NODiff: Neural Operator Diffusion for Multispectral Image Fusion

Junming Hou[1*], Ran Ran[2*], Sixing Chen[1], Zihao Chen[1], Xiaofeng Cong[1], Junling Li[1†], Liang-Jian Deng[2, 3†]

[1]Southeast University, Nanjing 211189, China
[2]University of Electronic Science and Technology of China, Chengdu 611731, China
[3]Multi-Hazard Early Warning Key Laboratory of Sichuan Province
junming_hou@seu.edu.cn, ranran@std.uestc.edu.cn, sixingchen@seu.edu.cn, 213233966@seu.edu.cn, cxf_svip@163.com, junlingli@seu.edu.cn, liangjian.deng@uestc.edu.cn

Abstract

Pansharpening is a powerful technique for generating high-resolution multispectral (HRMS) images by fusing currently available image pairs of low-resolution multispectral (LRMS) and texture-rich panchromatic (PAN) data, effectively addressing the physical constraints of satellite sensors. While recent generative diffusion models have demonstrated impressive performance gains in this domain, their prohibitive computational demands and training costs hinder practicality in resource-constrained remote sensing satellite systems. In this work, we propose NODiff, a novel diffusion framework that replaces the conventional attention-based denoising backbone with a neural operator, seamlessly integrating operator learning and generative modeling into an efficient yet effective solution for pansharpening. In practice, we implement our approach through a two-stage learning paradigm: First, we pretrain the proposed Neural Operator-based diffusion model to learn the high-resolution texture priors essential for pansharpening. Afterward, we freeze the pretrained parameters, and design a lightweight conditional detail guidance adapter to enable efficient fine-tuning for generating desired HRMS images. Meanwhile, a time-aware low-rank adaptation is introduced to dynamically refine high-frequency details potentially affected by spectral mode truncation. Extensive experiments on multiple benchmark datasets demonstrate that NODiff achieves competitive pansharpening performance while significantly reducing training and inference costs. Beyond pansharpening, our method provides new insights into building resource-efficient generative models.

Code — https://github.com/coder-JMHou/NODiff

Introduction

High-resolution remote sensing multispectral imagery has found widespread application across a broad spectrum of civilian and military domains, such as precision agriculture, environmental monitoring, mapping services, and target recognition (Meng et al. 2019; Vivone et al. 2024). Due to physical imaging constraints, current satellite multispectral sensors cannot directly acquire high-resolution multispectral (HRMS) data. As a common solution, they often observe a low-resolution (LR) MS image and its paired texture-rich panchromatic (PAN) image of the same scene, providing complementary spectral and spatial information, respectively. Pansharpening technology is developed to fuse the complementary information from PAN and MS modalities to generate HRMS images, thereby benefiting the downstream remote sensing applications.

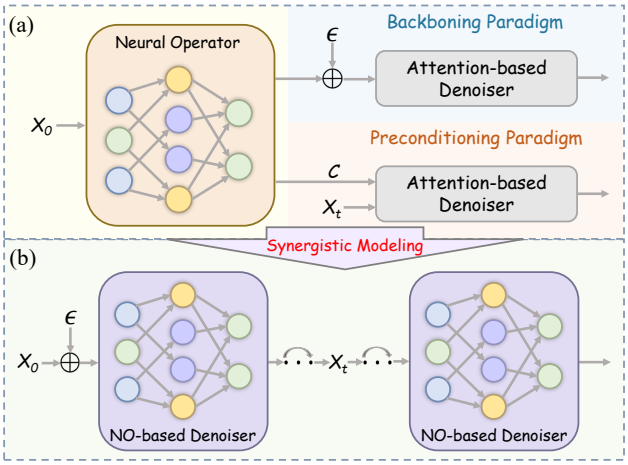

Figure 1: Comparison of integration schemes between Neural Operators and Diffusion Models: (a) Existing approaches (Liu and Tang 2025; Xu et al. 2025), and (b) Our design. Rather than the simple integration, our design delves into synergizing the strengths of two emerging neural architectures, resulting in an efficient and effective framework.

To date, the pansharpening field has undergone substantial methodological development, evolving from traditional model-driven algorithms to deep learning paradigms. Traditional approaches primarily include Component Substitution (CS), Multi-Resolution Analysis (MRA), and Variational Optimization (VO), according to their underlying fusion principles and algorithmic paradigms. However, these approaches often heavily rely on manually designed priors and unrealistic assumptions, leading to suboptimal fusion quality. With the growing availability of large-scale datasets, recent years have witnessed substantial advances in deep learning-based pansharpening methods, which significantly outperform traditional algorithms. These approaches exploit

*Contributed equally.
†Corresponding authors.

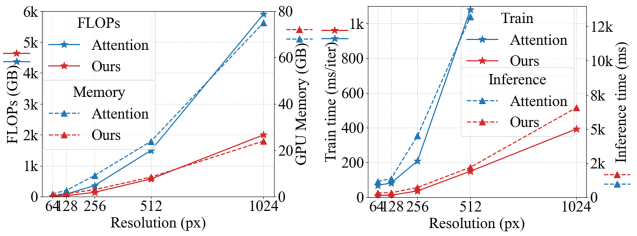

Figure 2: Comparison of computational cost and efficiency between our design and conventional attention-based denoising backbone, in terms of FLOPs, GPU memory usage, training time per iteration, and inference time.

rich training data to directly capture the complex non-linear mapping from input PAN–LRMS pairs to their corresponding HRMS outputs by training deep neural networks. In pursuit of improved performance, a variety of advanced network architectures have been proposed, spanning from convolutional neural networks (CNNs) to global transformers.

Very recently, diffusion models have been applied to pansharpening, yielding superior performance over conventional deep learning models. However, such performance gains come at the expense of substantial computational demands and training costs, posing significant challenges for practical applications, particularly in resource-constrained remote sensing satellites. Neural operators (NOs) are a recently developed neural architecture designed to solve partial differential equations (PDEs). They directly learn mappings between infinite-dimensional function spaces, providing resolution-invariant paradigms that generalize across varying input resolutions. Recent efforts have been made to integrate NOs and diffusion models for image super-resolution, as summarized in Figure 1: (1) Preconditioning: NOs act as preprocessors (*e.g.,* up-sample operator) to produce conditional detail guidance for diffusion model (Liu and Tang 2025); (2) Backboning: NOs serve as the super-resolution backbone, refined subsequently by diffusion processes (Xu et al. 2025). While promising, these approaches inherit the computational burden and inefficiency of diffusion models, limiting their practical applicability.

