PanDiT: A Few-Step Diffusion Transformer for High-Fidelity and Efficient Pansharpening

Jiabin Fang, Ke Cao, Xuanhua He, Jie Zhang, Man Zhou, and Liu Liu

***Abstract*—Pansharpening plays a crucial role in remote sensing by fusing low-resolution multispectral (LRMS) images and high-resolution panchromatic (PAN) images to generate high-resolution multispectral (HRMS) images. While denoising diffusion models offer potential for high-fidelity image generation, their practical application in pansharpening has been severely hindered by huge computational costs from iterative sampling and naive conditioning strategies that struggle to fuse multimodal information effectively. In this article, we introduce PanDiT, a novel diffusion transformer (DiT) framework designed to address these challenges. PanDiT is built on the core principle of a decoupled conditioning mechanism, which explicitly disentangles and injects spatial and spectral guidance, and is engineered for practical, few-step inference. Our framework leverages a powerful DiT backbone, where conditioning is achieved through two specialized injection blocks that capture spatial and time–frequency features. Crucially, by integrating an implicit sampling strategy, we accelerate the inference process to as few as two steps. Extensive experiments on multiple benchmark datasets demonstrate that PanDiT not only establishes a new state-of-the-art in fusion quality and achieves a good quality-efficiency trade-off. Code is available at https://github.com/para133/PanDIT**

***Index Terms*—Deep learning, diffusion transformer (DiT), image fusion, pansharpening.**

I. Introduction

WITH the rapid advancement of remote sensing imaging technology, satellite remote sensing images have garnered increasing attention from enterprises and organizations, and have been widely applied across diverse fields such as mineral exploration [1], environmental monitoring [2], and military defense [3]. However, due to current technological limitations, it remains challenging to directly acquire high-resolution multispectral (HRMS) images. To address this, satellites are typically equipped with two types of sensors to simultaneously capture low-resolution multispectral (LRMS) images and high-resolution, texture-rich panchromatic (PAN) images. Pansharpening techniques are then employed to integrate these complementary sources of information, enabling the generation of HRMS images with enhanced spatial and spectral quality.

Over the past decades, various pansharpening techniques have been proposed, which can be broadly categorized into traditional physics-based approaches [4], [5] and deep learning-based methods. Traditional approaches, although physically interpretable, rely on manually designed feature extraction, limiting their ability to model complex scenes and nonlinear relationships, thus leading to issues like spectral distortion and spatial artifacts. In recent years, inspired by the powerful feature extraction capabilities of deep learning for images, deep learning-based pansharpening methods have gained significant attention. Early works such as PNN [6] demonstrated the superiority of neural networks over traditional methods. Subsequent models, such as PanNet [7], enhanced fusion by injecting high-frequency details from PAN into upsampled LRMS images, while MSDCNN [8] improved multiscale feature representation through multiscale convolution. More recently, INNformer [9] introduced vision transformers (ViTs) [10] into pansharpening, leveraging long-range dependency modeling to achieve impressive results. Concurrently, the remarkable success of denoising diffusion probabilistic models (DDPMs) [11] in high-fidelity image generation has inspired their application to pansharpening [12], [13], [14]. These methods frame the task as a conditional generation process, offering a powerful new pathway for detail restoration, surpassing other methods.

Despite these advancements, the application of diffusion models to pansharpening is constrained by two critical bottlenecks that prevent them from reaching their full potential.

The first is the computational cost inherent in the diffusion process. The high generative fidelity of these models is achieved through a slow, iterative refinement process, starting from pure Gaussian noise and gradually denoising it over hundreds or even thousands of sequential steps [15]. For typical high-resolution remote sensing images, which can be thousands of pixels wide, this inference requires minutes per image and massive GPU memory consumption. This computational overhead makes them largely impractical for real-world scenarios. The second lies in the naive conditioning strategies employed by existing diffusion-based approaches. These methods typically resort to simple channel-wise concatenation, treating the guiding PAN and LRMS images as additional input channels to a standard U-Net backbone. This

Received 3 July 2025; revised 14 October 2025; accepted 5 November 2025. Date of publication 10 November 2025; date of current version 19 November 2025. This work was supported by Anhui Provincial Natural Science Foundation under Grant 2408085MD090. *(Jiabin Fang and Ke Cao contributed equally to this work.) (Corresponding authors: Jie Zhang; Liu Liu.)*

Jiabin Fang and Liu Liu are with Hefei University of Technology, Hefei 230002, China (e-mail: 2004fjb@mail.hfut.edu.cn; liuliu@hfut.edu.cn).

Ke Cao, Xuanhua He, and Jie Zhang are with Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230088, China (e-mail: caoke200820@mail.ustc.edu.cn; hexuanhua@mail.ustc.edu.cn; zhangjie@iim.ac.cn).

Man Zhou is with the University of Science and Technology of China, Hefei 230026, China (e-mail: manman@mail.ustc.edu.cn).

Digital Object Identifier 10.1109/TGRS.2025.3630637

simple fusion approach fails to consider the distinct physical properties of each modality. It forces a single network architecture to implicitly untangle two different signals: a single-channel, high-resolution structural map and a multichannel, low-resolution spectral data. This often leads to inefficiency in information fusion. An effective model requires a tailored method to explicitly disentangle and fuse these information sources, rather than leaving the task to chance during the training process.

In this article, we argue that unlocking the power of diffusion models for pansharpening requires a paradigm that simultaneously addresses these two challenges. To this end, we propose PanDiT, a novel framework built upon two core principles. First, we introduce a decoupled conditioning mechanism designed to orthogonally inject guidance from different physical domains—spatial structure and spectral information—into the diffusion process. This allows the model to intelligently integrate complementary features while minimizing modal conflicts. Second, we integrate an accelerated denoising strategy based on implicit sampling, which enables practical few-step inference and makes our high-performance model usable.

Our framework is built around a powerful diffusion transformer (DiT) [16] backbone, which excels at generative modeling. To realize our decoupled conditioning principle, we design two specialized injection blocks. One captures fine-grained spatial features using multiscale convolutions, while the other extracts directional frequency information via wavelet transforms. These distinct feature streams are then intelligently fused using a cross-attention mechanism before being injected into the DiT. Critically, our adoption of implicit sampling reduces the number of required inference steps to as few as two, representing a great leap in efficiency.

Our contributions are summarized as follows.

1) We propose a novel framework that, for the first time, successfully adapts the DiT architecture to pansharpening.
2) We introduce a decoupled conditioning mechanism that orthogonally injects spatial and time–frequency features. This approach to information fusion allows for superior preservation of both spatial details and spectral fidelity compared to previous methods.
3) We achieve a noticeable improvement in efficiency by integrating implicit sampling, reducing the inference process to as few as two steps.
4) Through extensive experiments on multiple benchmark datasets, our proposed PanDiT establishes a new state-of-the-art in both qualitative and quantitative evaluations, and we provide a systematic analysis of key architectural choices to guide future research.

II. Related Work

A. Pansharpening

Traditional pansharpening methods primarily rely on prior-based, handcrafted designs. Among these, component substitution (CS) algorithms enhance the spatial details of LRMS images by substituting their spatial features with those from the PAN image. However, these methods often suffer from significant spectral distortion, as they fail to account for the inherent correlations between LRMS and PAN images. In contrast, variational optimization (VO) techniques extract spatial information from the PAN image using multiresolution decomposition and inject it into the LRMS image by solving an optimization problem defined through an objective function. Multiresolution analysis (MRA) methods decompose images at multiple scales to extract and fuse spatial and spectral features. While these approaches attempt to leverage the complementary feature between the two modalities, their ability to represent features remains limited.

The advent of deep learning has significantly advanced pansharpening by overcoming the representational limitations of traditional methods. Inspired by the success of SRCNN [17], PNN was the first to apply deep learning to pansharpening, utilizing a three-layer convolutional network to fuse PAN and upsampled LRMS images, leading to significant improvements in performance. Subsequent works have built upon and refined this framework. For instance, PanNet introduced residual learning to better restore high-frequency spatial details, while SFINet [18] applied the Fourier transform for high-frequency feature extraction and employed an invertible network for spatial feature fusion. Zhou et al. [19] proposed a mutual information-driven approach, optimizing the loss function to promote complementary feature extraction from both LRMS and PAN images.

The introduction of ViTs has further revolutionized pansharpening techniques. Panformer [20] leveraged self-attention mechanisms to effectively capture global spatial dependencies. GPPNN [21] incorporated multimodal prior knowledge to enhance fusion quality. More recently, diffusion models have been incorporated into pansharpening. Cao et al. [22] proposed DDRF, which employs a denoising autoencoder structure. By conditioning on the upsampled LRMS and PAN images, the model progressively refines HRMS images from noise, demonstrating the strong potential of diffusion-based methods in pansharpening tasks.

