \mu_{t_1})} \\
=& \ln \frac{1}{(2\pi)^n |\hat{\Sigma}|^{\frac{1}{2}} \delta} - \frac{1}{2}(\hat{x}_0 - \mu_{t_1})^{\mathrm{T}} |\hat{\Sigma}|^{-1} (\hat{x}_0 - \mu_{t_1}) \\
\propto& -\frac{1}{2}(\hat{x}_0 - \mu_{t_1})^{\mathrm{T}} |\hat{\Sigma}|^{-1} (\hat{x}_0 - \mu_{t_1}) \\
\propto& -\frac{1}{2} \|\hat{x}_0 - \mu_{t_1}\|_2^2
\end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{2}\|x_0 - (z + \mu_{t_1})\|_2^2 \\
&= -\frac{1}{2}\left\|x_0 - x_0^\theta\right\|_2^2,
\end{aligned} \tag{A3}
$$

where x_0^θ is the prediction of network, and μ_{t_1} and $\hat{\Sigma}$ are the mean and the variance, respectively. In (A3), z is absorbed into the prediction of network. Thus, to maximize $\ln p^\theta(\hat{x}_0|\hat{x}_1)$, we just need to minimize $\left\|x_0 - x_0^\theta\right\|_2^2$.

For the denoising matching term, we use $p^\theta(\hat{x}_{t-1}|\hat{x}_t)$ to approximate the real distribution $p(\hat{x}_{t-1}|\hat{x}_t, x_0)$. $p(\hat{x}_{t-1}|\hat{x}_t, x_0)$ is a conditional Gaussian distribution, which can be expressed as

$$
\begin{aligned}
&p(\hat{x}_{t-1}|\hat{x}_t, x_0) \\
=&\frac{q(\hat{x}_t|\hat{x}_{t-1}, x_0)q(\hat{x}_{t-1}|x_0)}{q(\hat{x}_t|x_0)} \\
=&\frac{\mathcal{N}(\hat{x}_t; \sqrt{\alpha}_t\hat{x}_{t-1}, (1-\alpha_t)\mathbf{I})}{\mathcal{N}(\hat{x}_t; \sqrt{\overline{\alpha}}_t x_0, (1-\overline{\alpha}_t + \delta^2\overline{\alpha}_t)\mathbf{I})} \cdot \mathcal{N}(\hat{x}_{t-1}; \sqrt{\overline{\alpha}}_{t-1}x_0, (1-\overline{\alpha}_{t-1} + \delta^2\overline{\alpha}_{t-1})\mathbf{I}) \\
\propto&\exp\left\{ -\left[\frac{(\hat{x}_t - \sqrt{\alpha_t}\hat{x}_{t-1})^2}{2(1-\alpha_t)} + \frac{(\hat{x}_{t-1} - \sqrt{\overline{\alpha}_t}x_0)^2}{2(1-\sqrt{\alpha}_{t-1} + \delta^2\overline{\alpha}_{t-1})} - \underbrace{\frac{(\hat{x}_t - \sqrt{\overline{\alpha}_t}x_0)^2}{2(1-\overline{\alpha}_t + \delta^2\overline{\alpha}_{t-1})}}_{\text{constant term}}\right]\right\} \\
\propto&\exp\left\{ -\frac{1}{2}\left[\frac{\alpha_t}{1-\alpha_t} + \frac{1}{1-\overline{\alpha}_{t-1} + \delta^2\overline{\alpha}_{t-1}}\left(\hat{x}_{t-1}^2 - 2\frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1} + \delta^2\overline{\alpha}_{t-1})\hat{x}_t + \sqrt{\overline{\alpha}_{t-1}}(1-\alpha_t)x_0}{1-\overline{\alpha}_t + \delta^2\overline{\alpha}_t} \cdot \hat{x}_{t-1}\right)\right]\right\} \\
\propto&\mathcal{N}(\hat{x}_{t-1}; \mu_{q_t}, \Sigma_{q_t}),
\end{aligned} \tag{A4}
$$

where μ_{q_t} and Σ_{q_t} are the mean and the variance, respectively. Here are their expressions:

$$
\mu_{q_t} = \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1} + \delta^2\overline{\alpha}_{t-1})\hat{x}_t + \sqrt{\overline{\alpha}_{t-1}}(1-\alpha_t)x_0}{1-\overline{\alpha}_t + \delta^2\overline{\alpha}_t}, \tag{A5}
$$

$$
\Sigma_{q_t} = \frac{1-\overline{\alpha}_t + \delta^2\overline{\alpha}_t}{(1-\alpha_t)(1-\overline{\alpha}_{t-1} + \delta^2\overline{\alpha}_{t-1})}\mathbf{I} = \sigma_{q_t}^2\mathbf{I}. \tag{A6}
$$