In this work, we build upon a novel perspective to synergize the strengths of these two emerging neural paradigms. Instead of simple integration, our approach assimilates the neural operator as the denoising backbone of the diffusion model by replacing conventional attention-based architectures, resulting in an efficient yet effective NO-based diffusion framework. As shown in Figure 2, our NO-based design demonstrates competitive computational efficiency compared to conventional attention-based counterparts. In practice, we implement it through a two-stage training paradigm: First, we pretrain the proposed NO-based diffusion model to steer the diffusion generative process toward learning high-resolution texture priors, which are pivotal for pansharpening. Second, we introduce a straightforward conditional detail guidance adapter that takes LRMS and PAN image pairs as inputs to fine-tune the pretrained NO-based diffusion model for generating corresponding HRMS images. Meanwhile, a lightweight time-aware low-rank adaptation (LoRA) technique is employed to further refine high-frequency details that may be affected due to spectral mode truncation. Our method, built upon these two core processes, achieves competitive pansharpening performance compared to existing advanced approaches. Notably, compared to conventional attention-based diffusion models, our design achieves a favorable trade-off between efficiency and effectiveness by integrating the complementary strengths of two emerging and promising neural paradigms. We summarize the main contributions as follows:

- We propose NODiff, a novel diffusion framework that replaces costly attention-based denoiser architectures with resolution-invariant neural operator, thereby synergizing operator learning with generative modeling into an efficient yet effective solution for pansharpening.
- We introduce a lightweight conditional detail guidance adapter for parameter-efficient fine-tuning of the pretrained knowledge-rich diffusion model, complemented by a time-aware low-rank adaptation mechanism that dynamically refines high-frequency textures affected by spectral mode truncation, enabling the generation of high-quality HRMS images.
- Extensive experiments on multiple benchmark datasets demonstrate that our method outperforms current advanced pansharpening approaches, and also delivers a promising performance-efficiency balance compared to conventional attention-based diffusion counterparts.

Related Works

Deep Learning-based Pansharpening. In recent years, deep learning has emerged as the prevailing paradigm in the pansharpening domain owing to its powerful nonlinear learning capability, greatly surpassing traditional algorithms (Dian et al. 2021; Vivone et al. 2024). The implementation of PNN indicates the superiority of deep learning methods, sparking the development of diverse advanced network architectures, including convolutional neural networks (CNNs) (Yang et al. 2017; Deng et al. 2020; Hu et al. 2021; Hou et al. 2025), global transformers (Zhou et al. 2022; Hou et al. 2024; Liu, Dian, and Li 2025; Wu et al. 2025), and recently emerging generative diffusion models (Meng et al. 2023; Cao et al. 2024; Kim et al. 2025).

Image Super-resolution Neural Operator. As newly developed data-driven solutions for PDEs, neural operators directly learn mappings between infinite-dimensional function spaces (Kovachki et al. 2023; Azizzadenesheli et al. 2024). By optimizing network training in function spaces, NOs achieve superior nonlinear approximation and exhibit invariance to discretization (Lu et al. 2021; Li et al. 2021; Tran et al. 2023). More importantly, this resolution-invariant property enables NOs to generalize well across discretizations, thus making them particularly well-suited for resolution-varying tasks. For example, SRNO first presents a NO-based continuous image super-resolution approach (Wei and Zhang 2023). HiNOTE introduces a hierarchical NO framework for arbitrary-scale super-resolution of scientific data (Luo, Qian, and Yoon 2024). He et al. apply this novel

neural architecture to continuous-resolution hyperspectral pansharpening (He et al. 2025a). Very recently, several studies have integrated neural operators with diffusion models to harness their complementary strengths, yielding promising results (Liu and Tang 2025; Xu et al. 2025). Nonetheless, these approaches still inherit the substantial computational overhead characteristic of diffusion models. By contrast, we employ NOs as highly efficient and effective noise predictors by replacing the conventional attention-based counterparts, achieving a desirable performance-efficiency balance. Most importantly, this offers a novel perspective for combining these two emerging neural architectures.

Diffusion Models for Pansharpening. As a prevailing generative model, denoising diffusion probabilistic models (DDPMs) (Ho, Jain, and Abbeel 2020) have found wide applications in various domains, such as text-to-image generation (Saharia et al. 2022; Gu et al. 2022) and image editing (Kawar et al. 2023; Couairon et al. 2022; Shi et al. 2024). In addition, DDPMs have also shown notable efficacy in image processing tasks. Building on this, Song et al. (Song, Meng, and Ermon 2021) propose the denoising diffusion implicit model (DDIM) which employs a non-Markovian sampling process to accelerate diffusion model inference. In recent years, diffusion models have attracted growing interest in the pansharpening field, with PAN and LRMS serving as conditioning information (Kim et al. 2025; Cao et al. 2024). Although these methods exhibit superior fusion capabilities over conventional neural networks, they often require greater computational resources and training costs, which hinder their broader applications. This motivates us to explore more efficient variants that are suitable for resource-constrained remote sensing satellites.

Methodology

Overview of the Proposed Method

Accurately reconstructing fine-grained texture details is a key objective of pansharpening learning, as required by downstream remote sensing applications. Very recently, diffusion models have been introduced to learn the high-resolution residuals between LRMS and HRMS images, demonstrating impressive pansharpening performance improvements (Cao et al. 2024; Kim et al. 2025). However, these methods require substantial computational resources and training costs, limiting their practicality on resource-constrained remote sensing satellites. Our approach seeks to remedy this by incorporating high-resolution texture-specific priors into a computationally efficient diffusion model. To achieve this, we implement our method through a two-stage learning paradigm illustrated in Figure 3: (1) High-resolution Texture Prior Learning: First, we pretrain the proposed Neural Operator-based conditional diffusion model to learn high-resolution texture priors. (2) Conditional Detail Guidance Fine-Tuning: Following that, we use a straightforward conditional detail guidance adapter to fine-tune the pretrained diffusion model for generating corresponding HRMS images, aided by a time-aware LoRA. Our method synergizes the unique strengths of two emerging and powerful neural architectures, offering new insights for building efficient yet effective generative pansharpening approaches suitable for resource-constrained satellites.

Learning High-Resolution Diffusion Texture Priors

In this stage, our primary objective is to integrate high-resolution texture-specific priors into a diffusion model, enhancing its ability to reconstruct fine-grained textures while maintaining computational efficiency. Motivated by recent advances in integrating diffusion models with neural operators for image super-resolution (Liu and Tang 2025; Xu et al. 2025), we seek to explore this promising synergy to construct an efficient generative pansharpening model. To this end, we revisit the unique strengths of these two emerging neural architectures and identify the key aspects particularly suited for super-resolution scenarios: 1) the Neural Operators' function-space mapping capacity and resolution-invariance; and 2) the diffusion model's generative ability to produce fine-grained details. Building upon these insights, we develop an efficient yet effective Neural Operator-based diffusion model, where the denoising network is constructed by cascaded Fourier Neural Operator (FNO) layers (Li et al. 2021) shown in Figure 3, replacing conventional attention-heavy architectures.