B. Diffusion Transformer

In recent years, diffusion models have made significant strides in generative modeling. Ho et al. [11] introduced the DDPM, which reconstructs high-quality images from random noise through a progressive denoising process, showcasing the powerful potential of diffusion models in image generation. Subsequently, score-based generative models [23] deepened the theoretical understanding of noise modeling and denoising dynamics, thereby propelling the field forward.

At the same time, the advent of the ViT introduced a new paradigm for image feature modeling. By employing self-attention mechanisms, ViT effectively overcomes the limitations of local receptive fields inherent in convolutional neural networks (CNNs), enabling the modeling of long-range dependencies across an image. However, ViT models are typically less robust to noise and degraded inputs compared to carefully designed CNN architectures.

To harness the strengths of both paradigms, the DiT architecture was proposed, integrating diffusion models with

Transformer-based self-attention mechanisms. Unlike traditional diffusion models, which use a U-Net backbone, DiTs employ the global modeling capabilities of self-attention to better fuse multiscale and hierarchical features, thus improving the quality of image generation. DiTs have achieved remarkable performance across various tasks, including text-to-image generation [24], image super-resolution [25], and image restoration [26]. Despite these successes, their application in image fusion tasks, such as pansharpening, remains a promising avenue for future exploration.

C. Denoising Diffusion Implicit Models

Denoising diffusion implicit models (DDIMs) represent a notable advancement in implicit sampling for diffusion models. Through a non-Markovian formulation, DDIM alters the traditional diffusion process, enabling the omission of intermediate denoising steps during sampling. This approach leverages a deterministic reverse process, which accelerates image generation by producing high-quality results with significantly fewer sampling steps than DDPM. As a result, DDIM reduces computational costs while maintaining image fidelity.

III. Proposed Method

In this section, we denote the PAN image as $\mathbf{P} \in \mathbb{R}^{H\times W\times 1}$. In the model input, the LRMS image is upsampled using bicubic interpolation to match the spatial resolution of the PAN image, yielding $\mathbf{M} \in \mathbb{R}^{H\times W\times C}$.

A. Preliminaries

The diffusion model consists of two primary processes: forward diffusion and reverse denoising. Both processes are parameterized as Markov chains. In the forward diffusion process, Gaussian noise is progressively added to the image, leading to a noisy version of the original image. In contrast, the reverse denoising process reconstructs the original image by progressively denoising the noisy image through a specialized denoising network.

The forward diffusion process generates a Gaussian-distributed noisy image $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ by sequentially injecting Gaussian noise into the original image $\mathbf{x}_0$. This process is mathematically expressed as

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

where β_t represents the variance of the Gaussian noise at the tth step. Defining $\alpha_t = 1-\beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$, the distribution from $\mathbf{x}_0$ to $\mathbf{x}_t$ is given by

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t;\ \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right). \tag{2}$$

The core of the reverse denoising process involves learning a conditional distribution, which can be expressed as

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}\left(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I}\right) \tag{3}$$

where θ represents the model parameters to be trained. In the DDPM framework, this is typically achieved by directly predicting the added noise ϵ. Consequently, the training objective becomes

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{\mathbf{x}_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|_1\right]. \tag{4}$$

During inference, the diffusion model begins with a random Gaussian noise sample $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ and iteratively denoises it to reconstruct the original image. The sampling step is described by the following equation:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right) + \sigma_t\mathbf{z} \tag{5}$$

where $\sigma_t^2 = \beta_t$ and $\mathbf{z}$ represents standard Gaussian noise.

For pansharpening tasks, the diffusion model can be extended to a conditional diffusion model by incorporating the LRMS and PAN images as conditioning inputs, denoted by $\mathbf{c}$. In this setting, the model ϵ_θ takes the conditioning information into account, as formulated by

$$\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c}). \tag{6}$$

The forward diffusion process remains unchanged, with the HRMS image used as the original image $\mathbf{x}_0$, to which Gaussian noise is progressively added. However, during the reverse denoising process, the conditioning input $\mathbf{c}$ guides the generation. Consequently, the training objective is reformulated as a conditional loss function to incorporate this additional information.

B. PanDiT Framework

1) Overview: The overall architecture of our model, as shown in Fig. 1, consists of three primary components: the DiT Block, the feature extraction block, and the modality fusion block (MFB). The feature extraction block is further divided into the spatial extraction block (SEB), which extracts spatial features, and the time–frequency extraction block (TFEB), which captures time–frequency domain features.

Our model adopts a residual learning approach. Instead of directly predicting the HRMS image, we predict the residual difference between the HRMS image and the LRMS image. The final HRMS image is obtained by adding the predicted residual to the LRMS image. In line with the diffusion model framework, we initially generate random noise sampled from a Gaussian distribution, denoted as $\mathbf{N} \in \mathbb{R}^{H\times W\times C}$. To preserve image details during the reconstruction process, we refrain from using a variational autoencoder. Additionally, to enable comprehensive feature fusion, a preprocessing convolutional layer is applied to upsample the number of image channels, resulting in the noise representation $\mathbf{N_h} \in \mathbb{R}^{H\times W\times C_h}$, where C_h denotes the intermediate hidden channel dimension.

The noise $\mathbf{N_h}$ is then processed through a series of DiT Blocks, where conditional information is progressively injected to restore the feature maps corresponding to the residuals between the HRMS and LRMS images. The first half of the DiT Blocks utilizes the SEB for spatial feature extraction, while the latter half incorporates the TFEB to capture time–frequency domain features. In the final layer, the channel dimensions are restored to match the original multispectral image, and the LRMS image is added to the predicted residual to produce the final HRMS output.

Eventually, the overall process can be summarized as follows:

$$\text{HRMS} = \theta(\varphi_i \ldots (\varphi_1(\phi(N)))) + M. \tag{7}$$

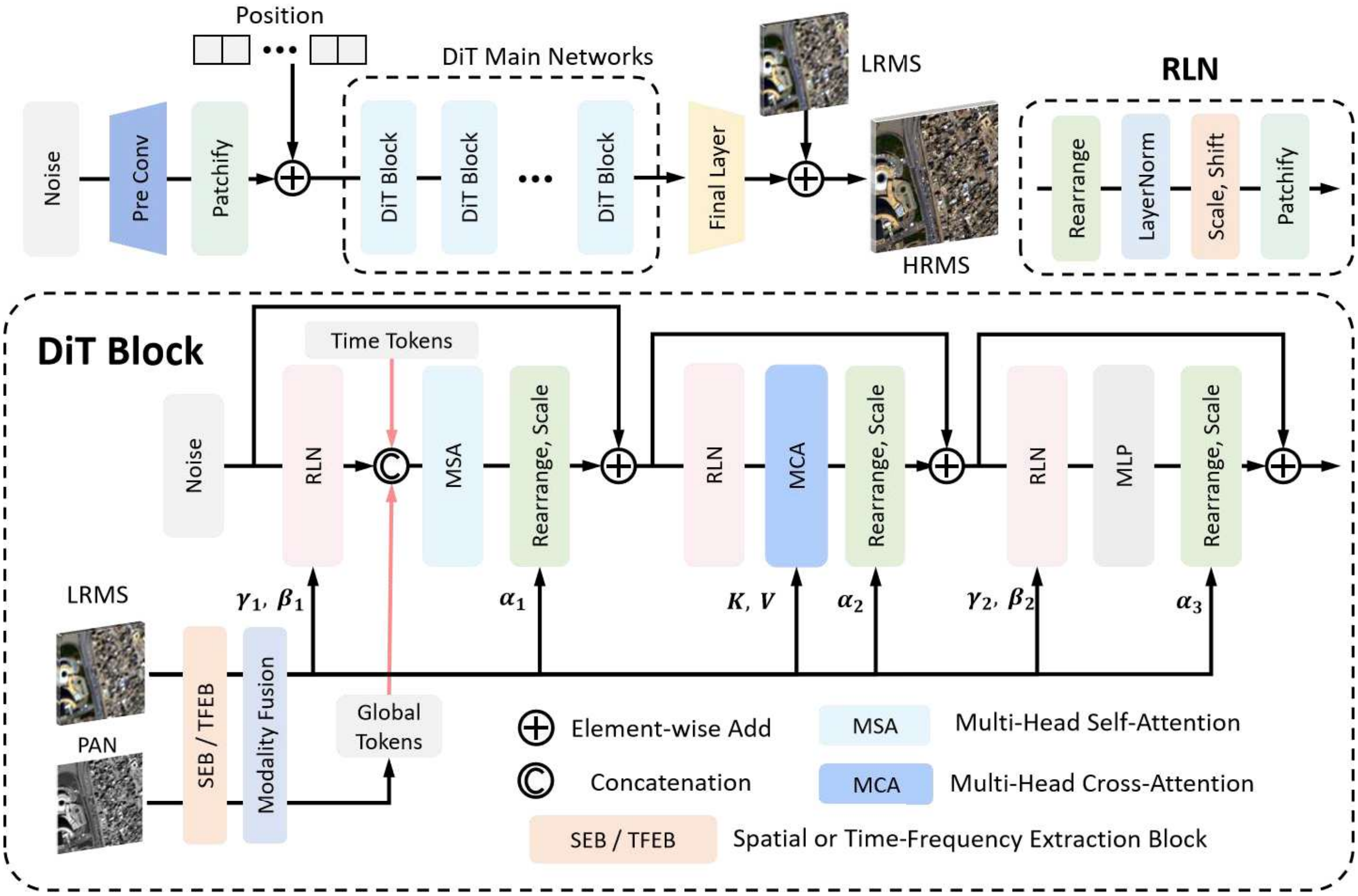

Fig. 1. Overall architecture of PanDiT is composed of three main components: the DiT Block, the feature extraction block, and the MFB. The upsampled noise is passed through multiple DiT blocks and subsequently downsampled through the final layer, generating the residual difference between HRMS and LRMS images.

In this equation, $\varphi_i(.)$ denotes the ith DiT Block, $\theta(.)$ represents the final downsampling output layer, and $\phi(.)$ refers to the preconvolution layer employed in the network.

2) DiT Block Design: In this work, we adopt the adaLN-Zero strategy, following the approach outlined by Peebles and Xie [16]. However, unlike the original implementation, we apply LayerNorm to $\mathbf{N_h}$ after the rearrange operation rather than to the individual tokens. This modification ensures better preservation of the relative relationships among the parameters in the feature maps. The scale and shift parameters in adaLN are generated by the SEB or TFEB, and their dimensions are aligned with those of $\mathbf{N_h}$.

Within the DiT Block, unlike the conventional Transformer encoder, we introduce a Rearrange operation before each LayerNorm to reconstruct the token sequence back into its original spatial configuration $I \in \mathbb{R}^{H \times W \times C}$. This operation restores the 2-D spatial layout of the image, allowing the model to maintain the relative spatial relationships among tokens throughout the encoding process. By preserving spatial consistency, the model can better capture local structural dependencies and reduce the loss of spatial contextual information that typically occurs during sequential token processing in standard Transformers.

We integrate multihead cross-attention to facilitate effective feature fusion. To maintain consistent feature flow, the key ($\mathbf{K}$) and value ($\mathbf{V}$) are derived from the feature maps produced by the SEB or TFEB. These feature maps are first upsampled using a convolutional layer, then split equally. The query ($\mathbf{Q}'$) is generated from $\mathbf{N_h}$ within the multihead cross-attention using a linear transformation. Before the multihead cross-attention, the scale and shift operations are applied within the RLN Block. The feature maps of $\mathbf{Q}'$, $\mathbf{K}$, and $\mathbf{V}$ are then split into $\mathbf{h_{num}}$ heads, denoted as $\mathbf{Q}'_\mathbf{i}$, $\mathbf{K_i}$, and $\mathbf{V_i}$, with each head computing attention values independently. The results are subsequently concatenated. Eventually, the multihead attention (MHA) operation MHA is defined as

$$\text{Attention}_i\left(Q'_i, K_i, V_i\right) = \text{softmax}\left(\frac{Q'_i {K_i}^T}{\sqrt{d_k}}\right) \tag{8}$$

$$\text{MHA}\left(Q', K, V\right) = \text{Concat}\left(\text{Attention}_i\right)_{i=1}^{h_{\text{num}}}. \tag{9}$$

Before applying multihead self-attention within the DiT Block, we concatenate time tokens and global tokens with the token sequence generated by the patchify operation. The time token is generated by feeding the time step, produced by a cosine scheduler, into a multilayer perceptron (MLP). The global token is generated by applying the patchify operation to the feature map obtained from the MFB. This concatenated sequence is then processed through an MLP and an average pooling layer. The global token provides the model with the

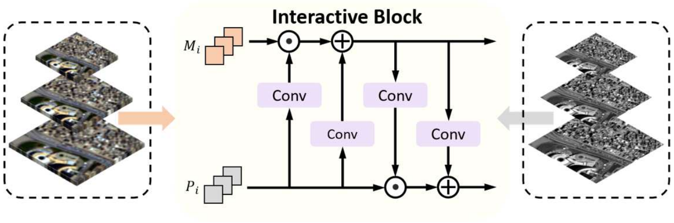

Fig. 2. Schematic of the proposed method for extracting spatial features. A spatial pyramid is utilized to extract multiscale spatial features from both LRMS and PAN images.

relative relationships between each token in the sequence and the entire image, enabling the self-attention mechanism to incorporate global image features. After the multihead self-attention, both the time and global tokens are discarded, and only the remaining tokens continue through the DiT Block.

In the final layer, only the last RLN and MLP components of the DiT Block are retained. The output dimension of the final MLP layer is adjusted to match the number of channels in the multispectral image, ensuring that each output token $\mathbf{token_O} \in \mathbb{R}^{H\times W\times C}$ corresponds to the final prediction. Subsequently, the predicted residuals between the HRMS and LRMS images are obtained through the rearrangement operation.

3) Decoupled Conditioning Mechanism: In prior conditional diffusion models, the incorporation of conditional features was typically achieved through a straightforward concatenation along the channel dimension. However, the LRMS and PAN images in pansharpening contain fundamentally different types of information, and as such, direct concatenation may not effectively capture these distinctions. Specifically, LRMS images are primarily characterized by spectral information, while PAN images offer high-resolution spatial details. To address this, our proposed PanDiT model adopts a more nuanced approach. We introduce a spatial pyramid structure that extracts spatial features from both LRMS and PAN images across multiple scales. Additionally, we apply wavelet transforms to extract principal component features from the LRMS images and directional high-frequency features from the PAN images. To fully leverage these diverse features, we implement a fusion strategy that considers both the discrepant and common features between the two modalities. This strategy ensures that the complementary features from both image types are maximally utilized, thereby enhancing the model's performance.

a) Spatial feature extraction: The SEB, illustrated in Fig. 2, is designed to extract spatial features by leveraging the local aggregation capability of convolutional operations. It captures regional details through localized receptive fields and constructs multiscale feature representations via progressive downsampling.

Specifically, the input images are first processed by a convolutional layer to obtain initial feature maps $\mathbf{M_1}, \mathbf{P_1} \in \mathbb{R}^{H\times W\times C_h}$. Subsequent downsampling operations halve the spatial resolution and double the channel dimension at each stage, resulting in feature maps $\mathbf{M_2}, \mathbf{P_2} \in \mathbb{R}^{H/2\times W/2\times 2C_h}$ and $\mathbf{M_3}, \mathbf{P_3} \in \mathbb{R}^{H/4\times W/4\times 4C_h}$.

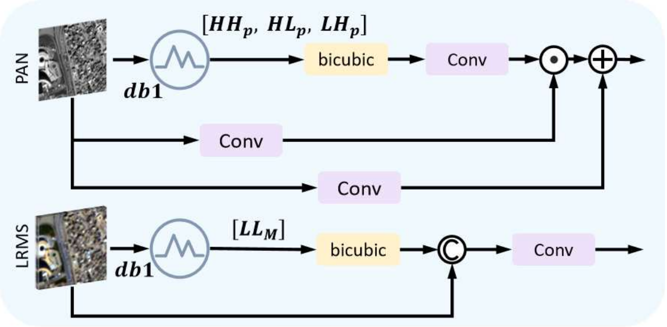

Fig. 3. Proposed method for time–frequency domain feature extraction. The wavelet transform is employed to extract low-frequency components from LRMS and high-frequency components from PAN, enabling complementary feature representation in the time–frequency domain.

To facilitate the fusion of features from LRMS and PAN images, we apply an element-wise affine transformation at each scale. The computation is described as follows:

$$\text{Scale}_M, \text{Shift}_M = \text{Conv}_{P_1}(P_i), \text{Conv}_{P_2}(P_i) \tag{10}$$

$$M_i = M_i \odot (1 + \text{Scale}_M) + \text{Shift}_M \tag{11}$$

$$\text{Scale}_P, \text{Shift}_P = \text{Conv}_{M_1}(M_i), \text{Conv}_{M_2}(M_i) \tag{12}$$

$$P_i = P_i \odot (1 + \text{Scale}_P) + \text{Shift}_P. \tag{13}$$

Here, $\odot$ represents element-wise multiplication. Finally, transposed convolutions are applied to upsample the lower resolution feature maps, restoring them to their original spatial size.

b) Time–frequency feature extraction: We adopt a simple yet effective strategy to capture complementary features from different modalities: high-frequency features are extracted from the PAN image, which contains rich structural details, while low-frequency features are derived from the LRMS image, which primarily encodes the dominant spectral content of the target HRMS image. The overall structure of the TFEB is illustrated in Fig. 3.