We assume that $p^\theta(\hat{x}_{t-1}|\hat{x}_t) = \mathcal{N}(\hat{x}_{t-1}; \mu_p, \Sigma_p)$. Similar to DDPM, we treat the variance as a constant that does not need to be learned by the network, setting $\Sigma_p = \Sigma_{q_t}$. Define the form of μ_p as follows:

$$
\mu_p = \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1} + \delta^2\overline{\alpha}_{t-1})\hat{x}_t + \sqrt{\overline{\alpha}_{t-1}}(1-\alpha_t)x_0^\theta}{1-\overline{\alpha}_t + \delta^2\overline{\alpha}_t}, \tag{A7}
$$

where x_0^θ is the prediction of network. Thus, all KL divergences can be expressed as follows:

$$
\begin{aligned}
D_{\mathrm{KL}}\left(q(\hat{x}_{t-1}|\hat{x}_t, x_0)||p^\theta(\hat{x}_{t-1}|\hat{x}_t)\right) =& D_{\mathrm{KL}}\left(\mathcal{N}(\hat{x}_{t-1}; \mu_{q_t}, \Sigma_{q_t})||\mathcal{N}(\hat{x}_{t-1}; \mu_p, \Sigma_{q_t})\right) \\
=& \frac{1}{2}\left[\ln\frac{|\Sigma_{q_t}|}{|\Sigma_{q_t}|} - d + \mathrm{tr}(\Sigma_{q_t}^{-1}\Sigma_{q_t}) + (\mu_p - \mu_{q_t})^{\mathrm{T}}\Sigma_{q_t}^{-1}(\mu_p - \mu_{q_t})\right] \\
=& \frac{1}{2\sigma_{q_t}^2}\left[\|\mu_p - \mu_{q_t}\|_2^2\right] \\
=& \frac{\sqrt{\overline{\alpha}_{t-1}}(1-\alpha_t)}{2\sigma_{q_t}^2(1-\overline{\alpha}_t + \delta^2\overline{\alpha}_t)}\left[\|x_0^\theta - x_0\|_2^2\right] \\
\propto& \|x_0^\theta - x_0\|_2^2.
\end{aligned} \tag{A8}
$$

According to (A3) and (A8), we can obtain

$$L_{\mathrm{ELBO}} \propto -\|x_0^\theta - x_0\|_2^2. \tag{A9}$$

Thus, to maximize L_{ELBO}, we only need to minimize $\|x_0^\theta - x_0\|_2^2$.