Let $\mathbf{L} \in \mathbb{R}^{H\times W\times s}$ denote the up-sampled LRMS image, and $\mathbf{H} \in \mathbb{R}^{H\times W\times s}$ the corresponding reference HRMS image. We define the residual image $\mathbf{X}_0 \in \mathbb{R}^{H\times W\times s}$ as $\mathbf{H}-\mathbf{L}$, which contains rich texture details absent in $\mathbf{L}$. Similar to (Cao et al. 2024; Kim et al. 2025), the forward diffusion process adds Gaussian noise to $\mathbf{X}_0$ over T iterations in the Markov steps:

$$q(\mathbf{X}_t|\mathbf{X}_{t-1}) = \mathcal{N}\left(\mathbf{X}_t; \sqrt{1-\beta_t}\mathbf{X}_{t-1}, \beta_t I\right), \quad (1)$$

where $\mathbf{X}_t$ denotes the noisy residual image at step t, and β_t controls the noise variance. Through reparameterization, $\mathbf{X}_t$ can be obtained directly by adding noise in one step to $\mathbf{X}_0$:

$$\mathbf{X}_t = \sqrt{\bar{\alpha}_t}\mathbf{X}_0 + \sqrt{(1-\bar{\alpha}_t)}\boldsymbol{\epsilon}, \;\; \boldsymbol{\epsilon} \sim \mathcal{N}(0, I), \quad (2)$$

where $\boldsymbol{\epsilon}$ denotes a standard Gaussian distribution and $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$, $\alpha_t = 1-\beta_t$. The reverse process samples $\mathbf{X}_0$ from pure noise $\mathbf{X}_T \sim \mathcal{N}(0, I)$ step-by-step. The posterior distribution of the denoising process is written as:

$$\begin{aligned} q(\mathbf{X}_{t-1}|\mathbf{X}_t, \mathbf{X}_0) &= \mathcal{N}\left(\mathbf{X}_{t-1}; \mu_t(\mathbf{X}_t, \mathbf{X}_0), \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t I\right), \\ \mu_t(\mathbf{X}_t, \mathbf{X}_0) &= \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{X}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_t\right), \end{aligned} \quad (3)$$

where $\boldsymbol{\epsilon}_t$ is the t-th step added noise in the forward process.

To address the efficiency–performance trade-off in conventional attention-based denoising architectures, we design an efficient FNO-based denoiser ϵ_θ that predicts the noise at each step. We use the $\mathbf{H}$ as high-resolution conditional guidance by concatenating it with the noisy image $\mathbf{X}_t$:

$$\mathbf{F}_t = \mathrm{Conv}([\mathbf{X}_t, \mathbf{H}]). \quad (4)$$

Subsequently, the resulting $\mathbf{F}_t$ is processed by the denoising network composed of cascaded FNO layers, as follows:

$$\mathbf{F}_t^l = \sigma\Big(\underbrace{\mathcal{F}^{-1}\big(\mathcal{P}_{\mathcal{K}}(\mathcal{F}(\mathbf{F}_t^{l-1}))\odot \mathrm{R}\big)}_{\text{Global Spectral Operator}} + \underbrace{\mathrm{W}\,\mathbf{F}_t^{l-1}}_{\text{Linear Transform}}\Big), \quad (5)$$

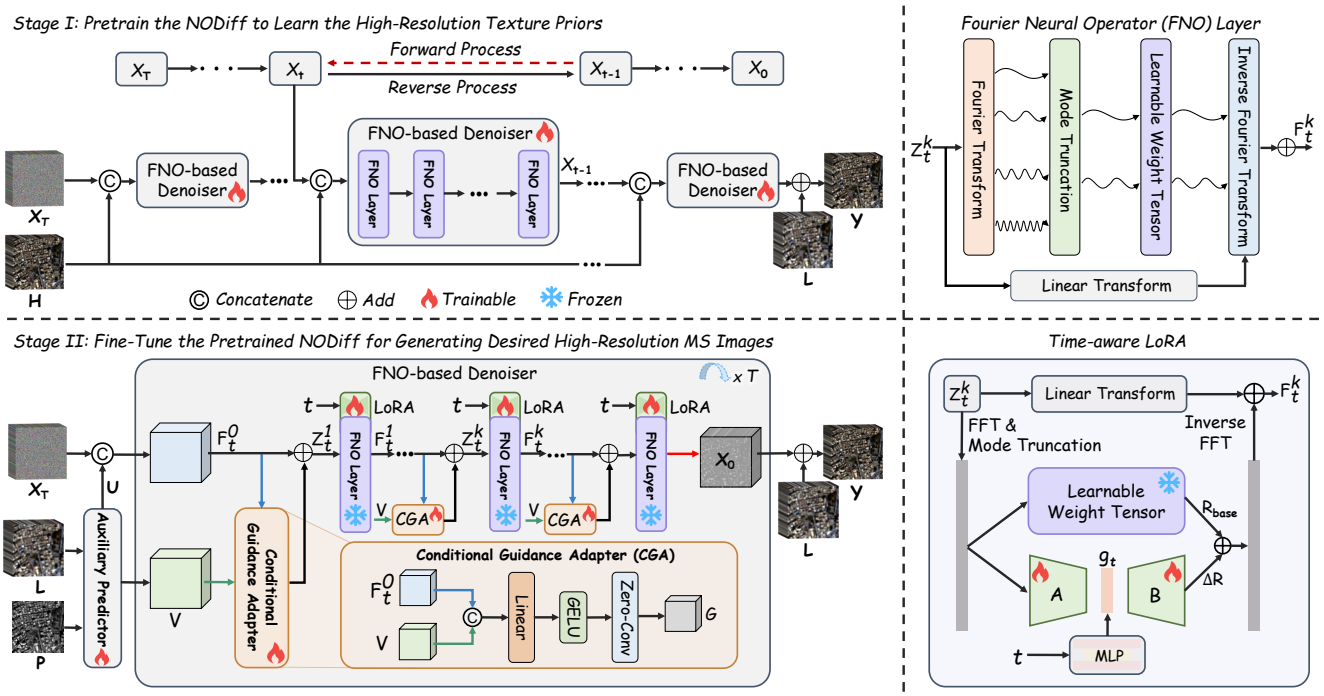

Figure 3: The framework of the proposed NODiff which is implemented through two main procedures. Firstly, we pretrain the NODiff to learn high-resolution diffusion texture priors conditioned on the HRMS images. Subsequently, we introduce a lightweight conditional detail guidance adapter using LRMS and PAN image pairs as inputs to fine-tune the pretrained NODiff for generating corresponding HRMS images, assisted by a time-aware LoRA design.

where $\mathbf{F}_t^l (l = 1, 2, \cdots, m)$ represents the features of the l-th FNO layer. $\mathcal{F}(\cdot)$ and $\mathcal{F}^{-1}(\cdot)$ denote the Fourier and inverse Fourier transforms, respectively. $\mathcal{P}_{\mathcal{K}}(\cdot)$ signifies the spectral mode truncation operation. R is a learnable complex weight tensor. W and $\sigma(\cdot)$ represent the linear transformation and nonlinear activation.