Specifically, we apply single-level wavelet decomposition to the PAN image to obtain its diagonal (HH), vertical (HL), and horizontal (LH) detail subbands. Similarly, wavelet decomposition is applied to the LRMS image; however, only the low-frequency approximation component (LL) is retained, given the LRMS image's predominantly low-frequency nature. The decomposition process is defined as

$$[(\text{HH}_P, \text{HL}_P, \text{LH}_P), \text{LL}_P] = \text{db1}(P) \tag{14}$$

$$[\text{HH}_M, \text{HL}_M, \text{LH}_M, (\text{LL}_M)] = \text{db1}(M) \tag{15}$$

where the components enclosed in parentheses denote the sub-bands that are preserved for subsequent processing. Each extracted subband of image I has a spatial size of $I_w \in \mathbb{R}^{H/2\times W/2\times C_I}$. To ensure spatial alignment, all components are upsampled to the original image resolution via bicubic interpolation.

The high-frequency components of PAN are concatenated along the channel dimension and passed through a convolutional layer, followed by an element-wise affine transformation with features extracted directly from the original PAN image. This operation embeds rich spatial detail into the final feature

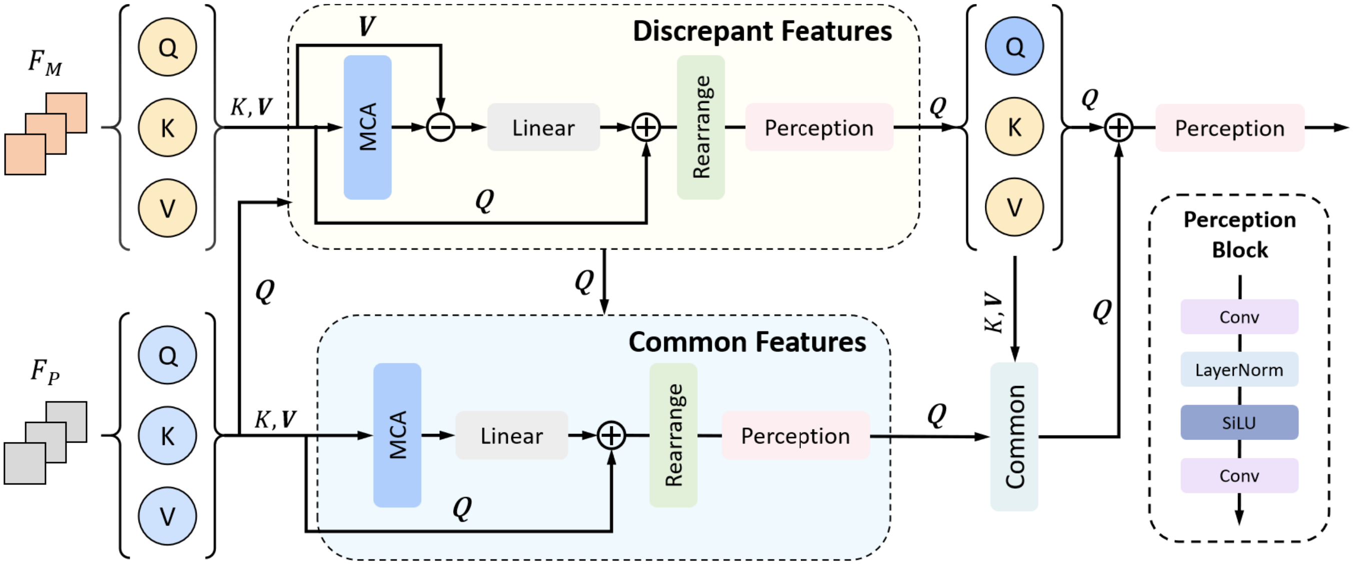

Fig. 4. Proposed modality fusion strategy. Both discrepant and common features are extracted from LRMS and PAN images to ensure the final output preserves the low-frequency spectral content of LRMS and the high-frequency spatial details of PAN.

representation. As for the LRMS input, which predominantly conveys low-frequency information, we concatenate the LL component with the original LRMS image and feed the result into a convolutional layer. This enables the model to learn a more expressive low-frequency representation with minimal redundancy.

c) Modality fusion: The complete condition injection block is responsible for extracting spatial or time–frequency domain features from the inputs and fusing these multimodal features. To achieve effective modality fusion, we draw inspiration from the ATFuse block [27], with appropriate modifications tailored to the pansharpening setting. We hypothesize that superior fusion quality can be attained by explicitly modeling both the discrepancies and the commonalities between LRMS and PAN inputs. The architecture of the MFB is depicted in Fig. 4.

We begin by splitting the extracted feature maps $\mathbf{F_M}$ and $\mathbf{F_P}$ from LRMS and PAN into three equal segments, corresponding to the query, key, and value for each modality. Following the MHA mechanism, each set of features is further divided into $\mathbf{h_{num}}$ attention heads. To extract modality-specific differences, we employ a cross-attention strategy where the query originates from the PAN branch while the key and value come from LRMS. The attention mechanism computes the similarity between $\mathbf{Q_P}$ and $\mathbf{K_M}$ using dot-product attention, and the resulting weights are applied to $\mathbf{V_M}$ to estimate the common features between the two modalities. The discrepant feature between PAN and LRMS is then obtained by subtracting the common representation from the original PAN features, followed by a linear projection to adjust dimensionality. This can be formally expressed as

$$D_{MP} = \text{Linear}\left(V_M - \text{MHA}\left(Q_P, K_M, V_M\right)\right) \tag{16}$$

where D_{MP} represents the discrepant feature. To further leverage the unique high-frequency information from the PAN image, this discrepant feature is reintegrated into the PAN representation. After a rearrangement operation, we employ a convolutional perception block to refine and enhance this output, yielding the enhanced feature map $\mathbf{F_D}$.

Although the discrepant feature derived via cross-attention captures detailed spatial structures, it is not sufficient for reconstructing complete HRMS content. Spectral fidelity and shared semantic structure must also be preserved. To this end, we introduce two additional common feature extraction modules: one conditioned on the PAN features and the other on the LRMS features. In both cases, the Query is derived from the original features, while the key and value originate from the alternate modality. The resulting common feature is denoted as $\mathbf{F_C}$. Finally, the enhanced output is obtained by combining the discrepancy and common features as follows:

$$F_D = \text{Perception}\left(\text{Rearrange}\left(Q_p + D_{MP}\right)\right) \tag{17}$$

$$F_{C_1} = \text{Common}\left(F_D, K_M, V_M\right) \tag{18}$$

$$F_{C_2} = \text{Common}\left(F_{C_1}, K_P, V_P\right) \tag{19}$$

$$F_O = F_D + F_{C_2}. \tag{20}$$

C. Denoising Diffusion Implicit Models

In DDPM, as described in (1), the forward diffusion process progressively adds Gaussian noise to the data through a Markov chain. In contrast, in PanDiT, the residual between the target HRMS and LRMS images is deterministic rather than stochastic. Hence, we adopt the deterministic variant of the diffusion model (DDIM) by setting $\sigma_t = 0$, which removes the random noise component in the reverse process.

Under this setting, the sampling trajectory becomes fully determined by the model's prediction $\boldsymbol{\epsilon}_t$ and the initial latent $\mathbf{x}_T$. The reverse process can be compactly expressed as

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{\mathbf{x}}_0 + \sqrt{1-\bar{\alpha}_{t-1}} \cdot \frac{\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\hat{\mathbf{x}}_0}{\sqrt{1-\bar{\alpha}_t}} \tag{21}$$

where $\hat{\mathbf{x}}_0 = \mathbf{x}_\theta(\mathbf{x}_t, t, \mathbf{c})$ denotes the model's predicted clean image at step $t = 0$.

Given the noisy input $\mathbf{x}_t$, time step t, and conditioning information $\mathbf{c}$. To encourage faithful HRMS reconstruction, we employ an ℓ_1 reconstruction objective

$$\mathcal{L}_{\text{recon}} = \mathbb{E}\left[\left|\mathbf{x}_0 - \mathbf{x}_\theta\left(\mathbf{x}_t, t, \mathbf{c}\right)\right|_1\right] \tag{22}$$

which minimizes the distance between the ground truth and the model prediction, thereby promoting high-fidelity reconstruction in a fully deterministic denoising process.