References

1. Gong, P.; Wang, J.; Huang, H. Stable Classification with Limited Samples in Global Land Cover Mapping: Theory and Experiments. *Sci. Bull.* **2024**, *69*, 1862–1865. [CrossRef] [PubMed]
2. Wang, Y.; Sun, Y.; Cao, X.; Wang, Y.; Zhang, W.; Cheng, X. A Review of Regional and Global Scale Land Use/Land Cover (LULC) Mapping Products Generated from Satellite Remote Sensing. *ISPRS J. Photogramm. Remote Sens.* **2023**, *206*, 311–334. [CrossRef]
3. Zhao, Q.; Ji, L.; Su, Y.; Yu, K.; Zhao, Y. Monitoring changes to small-sized lakes using high spatial and temporal satellite imagery in the Badain Jaran Desert from 2015 to 2020. *Int. J. Crowd Sci.* **2025**, *in press*. [CrossRef]
4. Vivone, G.; Alparone, L.; Chanussot, J.; Dalla Mura, M.; Garzelli, A.; Licciardi, G.A.; Restaino, R.; Wald, L. A Critical Comparison Among Pansharpening Algorithms. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 2565–2586. [CrossRef]
5. Tu, T.M.; Huang, P.; Hung, C.L.; Chang, C.P. A Fast Intensity-Hue-Saturation Fusion Technique with Spectral Adjustment for IKONOS Imagery. *IEEE Geosci. Remote Sens. Lett.* **2004**, *1*, 309–312. [CrossRef]
6. Aiazzi, B.; Baronti, S.; Selva, M. Improving Component Substitution Pansharpening Through Multivariate Regression of MS +Pan Data. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 3230–3239. [CrossRef]
7. Meng, X.; Xiong, Y.; Shao, F.; Shen, H.; Sun, W.; Yang, G.; Yuan, Q.; Fu, R.; Zhang, H. A Large-Scale Benchmark Data Set for Evaluating Pansharpening Performance: Overview and Implementation. *IEEE Geosci. Remote Sens. Mag.* **2021**, *9*, 18–52. [CrossRef]
8. Aiazzi, B.; Alparone, L.; Baronti, S.; Garzelli, A.; Selva, M. MTF-tailored Multiscale Fusion of High-resolution MS and Pan Imagery. *Photogramm. Eng. Remote Sens.* **2006**, *72*, 591–596. [CrossRef]
9. Otazu, X.; Gonzalez-Audicana, M.; Fors, O.; Nunez, J. Introduction of Sensor Spectral Response into Image Fusion Methods. Application to Wavelet-Based Methods. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 2376–2385. [CrossRef]
10. Ballester, C.; Caselles, V.; Igual, L.; Verdera, J.; Rougé, B. A Variational Model for P+XS Image Fusion. *Int. J. Comput. Vis.* **2006**, *69*, 43–58. [CrossRef]
11. Vicinanza, M.R.; Restaino, R.; Vivone, G.; Dalla Mura, M.; Chanussot, J. A Pansharpening Method Based on the Sparse Representation of Injected Details. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 180–184. [CrossRef]
12. Palsson, F.; Sveinsson, J.R.; Ulfarsson, M.O. A New Pansharpening Algorithm Based on Total Variation. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 318–322. [CrossRef]
13. Chen, Y.; Wan, Z.; Chen, Z.; Wei, M. CSLP: A Novel Pansharpening Method Based on Compressed Sensing and L-PNN. *Inf. Fusion* **2025**, *118*, 103002. [CrossRef]
14. Wang, H.; Zhang, H.; Tian, X.; Ma, J. Zero-Sharpen: A Universal Pansharpening Method across Satellites for Reducing Scale-Variance Gap via Zero-Shot Variation. *Inf. Fusion* **2024**, *101*, 102003. [CrossRef]
15. Deng, L.j.; Vivone, G.; Paoletti, M.E.; Scarpa, G.; He, J.; Zhang, Y.; Chanussot, J.; Plaza, A. Machine Learning in Pansharpening: A Benchmark, from Shallow to Deep Networks. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 279–315. [CrossRef]
16. Ciotola, M.; Poggi, G.; Scarpa, G. Unsupervised Deep Learning-Based Pansharpening With Jointly Enhanced Spectral and Spatial Fidelity. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5405417. [CrossRef]
17. Scarpa, G.; Ciotola, M. Full-Resolution Quality Assessment for Pansharpening. *Remote Sens.* **2022**, *14*, 1808. [CrossRef]
18. Zhou, H.; Liu, Q.; Weng, D.; Wang, Y. Unsupervised Cycle-Consistent Generative Adversarial Networks for Pan Sharpening. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5408814. [CrossRef]
19. Lin, H.; Dong, Y.; Ding, X.; Liu, T.; Liu, Y. Unsupervised Pan-Sharpening via Mutually Guided Detail Restoration. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 3386–3394. [CrossRef]
20. Cao, Z.; Cao, S.; Deng, L.J.; Wu, X.; Hou, J.; Vivone, G. Diffusion Model with Disentangled Modulations for Sharpening Multispectral and Hyperspectral Images. *Inf. Fusion* **2024**, *104*, 102158. [CrossRef]
21. Rui, X.; Cao, X.; Pang, L.; Zhu, Z.; Yue, Z.; Meng, D. Unsupervised Hyperspectral Pansharpening via Low-Rank Diffusion Model. *Inf. Fusion* **2024**, *107*, 102325. [CrossRef]
22. Jiang, H.; Chen, Z. Transformer-Based Diffusion and Spectral Priors Model for Hyperspectral Pansharpening. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 18962–18977. [CrossRef]
23. Xiao, J.L.; Huang, T.Z.; Deng, L.J.; Lin, G.; Cao, Z.; Li, C.; Zhao, Q. Hyperspectral Pansharpening via Diffusion Models with Iteratively Zero-Shot Guidance. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 10–17 June 2025; pp. 12669–12678. [CrossRef]