After iterative refinement, we project the last FNO layer's output to obtain the predicted noise at step t:

$$\hat{\boldsymbol{\epsilon}}_t = \text{Conv}(\mathbf{F}_t^m). \tag{6}$$

With the predicted noise $\hat{\boldsymbol{\epsilon}}_t$, the reverse process is parameterized as:

$$\begin{aligned} \mathbf{X}_{t-1} &= \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{X}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(\mathbf{X}_t, \mathbf{H}, t) \right) \\ &= \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{X}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \hat{\boldsymbol{\epsilon}}_t \right). \end{aligned} \tag{7}$$

After T-step denoising process, we can reconstruct the high-resolution multispectral image $\mathbf{Y} \in \mathbb{R}^{H \times W \times s}$ by adding the predicted residual image $\mathbf{X}_0$ to $\mathbf{L}$. Accordingly, the optimization objective is defined as:

$$\mathcal{L}_{\text{Stage I}} = \mathbb{E} \left\| \boldsymbol{\epsilon} - \epsilon_\theta \left(\sqrt{\bar{\alpha}_t}(\mathbf{H} - \mathbf{L}) + \sqrt{(1-\bar{\alpha}_t)}\boldsymbol{\epsilon}, \mathbf{H}, t \right) \right\|_1. \tag{8}$$

Conditional Detail Guidance Fine-Tuning

The pretraining stage enables NODiff to encapsulate rich prior knowledge, which we aim to fully retain and leverage during downstream pansharpening. To achieve this, we freeze the parameters of the pretrained FNO-based denoiser and introduce a flexible and parameter-efficient adaptation strategy implemented through the cross-modality detail guidance adapter and time-aware low-rank adaptation illustrated in Figure 3. Specifically, we first introduce a straightforward auxiliary predictor consisting of several convolution blocks, which takes the Sobel-filtered PAN and LRMS as inputs to produce: 1) The pseudo super-resolved (SR) image to replace the $\mathbf{H}$ in the Stage I; and 2) The cross-modality detail features as the input of the conditional detail guidance adapter along with the latent noisy feature $\mathbf{X}_t$:

$$\begin{aligned} \mathbf{V} &= \Phi(\mathcal{S}(\mathbf{P}, \mathbf{L})), \\ \mathbf{U} &= \text{Conv}(\mathbf{V}) + \mathbf{L}, \end{aligned} \tag{9}$$

where $\mathcal{S}(\cdot)$ denotes the mapping of Sobel operator, and $\Phi(\cdot)$ represents the backbone of the auxiliary predictor, where $\mathbf{V} \in \mathbb{R}^{H \times W \times C}$ is the extracted cross-modality detail features. $\mathbf{U}$ is the initially predicted pseudo SR used to replace the high-resolution reference image $\mathbf{H}$.

At the diffusion step t, we first concatenate $\mathbf{U}$ with the noisy feature $\mathbf{X}_t$ along the channel dimension, and then fuse them through a convolution layer, resulting in the input feature $\mathbf{F}_t^0 \in \mathbb{R}^{H \times W \times C}$ of the FNO-based denoiser:

$$\mathbf{F}_t^0 = \text{Conv}([\mathbf{X}_t, \mathbf{U}]). \tag{10}$$

Then, we inject the cross-modality detail feature $\mathbf{V}$ into each intermediate layer of the denoiser through a conditional detail guidance adapter, which is realized through a bidirectional information flow interaction. Specifically, for the l-th

Method	Reduced Resolution				Full Resolution		
	PSNR(± std)	SAM(± std)	ERGAS(± std)	$Q2^n$(± std)	D_λ(± std)	D_s(± std)	HQNR(± std)
BDSD-PC (Vivone 2019)	32.969±2.784	5.429±1.823	4.698±1.617	0.829±0.097	0.063±0.024	0.073±0.036	0.870±0.053
BT-H (Lolli et al. 2017)	33.080±2.880	4.920±1.425	4.579±1.496	0.832±0.094	0.057±0.023	0.081±0.037	0.867±0.054
LRTCFPan (Wu et al. 2023)	33.613±2.839	4.737±1.412	4.315±1.442	0.846±0.091	0.018±0.007	0.053±0.026	0.931±0.031
FusionNet (Deng et al. 2020)	38.042±2.592	3.325±0.698	2.467±0.645	0.904±0.090	0.024±0.009	0.036±0.014	0.941±0.020
LAGConv (Jin et al. 2022)	38.568±2.789	3.104±0.559	2.300±0.613	0.910±0.091	0.037±0.015	0.042±0.015	0.923±0.025
Fourmer (Zhou et al. 2023)	38.268±2.727	3.236±0.681	2.419±0.665	0.911±0.090	0.022±0.010	0.035±0.004	0.944±0.013
HFIN (Tan et al. 2024)	38.534±2.786	3.088±0.635	2.306±0.557	0.912±0.089	0.025±0.008	0.043±0.017	0.934±0.024
HOIF (Zhou et al. 2024)	38.352±2.855	3.186±0.643	2.385±0.668	0.913±0.086	0.039±0.021	0.039±0.010	0.924±0.026
PanDiff (Meng et al. 2023)	37.860±2.773	3.316±0.674	2.492±0.656	0.906±0.088	0.027±0.012	0.054±0.026	0.920±0.036
PanMamba (He et al. 2025b)	39.012±2.818	2.914±0.592	2.184±0.521	0.920±0.085	0.018±0.007	0.031±0.011	0.952±0.015
ADWM (Huang et al. 2025)	39.170±2.878	2.914±0.589	2.145±0.531	0.919±0.086	0.024±0.010	**0.029±0.015**	0.948±0.021
NODiff (ours)	**39.314±2.898**	**2.859±0.585**	**2.109±0.526**	**0.923±0.083**	**0.014±0.005**	0.030±0.010	**0.956±0.012**
Ideal value	∞	**0**	**0**	**1**	**0**	**0**	**1**

Table 1: Quantitative results for reduced and full resolution WV3 samples, comparing several representative state-of-the-art methods. Bold: Best; Underline: Second best.