In practical applications, we set the total number of diffusion steps to $\mathbf{T}$. To accelerate the inference process, a subsequence $[\tau_1, \tau_2, \ldots, \tau_s]$ is sampled from the original step sequence $[1, 2, \ldots, \mathbf{T}]$. The corresponding forward diffusion process for each sampled step remains a Markov chain, defined as

$$\mathbf{q}\left(\mathbf{x}_{\tau_\mathrm{i}}|\mathbf{x_0}\right) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\alpha_{\tau_i}}\mathbf{x_0}, \left(1 - \alpha_{\tau_\mathrm{i}}\right)\mathbf{I}\right). \tag{23}$$

Accordingly, the reverse process can be constructed as a Markov chain over the selected steps, enabling a significantly reduced number of iterations for sampling

$$\mathbf{x}_{\tau-1} = \sqrt{\bar{\alpha}_{\tau-1}}\hat{\mathbf{x}}_0 + \sqrt{1 - \bar{\alpha}_{\tau-1}} \cdot \frac{\mathbf{x}_\tau - \sqrt{\bar{\alpha}_\tau}\hat{\mathbf{x}}_0}{\sqrt{1 - \bar{\alpha}_\tau}}. \tag{24}$$

Conventional diffusion models typically adopt a noise prediction paradigm, which has proven effective for generating large-scale, high-resolution images. However, in the pansharpening scenario, the target outputs are relatively small, and the model benefits from rich conditional information provided by the LRMS and PAN images. To better exploit these informative priors, our approach introduces a residual learning strategy that shifts the prediction target from the entire HRMS image to its residual with respect to the LRMS image. This design substantially reduces the learning complexity, as the model only needs to reconstruct fine-grained differences rather than synthesizing the full image from random noise.

Theoretically, this residual formulation transforms the diffusion trajectory into a low-variance mapping, where each reverse step predicts a deterministic correction to the conditional inputs. Under this setting, the signal-to-noise ratio (SNR) of the intermediate states remains high throughout the process, making it possible to recover high-fidelity details with only a few denoising iterations. Moreover, the deterministic reverse process ensures that the predicted residuals are consistent across diffusion steps, effectively eliminating stochastic perturbations and further stabilizing reconstruction. Built upon a ViT backbone with strong representation capability, the proposed model thus achieves accurate and consistent residual estimation in each step. As a result, high-quality pansharpened images can be generated with only a small number of sampling steps, maintaining both fidelity and efficiency during inference.

IV. Experiment

A. Datasets and Benchmark

Our experiments were conducted on three datasets: WorldView-II (WV2), Gaofen-2 (GF2), and WorldView-III (WV3), which encompass a diverse range of natural and urban scenes. Since ground-truth data is not available, we follow the Wald protocol [28] to generate the datasets. For training, we extract LRMS patches of size 32 × 32 and PAN patches of size 128 × 128 from the images. We compare the proposed method against both classical traditional approaches, including GFPCA [29], GS, Brovey [4], IHS [30], and SFIM [5], as well as state-of-the-art deep learning methods, such as PanNet [7], MSDCNN [8], SRPPNN [31], INNformer [9], SFINet [18], MSDDN [32], PanFlowNet [33], Pan-Mamba [34], CrossDiff [35], and DDRF [22]. The evaluation metrics used in our experiments include PSNR, SSIM, SAM [36], and ERGAS, alongside no-reference metrics such as D_S, D_λ, and QNR.

B. Implement Details

We trained our model using the PyTorch framework on an RTX 4090 GPU. The training process employed the AdamW optimizer with an initial learning rate of 1×10^{-4}. The model consists of 12 DiT blocks, with a patch size of 8 and a token size of 1024. The exponential moving average (EMA) ratio was set to 0.995. A total of 20 000 iterations were performed, with the learning rate reduced to 2×10^{-5} after 10 000 iterations. During training, the total number of diffusion steps was set to 500, while for prediction, only two sampling steps were performed.

C. Comparison With State of the Art Methods

1) Evaluation on Reduced-Resolution Scene: The evaluation results on the three datasets are presented in Table I. We benchmarked our proposed method against state-of-the-art pansharpening approaches, and the results demonstrate that our model achieves significant improvements, outperforming all other methods across all evaluation metrics. Notably, on the WV2 and GF2 datasets, our method achieves a substantial PSNR improvement, outperforming the second-best method by 0.35 and 0.30 dB, indicating superior consistency with the ground truth. Additional metrics further validate the effectiveness of our approach: higher SSIM indicates better preservation of contrast and structural details, while lower SAM and ERGAS suggest more accurate spectral fusion and reduced overall error, resulting in clearer textures.

The quantitative results of the model are illustrated in Figs. 5 and 6. Representative samples were selected for visualization, with the last row of each image displaying the mean squared error (mse) distribution between the fusion and the ground truth. In these visualizations, brighter areas indicate higher discrepancies. Our method consistently showcases the lowest error across all datasets, strongly affirming that its results are the closest to the ground truth.

2) Evaluation on Full-Resolution Scene: To assess the generalization capability of our model, we evaluated it on the full-resolution WV2 dataset without downsampling. Given the absence of ground-truth data, we evaluated our model's performance using no-reference quality metrics, including D_S, D_λ, and QNR. As shown in Table II, our method outperforms all other state-of-the-art approaches across all evaluation metrics, showcasing its strong potential for practical applications. Visual results of different methods on the full-resolution dataset are provided in Fig. 7. These results demonstrate that our method effectively preserves fine spatial details while maintaining high spectral fidelity.

TABLE I

Quantitative Comparison on Three Datasets. Best Results Highlighted in Red. ↑ Indicates That the Larger the Value, the Better the Performance, and ↓ Indicates That the Smaller the Value, the Better the Performance

Method	WorldView-II				GaoFen2				Worldview-III			
	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
SFIM	34.1297	0.8975	0.0439	2.3449	36.9060	0.8882	0.0318	1.7398	21.8212	0.5457	0.1208	8.9730
Brovey	35.8646	0.9216	0.0403	1.8238	37.7974	0.9026	0.0218	1.3720	22.5060	0.5466	0.1159	8.2331
IHS	35.2962	0.9027	0.0461	2.0278	38.1754	0.9100	0.0243	1.5336	22.5579	0.5354	0.1266	8.3616
GS	35.6376	0.9176	0.0423	1.8774	37.2260	0.9034	0.0309	1.6736	22.5608	0.5470	0.1217	8.2433
GFPCA	34.5581	0.9038	0.0488	2.1411	37.9443	0.9204	0.0314	1.5604	22.3344	0.4826	0.1294	8.3964
PANNET	40.8176	0.9626	0.0257	1.0557	43.0659	0.9685	0.0178	0.8577	29.6840	0.9072	0.0851	3.4263
MSDCNN	41.3355	0.9664	0.0242	0.9940	45.6847	0.9827	0.0135	0.6389	30.3038	0.9184	0.0782	3.1884
SRPPNN	41.4538	0.9679	0.0233	0.9899	47.1998	0.9877	0.0106	0.5586	30.4346	0.9202	0.0770	3.1553
INNformer	41.6903	0.9704	0.0227	0.9514	47.3528	0.9893	0.0102	0.5479	30.5365	0.9225	0.0747	3.0997
SFINet	41.7244	0.9725	0.0220	0.9506	47.4712	0.9901	0.0102	0.5462	30.5971	0.9236	0.0741	3.0798
MSDDN	41.8435	0.9711	0.0222	0.9478	47.4101	0.9895	0.0101	0.5414	30.8645	0.9258	0.0757	2.9581
PanFlowNet	41.8548	0.9712	0.0224	0.9335	47.2533	0.9884	0.0103	0.5512	30.4873	0.9221	0.0751	2.9531
Pan-Mamba	42.2354	0.9729	0.0212	0.8975	47.6453	0.9894	0.0103	0.5286	31.1551	0.9299	0.0702	2.8942
CrossDiff	40.4803	0.9604	0.0258	1.0823	46.4624	0.9854	0.0115	0.6027	29.4448	0.8975	0.0856	3.5246
DDRF	42.2880	0.9732	0.0210	0.8880	47.2665	0.9884	0.0104	0.5511	30.6744	0.9216	0.0759	3.0688
Ours	42.6385	0.9748	0.0202	0.8564	47.9487	0.9902	0.0095	0.5110	31.2684	0.9319	0.0732	2.8620