24. Xiong, Z.; Li, W.; Zhao, X.; Zhang, B.; Tao, R.; Du, Q. PRF-Net: A Progressive Remote Sensing Image Registration and Fusion Network. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *36*, 9437–9450. [CrossRef]
25. Lee, J.; Seo, S.; Kim, M. SIPSA-Net: Shift-Invariant Pan Sharpening With Moving Object Alignment for Satellite Imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10166–10174.
26. Seo, S.; Choi, J.S.; Lee, J.; Kim, H.H.; Seo, D.; Jeong, J.; Kim, M. UPSNet: Unsupervised Pan-Sharpening Network With Registration Learning Between Panchromatic and Multi-Spectral Images. *IEEE Access* **2020**, *8*, 201199–201217. [CrossRef]
27. Dai, H.; Liu, X.; Qiao, Y.; Zheng, K.; Xiao, X.; Cai, Z. UFN-GAN: An Unsupervised Generative Adversarial Network for Remote Sensing Image Fusion. In Proceedings of the 2021 China Automation Congress (CAC), Beijing, China, 22–24 October 2021; pp. 1803–1808. [CrossRef]
28. Ciotola, M.; Vitale, S.; Mazza, A.; Poggi, G.; Scarpa, G. Pansharpening by Convolutional Neural Networks in the Full Resolution Framework. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5408717. [CrossRef]
29. Aiazzi, B.; Alparone, L.; Baronti, S.; Garzelli, A. Context-Driven Fusion of High Spatial and Spectral Resolution Images Based on Oversampled Multiresolution Analysis. *IEEE Trans. Geosci. Remote Sens.* **2002**, *40*, 2300–2312. [CrossRef]
30. Aimé, P.; Drumetz, L.; Mura, M.D.; Bajjouk, T.; Garello, R. Consistency and Ambiguities of Quality No Reference Metric for Pansharpening. In Proceedings of the IGARSS 2023—IEEE International Geoscience and Remote Sensing Symposium, Pasadena, CA, USA, 16–21 July 2023; pp. 5583–5586. [CrossRef]
31. Li, S.; Li, S.; Zhang, L. Hyperspectral and Panchromatic Images Fusion Based on the Dual Conditional Diffusion Models. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5526315. [CrossRef]
32. Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D.J.; Norouzi, M. Image Super-Resolution via Iterative Refinement. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 4713–4726. [CrossRef]
33. Niu, A.; Zhang, K.; Pham, T.X.; Sun, J.; Zhu, Y.; Kweon, I.S.; Zhang, Y. CDPMSR: Conditional Diffusion Probabilistic Models for Single Image Super-Resolution. *arXiv* **2023**, arXiv:2302.12831. [CrossRef]
34. Shang, S.; Shan, Z.; Liu, G.; Zhang, J. ResDiff: Combining CNN and Diffusion Model for Image Super-Resolution. *arXiv* **2023**, arXiv:2303.08714. [CrossRef]
35. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 6840–6851.
36. Song, J.; Meng, C.; Ermon, S. Denoising Diffusion Implicit Models. In Proceedings of the International Conference on Learning Representations, Online, 26 April–1 May 2020.
37. Liu, J.; Yuan, Z.; Pan, Z.; Fu, Y.; Liu, L.; Lu, B. Diffusion Model with Detail Complement for Super-Resolution of Remote Sensing. *Remote Sens.* **2022**, *14*, 4834. [CrossRef]
38. Zhang, L.; Rao, A.; Agrawala, M. Adding Conditional Control to Text-to-Image Diffusion Models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 3836–3847.
39. Shen, Z.; Zhang, M.; Zhao, H.; Yi, S.; Li, H. Efficient Attention: Attention With Linear Complexities. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Online, 5–9 January 2021; pp. 3531–3539.
40. Zampieri, A.; Charpiat, G.; Girard, N.; Tarabalka, Y. Multimodal Image Alignment through a Multiscale Chain of Neural Networks with Application to Remote Sensing. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 657–673.
41. Shekarforoush, H.; Berthod, M.; Zerubia, J. Subpixel Image Registration by Estimating the Polyphase Decomposition of Cross Power Spectrum. In Proceedings of the CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 18–20 June 1996; pp. 532–537. [CrossRef]
42. Stone, H. Progressive Wavelet Correlation Using Fourier Methods. *IEEE Trans. Signal Process.* **1999**, *47*, 97–107. [CrossRef]
43. Zhu, L.; Geng, X. A New Translation Matching Method Based on Autocorrelated Normalized Cross-Power Spectrum. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 6956–6968. [CrossRef]
44. Geng, X.; Yang, W. Cyclic Shift Matrix—A New Tool for the Translation Matching Problem. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8904–8913. [CrossRef]
45. Ciotola, M.; Scarpa, G. Fast Full-Resolution Target-Adaptive CNN-Based Pansharpening Framework. *Remote Sens.* **2023**, *15*, 319. [CrossRef]
46. Wald, L. *Data Fusion: Definitions and Architectures: Fusion of Images of Different Spatial Resolutions*; Presses des Mines: Paris, France, 2002.
47. Yuhas, R.H.; Goetz, A.F.H.; Boardman, J.W. *Discrimination among Semi-Arid Landscape Endmembers Using the Spectral Angle Mapper (SAM) Algorithm*; JPL: Grand Rapids, MI, USA, 1992.
48. Ji, L.; Geng, X.; Sun, K.; Zhao, Y.; Gong, P. Modified N-FINDR Endmember Extraction Algorithm for Remote-Sensing Imagery. *Int. J. Remote Sens.* **2015**, *36*, 2148–2162. [CrossRef]