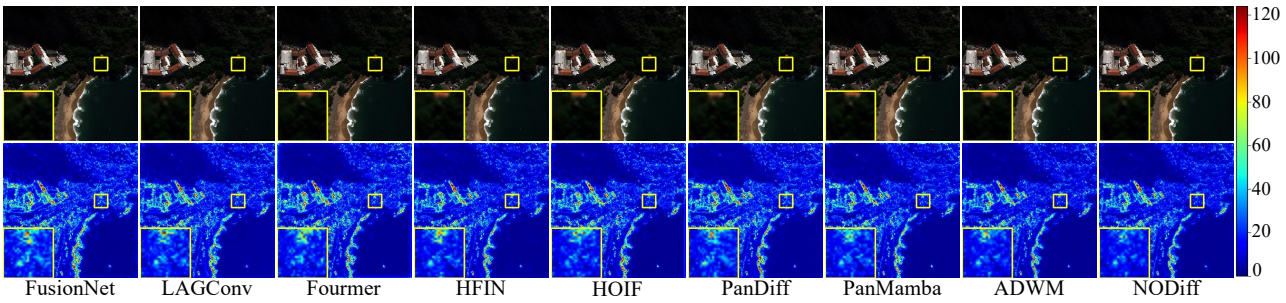

Figure 4: The visual results (top) and mean absolute error maps (bottom) of all compared DL-based methods on a reduced resolution WV3 testing sample.

FNO layer of the denoiser, the adapter first integrates the $(l-1)$-th output noisy feature $\mathbf{F}_t^{l-1} \in \mathbb{R}^{H\times W\times C}(l = 1, 2, \cdots, m)$ with $\mathbf{V}$ to produce a residual guidance feature $\mathbf{G} \in \mathbb{R}^{H\times W\times C}$, which is then added to $\mathbf{F}_t^{l-1}$ to form the input feature $\mathbf{Z}_t^l \in \mathbb{R}^{H\times W\times C}$ for the l-th FNO layer:

$$\begin{aligned} \mathbf{G} &= \phi([\mathbf{V}, \mathbf{F}_t^{l-1}]), \\ \mathbf{Z}_t^l &= \mathbf{F}_t^{l-1} + \mathbf{G}, \end{aligned} \tag{11}$$

where $\phi(\cdot)$ denotes the mapping of the detail guidance adapter as illustrated in Figure 3.

In parallel, we introduce a time-aware low-rank adaptation (Ta-LoRA) technique into the spectral convolution of the FNO layer to further refine high-frequency details that may be affected owing to mode truncation. Let R_{base} be the complex weight tensor corresponding to the retained mode set (frozen). A timestep embedding is first mapped into a rank-wise vector $\mathbf{g}_t \in \mathbb{R}^r$ using an MLP layer:

$$\mathbf{g}_t = 1 + \mathrm{Tanh}(\mathrm{MLP}(\mathrm{Embed}(t))). \tag{12}$$

Building on the diagonalized form $\mathrm{D}_t = \mathrm{diag}(\mathbf{g}_t) \in \mathbb{R}^{r\times r}$, the low-rank factors $\mathrm{A} \in \mathbb{R}^{r\times C_{\text{in}}}$ and $\mathrm{B} \in \mathbb{C}^{C_{\text{out}}\times r\times m_x\times m_y}$, and a scaling factor α, the time-modulated low-rank increment is computed as:

$$\Delta\mathrm{R} = \alpha \cdot \mathrm{AD}_t\mathrm{B}. \tag{13}$$

Thus, the final effective spectral weights are updated as:

$$\mathrm{R} = \mathrm{R}_{\text{base}} + \Delta\mathrm{R}. \tag{14}$$

Notably, the increment $\Delta\mathrm{R}$ is applied exclusively to the retained spectral modes including positive/negative bands, leaving the pretrained FNO dataflow unchanged. Thus, this Ta-LoRA incurs negligible additional parameters and computational overhead. Moreover, the time-aware vector g_t dynamically adapts the LoRA increment across different noise levels, making the later iterations more sensitive and effective for high-frequency details. The optimization objective is formally identical to that of Stage I and is expressed as:

$$\mathcal{L}_{\text{Stage II}} = \mathbb{E}\left\|\boldsymbol{\epsilon} - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}(\mathbf{H}-\mathbf{L}) + \sqrt{(1-\bar{\alpha}_t)}\boldsymbol{\epsilon}, \mathbf{P}, \mathbf{L}, t\right)\right\|_1. \tag{15}$$

Experiments

Datasets, Metrics and Implementation Details. We experiment on three popular benchmark datasets, including the WorldView-3 (WV3), GaoFen-2 (GF2), and QuickBird (QB), to evaluate our model's efficacy. The training datasets are simulated using observations from the original satellite imagery according to Wald's protocol. All the training and test data are publicly available at the PanCollection (Deng et al. 2022). The reduced resolution results are evaluated

Method	Reduced Resolution-GF2				Reduced Resolution-QB			
	PSNR(± std)	SAM(± std)	ERGAS(± std)	$Q2^n$(± std)	PSNR(± std)	SAM(± std)	ERGAS(± std)	$Q2^n$(± std)
BDSD-PC (Vivone 2019)	35.180±2.317	1.681±0.360	1.667±0.445	0.892±0.035	32.547±3.202	8.089±1.980	7.515±0.800	0.831±0.090
BT-H (Lolli et al. 2017)	36.054±2.236	1.649±0.360	1.528±0.409	0.918±0.025	32.648±3.273	7.194±1.552	7.401±0.838	0.833±0.088
LRTCFPan (Wu et al. 2023)	37.599±2.331	1.298±0.312	1.272±0.343	0.935±0.030	33.260±3.272	7.187±1.711	6.928±0.812	0.855±0.087
FusionNet (Deng et al. 2020)	39.639±2.270	0.974±0.212	0.988±0.222	0.964±0.009	37.532±2.518	4.923±0.908	4.159±0.321	0.925±0.090
LAGConv (Jin et al. 2022)	42.735±1.447	0.786±0.148	0.687±0.113	0.980±0.009	38.181±2.456	4.547±0.830	3.826±0.420	0.934±0.088
Fourmer (Zhou et al. 2023)	40.670±1.903	0.976±0.209	0.885±0.185	0.970±0.011	36.797±2.597	5.079±1.009	4.494±0.410	0.924±0.080
HFIN (Tan et al. 2024)	42.189±1.752	0.843±0.148	0.735±0.126	0.977±0.011	38.247±2.403	4.542±0.805	3.813±0.322	0.934±0.085
HOIF (Zhou et al. 2024)	40.982±1.802	0.943±0.205	0.841±0.162	0.974±0.009	38.242±2.119	4.521±0.811	3.825±0.510	0.933±0.094
PanDiff (Meng et al. 2023)	42.326±1.635	0.875±0.133	0.727±0.115	0.981±0.008	38.538±2.401	4.524±0.792	3.688±0.343	0.937±0.084
PanMamba (He et al. 2025b)	42.907±1.811	0.743±0.156	0.684±0.129	0.982±0.008	37.356±2.570	4.625±0.904	4.277±0.779	0.929±0.085
ADWM (Huang et al. 2025)	43.884±1.714	0.672±0.130	0.597±0.107	0.985±0.006	38.466±2.420	4.450±0.809	3.705±0.346	0.937±0.085
NODiff (ours)	**44.262±1.833**	**0.645±0.129**	**0.573±0.110**	**0.986±0.006**	**38.700±2.446**	**4.372±0.787**	**3.607±0.328**	**0.938±0.085**
Ideal value	∞	**0**	**0**	**1**	∞	**0**	**0**	**1**