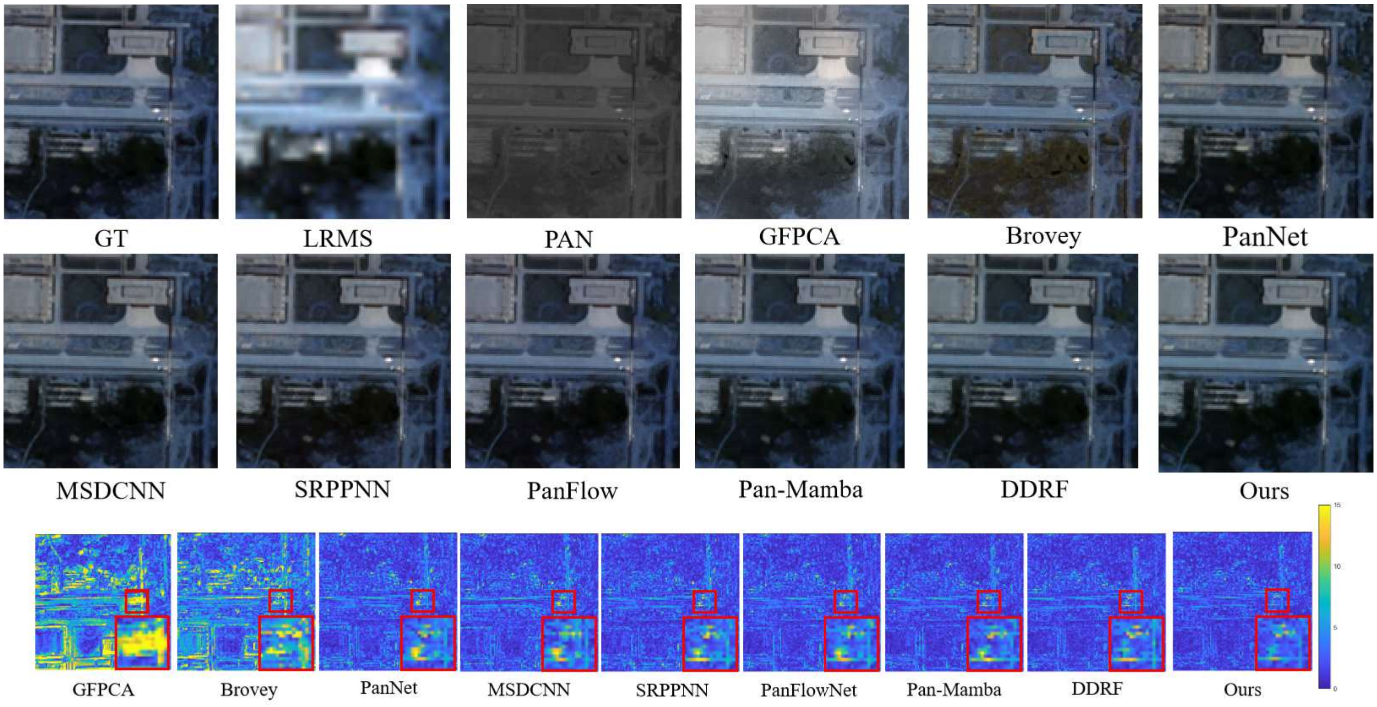

Fig. 5. Result of our approach was compared against nine other methods on the WorldView-II dataset.

D. *Comparison of Model Configurations*

We explored various configurations of the PanDiT and examined the scaling properties of our model. The model are named based on their patch size, token size, and the number of layers. Specifically, "D" and "S" denote models with 12 and 8 DiT blocks, respectively, while "L" and "M" refer to

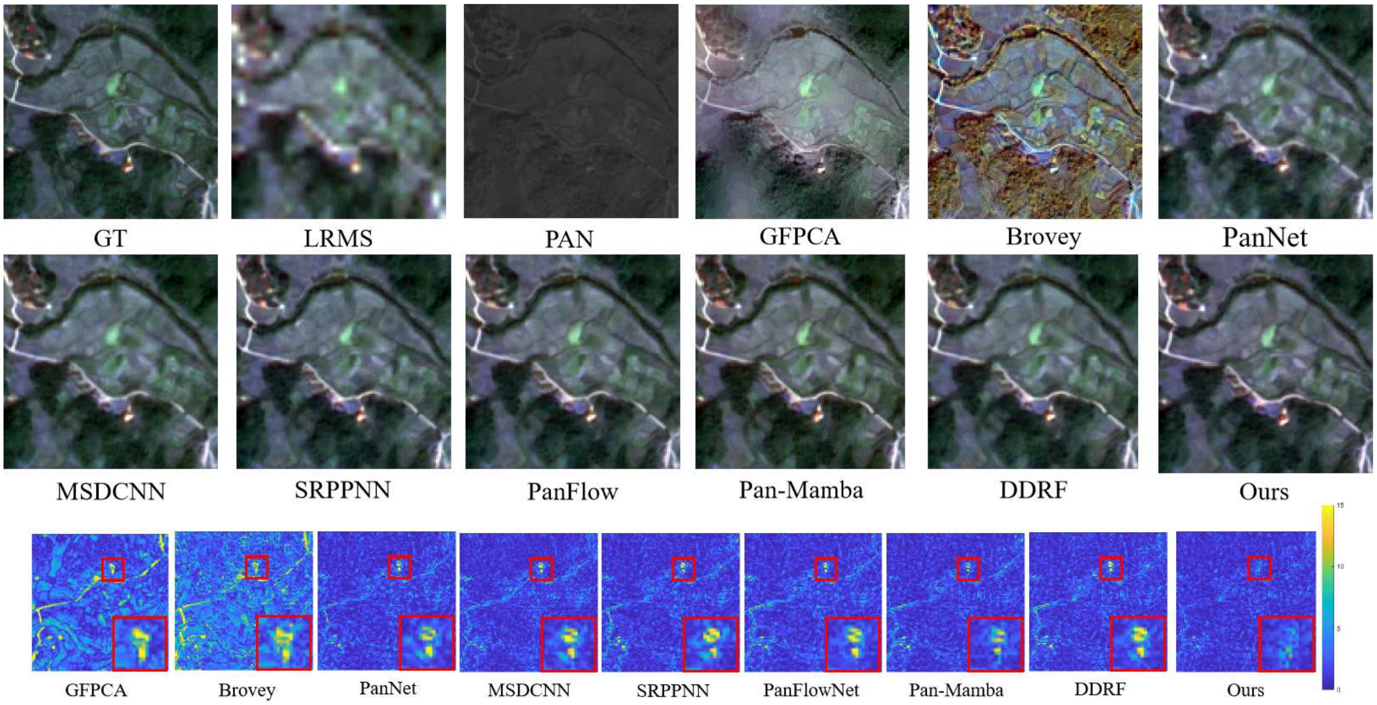

Fig. 6. Result of our approach was compared against nine other methods on the Gaofen-2 dataset.

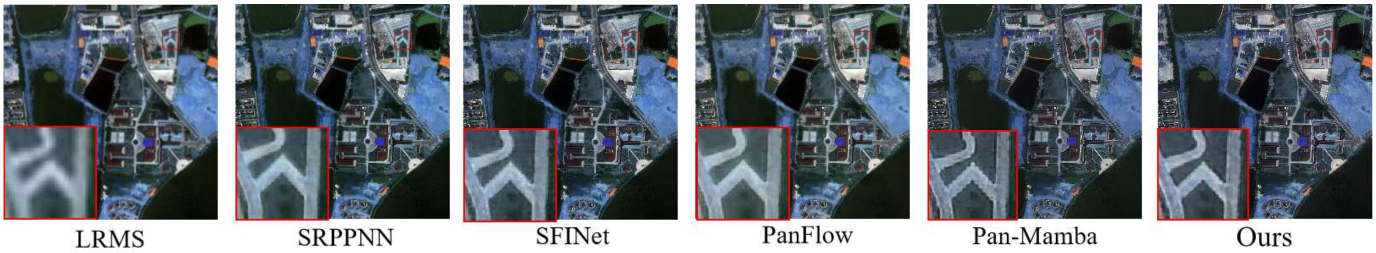

Fig. 7. Result of our approach was compared against four other methods on the full-resolution WV2 dataset.

TABLE II

Evaluation of the Proposed Method on Real-World Full-Resolution Scenes From the WorldView-II Dataset. Best Results Highlighted in Red. ↑ Indicates That the Larger the Value, the Better the Performance, and ↓ Indicates That the Smaller the Value, the Better the Performance

Metric	SFIM	Brovey	IHS	GFPCA	MSDCNN	SRPPNN	INNformer	SFINet	PanFlowNet	Pan-Mamba	Ours
$D_\lambda \downarrow$	0.1403	0.1026	0.1110	0.1139	0.1063	0.0998	0.0995	0.1034	0.0966	0.0966	0.0917
$D_S \downarrow$	0.1320	0.1409	0.1556	0.1535	0.1443	0.1637	0.1305	0.1305	0.1274	0.1272	0.1222
QNR ↑	0.7826	0.7728	0.7527	0.7532	0.7683	0.7548	0.7858	0.7827	0.7910	0.7911	0.7997

models with token sizes of 1024 and 512, respectively. For instance, PanDiT-8/L/D corresponds to a PanDiT model with a patch size of 8, a token size of 1024, and 12 DiT blocks.