49. Wang, Z.; Bovik, A.; Sheikh, H.; Simoncelli, E. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]
50. Liu, Q.; Zhou, H.; Xu, Q.; Liu, X.; Wang, Y. PSGAN: A Generative Adversarial Network for Remote Sensing Image Pan-Sharpening. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 10227–10242. [CrossRef]
51. Xiao, J.L.; Huang, T.Z.; Deng, L.J.; Wu, Z.C.; Wu, X.; Vivone, G. Variational Pansharpening Based on Coefficient Estimation With Nonlocal Regression. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5406115. [CrossRef]
52. Wu, Z.C.; Huang, T.Z.; Deng, L.J.; Vivone, G. A Framelet Sparse Reconstruction Method for Pansharpening with Guaranteed Convergence. *Inverse Probl. Imaging* **2023**, *17*, 1277–1300. [CrossRef]
53. Zhang, K.; Wang, A.; Zhang, F.; Wan, W.; Sun, J.; Bruzzone, L. Spatial-Spectral Dual Back-Projection Network for Pansharpening. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5402216. [CrossRef]
54. Diao, W.; Hu, L.; Zhang, K.; Xiao, L. Dual-Conditionally Guided Diffusion Models for Fusion of Unregistered Multisource Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5404117. [CrossRef]
55. Cao, Q.; Deng, L.J.; Wang, W.; Hou, J.; Vivone, G. Zero-Shot Semi-Supervised Learning for Pansharpening. *Inf. Fusion* **2024**, *101*, 102001. [CrossRef]
56. Rui, X.; Cao, X.; Li, Y.; Meng, D. Variational Zero-shot Multispectral Pansharpening. *arXiv* **2024**, arXiv:2407.06633. [CrossRef]
57. Alparone, L.; Aiazzi, B.; Baronti, S.; Garzelli, A.; Nencini, F.; Selva, M. Multispectral and Panchromatic Data Fusion Assessment Without Reference. *Photogramm. Eng. Remote Sens.* **2008**, *74*, 193–200. [CrossRef]
58. Zhou, J.; Civco, D.L.; Silander, J.A. A Wavelet Transform Method to Merge Landsat TM and SPOT Panchromatic Data. *Int. J. Remote Sens.* **1998**, *19*, 743–757. [CrossRef]
59. Hasituya.; Chen, Z.; Wang, L.; Liu, J. Selecting Appropriate Spatial Scale for Mapping Plastic-Mulched Farmland with Satellite Remote Sensing Imagery. *Remote Sens.* **2017**, *9*, 265. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.