Table 2: Quantitative results for reduced resolution GF2 and QB samples, comparing several representative state-of-the-art methods. Bold: Best; Underline: Second best.

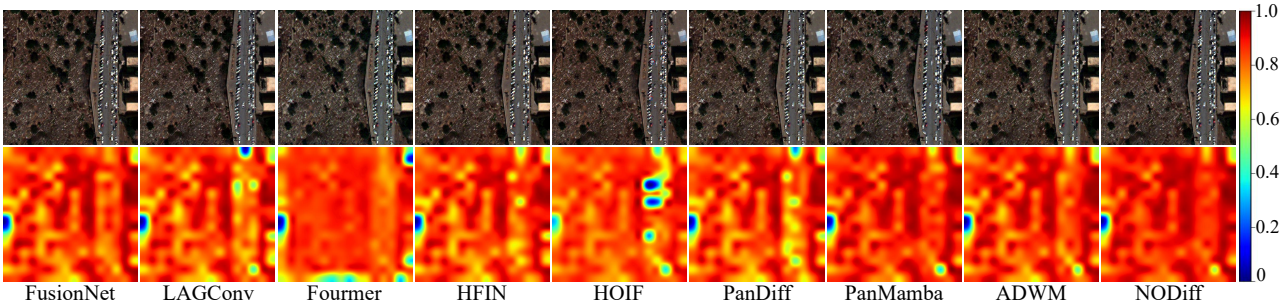

Figure 5: The visual results (top) and the corresponding HQNR maps (bottom) of all compared DL-based methods on a full resolution WV3 testing sample.

using four widely adopted reference-based metrics: PSNR, SAM (Yuhas, Goetz, and Boardman 1992), ERGAS (Wald 2002), and Q2n (Zhou, Civco, and Silander 1998). The real-world full resolution performance is compared using three non-reference indicators: D_λ, D_s, and HQNR (Arienzo et al. 2022). All experiments are conducted using the PyTorch framework on an NVIDIA GeForce RTX 4090 GPU. Additional training details and experimental results are provided in the supplementary material due to page limitations.

Baselines. We compare NODiff with several state-of-the-art DL-based pansharpening methods, including FusionNet (Deng et al. 2020), LAGConv (Jin et al. 2022), Fourmer (Zhou et al. 2023), HFIN (Tan et al. 2024), HOIF (Zhou et al. 2024), PanDiff (Meng et al. 2023), PanMamba (He et al. 2025b), and ADWM (Huang et al. 2025). Additionally, three classical algorithms are selected for a comprehensive comparison: BT-H (Lolli et al. 2017), BDSD-PC (Vivone 2019), and LRTCFPan (Wu et al. 2023).

Comparison With SOTA Methods

Evaluation on Reduced Resolution Scene. We begin by quantitatively evaluating the similarity between the fused images and ground truth in the reduced resolution setting. As presented in Table 1 (the left panel) and Table 2, our model consistently outperforms the compared cutting-edge pansharpening methods across all evaluation metrics on three

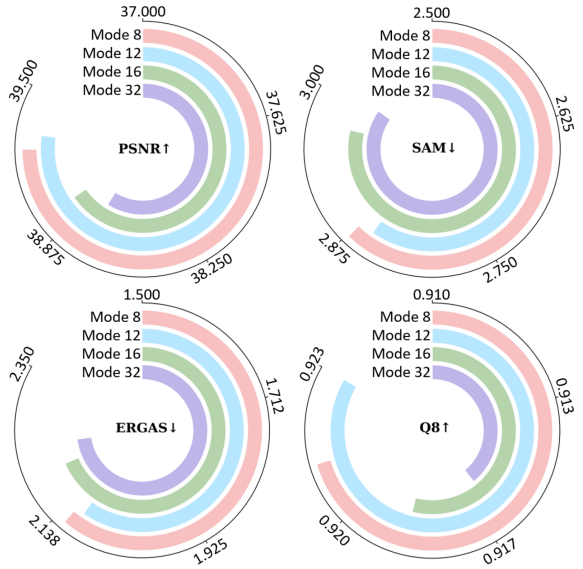

Figure 6: Ablation study on the spectral mode truncation.

datasets. Figure 4 provides the RGB visualizations and the corresponding Mean Absolute Error (MAE) residual maps compared to the ground truth for a WV3 example. Our

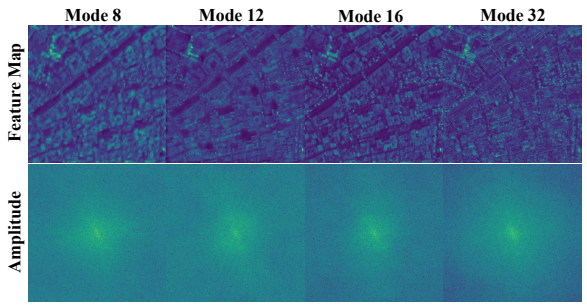

Figure 7: Features and Amplitude under mode truncation.