We trained and evaluated these different model configurations on the WV2 dataset, calculating performance metrics of fusion results, as well as the FLOPs and parameter counts. The result is shown in Table III.. Our findings indicate that reducing the patch size, which results in longer token sequences, significantly enhances model performance. This improvement can be attributed to the model's enhanced capacity to capture fine-grained features, which facilitates better recovery of high-frequency details. Additionally, increasing the token size and model depth further improves performance: larger tokens provide richer representations for each patch, while deeper networks enable more effective feature extraction.

The relationship between PanDiT and other methods in terms of parameter count and performance is illustrated in the Fig. 8. Although our model involves a slight increase in parameters, this increase results in a substantial improvement in performance metrics. Specifically, PanDiT achieves higher

TABLE III
COMPARISON OF DIFFERENT MODEL CONFIGURATIONS ON WORLDVIEW-II

Method	WorldView-II				Param(M)	Flops(G)
	PSNR↑	SSIM↑	SAM↓	ERGAS↓		
PanDiT-16/M/S	40.9872	0.9641	0.0244	1.0209	0.2178	0.9860
PanDiT-16/M/D	41.2656	0.9658	0.0239	0.9947	0.3241	1.4529
PanDiT-16/L/S	41.7145	0.9690	0.0227	0.9471	0.8419	3.5049
PanDiT-16/L/D	41.8430	0.9698	0.0223	0.9337	1.2550	5.1947
PanDiT-8/M/S	42.3091	0.9728	0.0210	0.8881	0.2588	3.2696
PanDiT-8/M/D	42.3500	0.9731	0.0209	0.8832	0.3818	4.8624
PanDiT-8/L/S	42.4778	0.9737	0.0206	0.8724	0.9239	12.6345
PanDiT-8/L/D	42.6385	0.9748	0.0202	0.8564	1.3703	18.8263

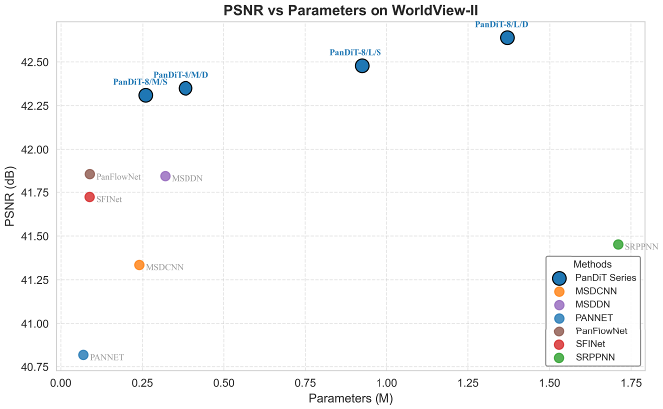

Fig. 8. Comparison of Parameter Count and PSNR Metrics between PanDiT and Other Methods on WorldView-II.

PSNR values compared to models with similar or even larger parameter counts, indicating that the additional parameters effectively enhance the model's ability to capture fine-grained image details. This also demonstrates that PanDiT's design enables a more efficient utilization of parameters, striking a favorable balance between model size and reconstruction quality.

It is important to note that the patch size directly influences the sequence length, with smaller patch sizes resulting in longer token sequences and consequently increasing computational complexity. On the other hand, the token size has a critical impact on the total number of model parameters. While longer sequences are introduced with smaller patch sizes, the parameter count remains relatively stable as long as the token size is fixed. Therefore, the model's configuration requires careful consideration to balance computational efficiency and performance. Notably, even lightweight models can deliver strong performance when appropriately configured.

E. Comparison With Other Diffusion Methods

To provide a comprehensive comparison, we briefly introduce the diffusion-based methods used in our experiments, including CrossDiff and DDRF, which represent two representative paradigms in pansharpening. CrossDiff leverages a cross-predictive diffusion process. In the pretraining stage, a cross-predictive task is constructed where the model learns to reconstruct PAN images from MS images and vice versa using DDPM. In the adaptation stage, the pretrained encoders are frozen to extract image features, and only a fusion head is trained. DDRF employs a conditional U-Net backbone and utilizes two conditional injection modulation modules to inject both global and fine-grained high-frequency features. Those design achieves high-quality fusion results but require relatively large iterative sampling steps, leading to high computational cost during inference.

We conducted a comparative analysis on the WV2 dataset. The test results include the time overhead required for model prediction on 50 images with varying numbers of sampling steps, with predictions made on a laptop equipped with an NVIDIA 3060 GPU. The results are presented in Table IV. Our experimental findings demonstrate that, in the context of pansharpening, the implicit denoising sampling method outperforms the explicit method. Compared to DDRF, our model maintains strong performance even when the number of steps for implicit denoising sampling is reduced to as low as two steps, with no obvious degradation in the evaluation metrics. Therefore, the conclusion can be summarized as follows: slightly increasing the model's computational cost to enhance single-step prediction performance, the implicit denoising sampling method can be more robust and less dependent on the number of steps, which results in improved prediction efficiency while preserving image quality and reducing overall time costs.

Our proposed PanDiT introduces a DiT backbone with a decoupled conditioning mechanism, which explicitly disentangles spatial and spectral features through dedicated extraction blocks. Moreover, by integrating implicit sampling, PanDiT dramatically reduces the number of inference steps to only a few (as few as two), achieving a superior trade-off between efficiency and fidelity. In contrast to DDRF's heavy iterative denoising and CrossDiff's pretext-based learning, PanDiT is designed for direct, efficient, and high-fidelity fusion under a unified framework.

TABLE IV
COMPARISON OF SAMPLING STRATEGIES IN DIFFUSION MODELS ON WORLDVIEW-II

Model	Sampling Strategy	steps	time(s)	WorldView-II			
				PSNR↑	SSIM↑	SAM↓	ERGAS↓
CrossDiff	Explicit Sampling	1000	128.69	40.4803	0.9604	0.0258	1.0823
CrossDiff	Explicit Sampling	100	54.92	35.0817	0.8835	0.0373	1.9101
DDRF	Explicit Sampling	100	93.41	41.2757	0.9659	0.0238	0.9976
DDRF	Implicit Sampling	25	28.27	42.2880	0.9732	0.0210	0.8880
DDRF	Implicit Sampling	2	8.90	42.0884	0.9719	0.0215	0.9050
PanDiT	Explicit Sampling	100	141.93	40.8406	0.9628	0.0249	1.0386
PanDiT	Implicit Sampling	25	38.66	42.6385	0.9749	0.0202	0.8565
PanDiT	Implicit Sampling	2	10.18	42.6385	0.9748	0.0202	0.8564

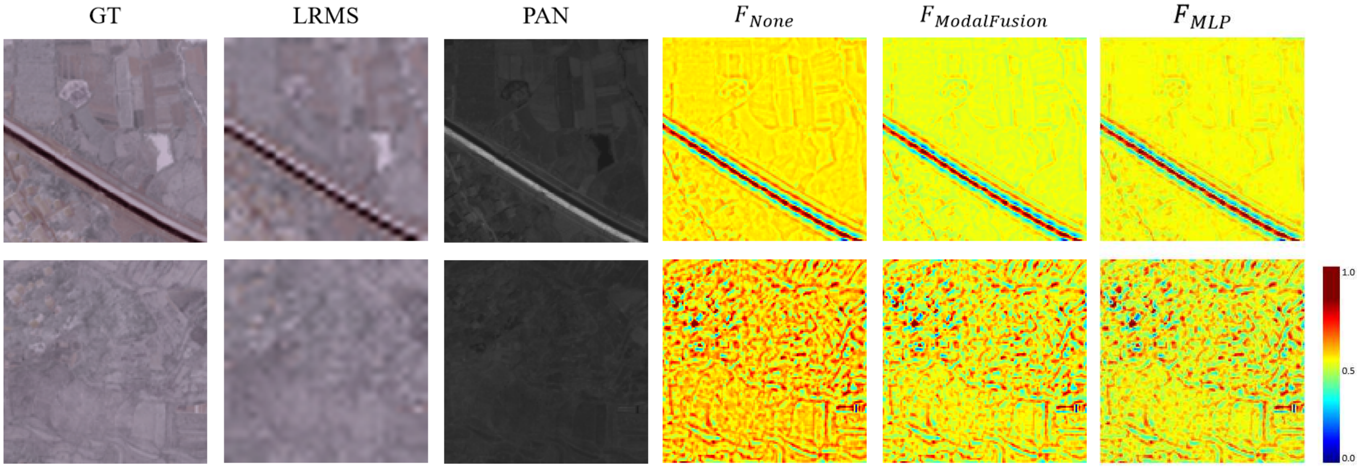

Fig. 9. Feature maps from the last DiT Block of different model variants.