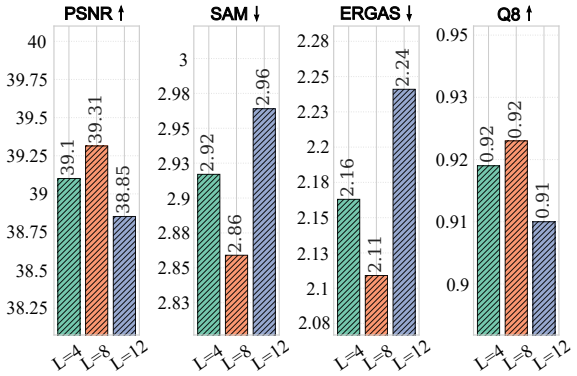

Figure 8: Ablation study on the FNO layer number.

model produces images with minimal aberrations, as evidenced by its sparse MAE residues. These results provide compelling evidence of the superiority of our method.

Evaluation on Full Resolution Scene. We evaluate the real-world applicability of NODiff using full resolution data. As shown in the right portion of Table 1, it consistently achieves leading performance across most metrics on the WV3 dataset, reaffirming the encouraging trends observed at reduced resolution assessment. In particular, it yields the highest HQNR score compared to both traditional and deep learning baselines, indicating superior fusion quality with respect to spectral fidelity and spatial detail. This is corroborated by the HQNR visualization shown in Figure 5, where our result exhibits a uniformly dark tone with sparse bright spots, indicating minimal spatial-spectral distortion. These full resolution results further substantiate the robust generalization capability of our method in real-world scenarios.

Ablation Study

Effect of the Spectral Mode Truncation. We first investigate the impact of mode truncation in spectral convolution. Specifically, we compare four configurations with mode values set to 8, 12, 16, and 32. As shown in Figure 6, our method with the mode=12 consistently achieves the best performance across all evaluation metrics. As visualized in Figure 7, the feature map of our method exhibits clear and finer textures, along with a more balanced frequency distribution. In comparison, higher spectral modes (*e.g.*, 16 and 32) introduce unwanted high-frequency noise, leading to noticeable artifacts in the feature maps; whereas a lower mode (8) results in overly smooth features and loss of important details.

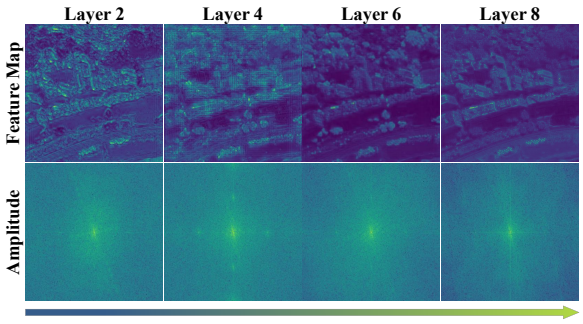

Figure 9: Features and Amplitude across FNO layers.

Method	PSNR	SAM	ERGAS	Q8
w/o Stage I	38.865	2.992	2.200	0.917
w/o CGA	38.681	3.028	2.356	0.911
w/o Ta-LoRA	<u>39.032</u>	<u>2.957</u>	<u>2.185</u>	<u>0.919</u>
NODiff (Ours)	**39.314**	**2.859**	**2.109**	**0.923**

Table 3: Effect of the training and fine-tuning designs.

Effect of the FNO Layer Number. We examine the effect of varying the number of FNO layers. As summarized in Figure 8, the model achieves optimal performance with 8 layers, and the feature maps become progressively clearer as depth increases (Figure 9), highlighting our model's expressiveness. Further increasing the depth leads to a performance drop, possibly because the amplified high-frequency deviation caused by spectral truncation diminishes sensitivity to finer textures. Notably, the model with only 4 layers also performs competitively over other cutting-edge approaches.

Effect of the Fine-tuning Designs. We analyze the impact of different training paradigms and fine-tuning designs. As reported in Table 3, the two-stage pretraining-fine-tuning pipeline significantly outperforms the one-stage training scheme. Moreover, incorporating both CGA and Ta-LoRA yields the best performance across all metrics, demonstrating the effectiveness of each fine-tuning component.

Conclusion

We propose an efficient yet effective Neural Operator–based diffusion framework that seamlessly synergizes operator learning and generative modeling to learn high-resolution diffusion texture priors for pansharpening. To enable task-specific adaptation, we introduce a conditional detail guidance adapter for parameter-efficient fine-tuning of texture generation, complemented by a time-aware low-rank adaptation to remedy the potential negative effects of spectral mode truncation. Building on these ingredients, our method achieves superior performance, and offers a scalable solution for resource-efficient generative modeling.

Acknowledgements

The work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62301151, in part by the Project of the Department of Science and Technology of Sichuan Province under Grant 2025YFNH0001.

References

Arienzo, A.; Vivone, G.; Garzelli, A.; Alparone, L.; and Chanussot, J. 2022. Full-resolution quality assessment of pansharpening: Theoretical and hands-on approaches. *IEEE Geoscience and Remote Sensing Magazine*, 10(3): 168–201.

Azizzadenesheli, K.; Kovachki, N.; Li, Z.; Liu-Schiaffini, M.; Kossaifi, J.; and Anandkumar, A. 2024. Neural operators for accelerating scientific simulations and design. *Nature Reviews Physics*, 6(5): 320–328.

Cao, Z.; Cao, S.; Deng, L.-J.; Wu, X.; Hou, J.; and Vivone, G. 2024. Diffusion model with disentangled modulations for sharpening multispectral and hyperspectral images. *Information Fusion*, 104: 102158.

Couairon, G.; Verbeek, J.; Schwenk, H.; and Cord, M. 2022. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*.

Deng, L.-J.; Vivone, G.; Jin, C.; and Chanussot, J. 2020. Detail injection-based deep convolutional neural networks for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 59(8): 6995–7010.

Deng, L.-J.; Vivone, G.; Paoletti, M. E.; Scarpa, G.; He, J.; Zhang, Y.; Chanussot, J.; and Plaza, A. 2022. Machine learning in pansharpening: A benchmark, from shallow to deep networks. *IEEE Geoscience and Remote Sensing Magazine*, 10(3): 279–315.

Dian, R.; Li, S.; Sun, B.; and Guo, A. 2021. Recent advances and new guidelines on hyperspectral and multispectral image fusion. *Information Fusion*, 69: 40–51.

Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; and Guo, B. 2022. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10696–10706.

He, L.; Fang, Z.; Li, J.; Ye, H.; and Plaza, A. 2025a. Arbitrary-Resolution Hyperspectral Pansharpening Neural Operators. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–19.

He, X.; Cao, K.; Zhang, J.; Yan, K.; Wang, Y.; Li, R.; Xie, C.; Hong, D.; and Zhou, M. 2025b. Pan-mamba: Effective pan-sharpening with state space model. *Information Fusion*, 115: 102779.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hou, J.; Cao, Z.; Zheng, N.; Li, X.; Chen, X.; Liu, X.; Cong, X.; Hong, D.; and Zhou, M. 2024. Linearly-evolved Transformer for Pan-sharpening. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1486–1494.

Hou, J.; Liu, X.; Wu, C.; Cong, X.; Huang, C.; Deng, L.-J.; and You, J. W. 2025. Bidomain uncertainty gated recursive network for pan-sharpening. *Information Fusion*, 118: 102938.

Hu, J.-F.; Huang, T.-Z.; Deng, L.-J.; Jiang, T.-X.; Vivone, G.; and Chanussot, J. 2021. Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12): 7251–7265.

Huang, J.; Chen, H.; Ren, J.; Peng, S.; and Deng, L. 2025. A General Adaptive Dual-level Weighting Mechanism for Remote Sensing Pansharpening. *arXiv preprint arXiv:2503.13214*.

Jin, Z.-R.; Zhang, T.-J.; Jiang, T.-X.; Vivone, G.; and Deng, L.-J. 2022. LAGConv: Local-Context Adaptive Convolution Kernels with Global Harmonic Bias for Pansharpening. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1113–1121. AAAI Press.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6007–6017.

Kim, S.; Do, J.; Lee, J.; and Kim, M. 2025. U-Know-DiffPAN: An Uncertainty-aware Knowledge Distillation Diffusion Framework with Details Enhancement for PAN-Sharpening. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 23069–23079.

Kovachki, N.; Li, Z.; Liu, B.; Azizzadenesheli, K.; Bhattacharya, K.; Stuart, A.; and Anandkumar, A. 2023. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89): 1–97.

Li, Z.; Kovachki, N. B.; Azizzadenesheli, K.; liu, B.; Bhattacharya, K.; Stuart, A.; and Anandkumar, A. 2021. Fourier Neural Operator for Parametric Partial Differential Equations. In *International Conference on Learning Representations*.

Liu, X.; and Tang, H. 2025. DiffFNO: Diffusion Fourier Neural Operator. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 150–160.

Liu, Y.; Dian, R.; and Li, S. 2025. Low-rank transformer for high-resolution hyperspectral computational imaging. *International Journal of Computer Vision*, 133(2): 809–824.

Lolli, S.; Alparone, L.; Garzelli, A.; and Vivone, G. 2017. Haze correction for contrast-based multispectral pansharpening. *IEEE Geoscience and Remote Sensing Letters*, 14(12): 2255–2259.

Lu, L.; Jin, P.; Pang, G.; Zhang, Z.; and Karniadakis, G. E. 2021. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature machine intelligence*, 3(3): 218–229.

Luo, X.; Qian, X.; and Yoon, B.-J. 2024. Hierarchical Neural Operator Transformer with Learnable Frequency-aware Loss Prior for Arbitrary-scale Super-resolution. In *Proceedings of the 41st International Conference on Machine Learn-*

ing, volume 235 of *Proceedings of Machine Learning Research*, 33466–33485. PMLR.

Meng, Q.; Shi, W.; Li, S.; and Zhang, L. 2023. PanDiff: A novel pansharpening method based on denoising diffusion probabilistic model. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–17.

Meng, X.; Shen, H.; Li, H.; Zhang, L.; and Fu, R. 2019. Review of the pansharpening methods for remote sensing images based on the idea of meta-analysis: Practical discussion and challenges. *Information Fusion*, 46: 102–113.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Shi, Y.; Xue, C.; Liew, J. H.; Pan, J.; Yan, H.; Zhang, W.; Tan, V. Y.; and Bai, S. 2024. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8839–8849.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Tan, J.; Huang, J.; Zheng, N.; Zhou, M.; Yan, K.; Hong, D.; and Zhao, F. 2024. Revisiting Spatial-Frequency Information Integration from a Hierarchical Perspective for Panchromatic and Multi-Spectral Image Fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25922–25931.

Tran, A.; Mathews, A.; Xie, L.; and Ong, C. S. 2023. Factorized Fourier Neural Operators. In *The Eleventh International Conference on Learning Representations*.

Vivone, G. 2019. Robust band-dependent spatial-detail approaches for panchromatic sharpening. *IEEE transactions on Geoscience and Remote Sensing*, 57(9): 6421–6433.

Vivone, G.; Deng, L.-J.; Deng, S.; Hong, D.; Jiang, M.; Li, C.; Li, W.; Shen, H.; Wu, X.; Xiao, J.-L.; et al. 2024. Deep Learning in Remote Sensing Image Fusion: Methods, protocols, data, and future perspectives. *IEEE Geoscience and Remote Sensing Magazine*.

Wald, L. 2002. *Data fusion: definitions and architectures: fusion of images of different spatial resolutions*. Presses des MINES.

Wei, M.; and Zhang, X. 2023. Super-resolution neural operator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18247–18256.

Wu, X.; Cao, Z.-H.; Huang, T.-Z.; Deng, L.-J.; Chanussot, J.; and Vivone, G. 2025. Fully-Connected Transformer for Multi-Source Image Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3): 2071–2088.

Wu, Z.-C.; Huang, T.-Z.; Deng, L.-J.; Huang, J.; Chanussot, J.; and Vivone, G. 2023. LRTCFPan: Low-rank tensor completion based framework for pansharpening. *IEEE Transactions on Image Processing*, 32: 1640–1655.

Xu, Z.; Tang, Y.; Xu, B.; and Li, Q. 2025. NeurOp-Diff: Continuous Remote Sensing Image Super-Resolution via Neural Operator Diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 12491–12501.

Yang, J.; Fu, X.; Hu, Y.; Huang, Y.; Ding, X.; and Paisley, J. 2017. PanNet: A deep network architecture for pansharpening. In *Proceedings of the IEEE international conference on computer vision*, 5449–5457.

Yuhas, R. H.; Goetz, A. F.; and Boardman, J. W. 1992. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm. In *JPL, Summaries of the Third Annual JPL Airborne Geoscience Workshop. Volume 1: AVIRIS Workshop*.

Zhou, J.; Civco, D. L.; and Silander, J. A. 1998. A wavelet transform method to merge Landsat TM and SPOT panchromatic data. *International journal of remote sensing*, 19(4): 743–757.

Zhou, M.; Huang, J.; Fang, Y.; Fu, X.; and Liu, A. 2022. Pan-sharpening with customized transformer and invertible neural network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 3553–3561.

Zhou, M.; Huang, J.; Guo, C.-L.; and Li, C. 2023. Fourmer: An efficient global modeling paradigm for image restoration. In *International conference on machine learning*, 42589–42601. PMLR.

Zhou, M.; Zheng, N.; He, X.; Hong, D.; and Chanussot, J. 2024. Probing synergistic high-order interaction for multimodal image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.