F. Ablation Study

To validate the effectiveness of the proposed components in our model, we conducted a series of ablation studies. All experiments were performed on the WorldView-II dataset and evaluated using implicit sampling with a sampling step of 2, ensuring consistency across all experimental variants. The ablation study can be divided into two parts. The first part involves replacing components of the condition injection blocks to assess the contributions of the proposed SEB, TFEB, and MFB. The second part compares our condition injection block with the previous approach used in DiT, where MLPs were directly applied to extract and inject conditional information.

The experimental results are summarized in Table V. In Fig. 9, we present feature maps from the final DiT Block and the direct prediction results of selected model variants in the ablation study on a representative sample. The visualizations from left to right include the ground truth, the LRMS image, the PAN image, the feature map from the original PanDiT model, the feature map from the variant without the MFB, and the feature map from the variant using an MLP as the condition injection block.

TABLE V
ABLATION FOR PANDIT ON WORLDVIEW-II DATASETS. "A → B" MEANS REPLACING A WITH B, "TOTAL" REPRESENTS THE ENTIRE CONDITION INJECTION BLOCK, AND "NONE" REPRESENTS THE ORIGINAL PANDIT MODEL

Variant	WorldView-II			
	PSNR↑	SSIM↑	SAM↓	ERGAS↓
SEB → Conv	42.4596	0.9735	0.0207	0.8737
TFEB → Conv	42.4090	0.9733	0.0208	0.8778
TFEB → SEB	42.4904	0.9740	0.0206	0.8691
SEB → TFEB	42.3372	0.9727	0.0210	0.8865
Total → MLP	41.6133	0.9681	0.0226	0.9561
None	42.6385	0.9748	0.0202	0.8564

1) Effectiveness of the Spatial and TFEB: We propose two distinct methods for extracting features from LRMS and PAN images: the SEB, based on spatial pyramids, and the TFEB, based on wavelet transforms. In this ablation, we replaced each feature extraction module with convolutional layers to maintain the overall parameters. As shown in the results, substituting either feature extraction method resulted in varying degrees of performance degradation, with PSNR dropping by 0.18 and 0.23 dB, respectively. These findings highlight the effectiveness of our feature extraction methods in enhancing the model's ability to capture essential features from LRMS and PAN images.

2) Effectiveness of the Dual-Branch Feature Extraction Mechanism: In PanDiT, we propose to explicitly extract both spatial and time–frequency domain features to help the model better analyze information from different frequency components, thereby achieving more accurate HRMS reconstruction. To verify the effectiveness of this design, we conducted ablation experiments in which the model was tested using only spatial features or only time–frequency features for image fusion. The results show that, compared with the full model, the PSNR decreased by 0.14 and 0.30 dB, respectively. These findings indicate that spatial and time–frequency features are highly complementary: spatial features focus on local structural details, while time–frequency features capture global frequency information. The combination of both enables a more comprehensive feature representation and significantly improves reconstruction quality.

3) Effectiveness of the MFB: The MFB is designed to integrate features from multiple modalities by extracting both the discrepant and common features between LRMS and PAN images. In the ablation study, we replaced the fusion mechanism by simply concatenating the features from LRMS and PAN images. The test results show that removing the MFB led to a noticeable decline in performance, with PSNR dropping by 0.33 dB.

The visualized feature maps further reveal that eliminating the MFB significantly weakens the model's ability to capture fine texture details. Compared to the original PanDiT model, the activation in high-frequency texture regions is substantially reduced. This suggests that, without the fusion module, the model's ability to capture the rich details in PAN images is severely diminished. Moreover, the model's capacity to represent continuous low-frequency structures from LRMS is compromised, leading to fragmented or discontinuous coherent regions. These observations underscore the critical role of the MFB in enabling the model to effectively perceive both detailed and structural information from multisource images.

4) Effectiveness of the Condition Injection Block: In most generative tasks, conditional information of DiTs is typically injected through an MLP to guide the backbone in denoising and reconstruction. However, the simple MLPs restrict the model's capability to model spatial correlations between image patches, impairing the recovery of high-frequency details and compromising fusion quality. To address this limitation, we propose an improved condition injection block, which augments the traditional MLP by incorporating a convolutional structure based on a sliding window mechanism. Leveraging the local receptive field of convolutions, this design enables the model to more effectively capture image details and local structures during conditional information injection, thereby enhancing its ability to model high-frequency content.

Experimental results corroborate the effectiveness of our approach. When the condition injection block is replaced with a pure MLP structure, the spectral reconstruction accuracy of the fusion results deteriorates notably, which leads to clear distortions. These findings highlight the critical role of conditional information injection in diffusion models, which determines the final reconstruction quality.

The visualized feature maps also reveal that although the MLP-based structures can partially perceive high-frequency features and activate certain texture areas, they still face difficulties in modeling edge structures. Specifically, the feature maps extracted by the MLP often show broken responses and incomplete structures in regions where edges are expected to be continuous. This demonstrates that MLP cannot directly model local spatial neighborhoods. They tend to compress features globally, overlooking the continuous and structured nature of local image features. As a result, MLPs struggle with edge contours and fine-grained textures, and are insufficient for tasks requiring spatial detail continuity. In contrast, the feature maps generated by the original PanDiT show clearer and more continuous responses to structural elements such as edges and contours. This suggests that our proposed condition injection block is more effective at preserving spatial structure and geometric information during feature extraction, particularly excelling at modeling detailed features.

V. Conclusion

In this study, we propose a novel pansharpening network, PanDiT, inspired by the widely adopted DiT architecture in generative tasks. In network design, we introduce two tailored modules to effectively extract conditional information from the input images: the SEB, constructed using a spatial feature pyramid, and the TFEB, which leverages wavelet transforms. Additionally, we design an MFB to capture both the complementary features between the LRMS and PAN modalities. From the perspective of sampling strategy, PanDiT adopts an implicit sampling mechanism within the diffusion framework, significantly reducing the number of inference steps required and thus enhancing generation efficiency. Extensive experiments demonstrate that PanDiT outperforms state-of-the-art methods, achieving superior fusion of spectral and spatial details. The proposed framework not only delivers high-quality pansharpening but also offers improved computational efficiency, making it a promising solution for real-world remote sensing image fusion.

References

[1] C. A. Bishop, J. G. Liu, and P. J. Mason, "Hyperspectral remote sensing for mineral exploration in Pulang, Yunnan Province, China," *Int. J. Remote Sens.*, vol. 32, no. 9, pp. 2409–2426, May 2011.

[2] J. Yang et al., "The role of satellite remote sensing in climate change studies," *Nature Climate Change*, vol. 3, no. 10, pp. 875–883, Oct. 2013.

[3] R. J. Lee and S. Steele, "Military use of satellite communications, remote sensing, and global positioning systems in the war on terror," *J. Air L. Com.*, vol. 79, no. 1, p. 69, 2014.

[4] A. R. Gillespie, A. B. Kahle, and R. E. Walker, "Color enhancement of highly correlated images. II. Channel ratio and 'chromaticity' transformation techniques," *Remote Sens. Environ.*, vol. 22, no. 3, pp. 343–365, Aug. 1987.

[5] J. G. Liu, "Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details," *Int. J. Remote Sens.*, vol. 21, no. 18, pp. 3461–3472, Jan. 2000.

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

[7] J. Zhong, B. Yang, G. Huang, F. Zhong, and Z. Chen, "Remote sensing image fusion with convolutional neural network," *Sens. Imag.*, vol. 17, no. 1, pp. 1–16, Dec. 2016.

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

[9] M. Zhou, J. Huang, Y. Fang, X. Fu, and A. Liu, "Pan-sharpening with customized transformer and invertible neural network," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 3553–3561.

[10] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[11] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. 38th Int. Conf. Mach. Learn.*, Jul. 2021, pp. 8162–8171.

[12] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," 2020, *arXiv:2011.13456*.

[13] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," 2020, *arXiv:2010.02502*.

[14] C. Lü, Y. Zhou, F. Bao, J. F. Chen, C. Li, and J. Zhu, "DPM-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 5775–5787.

[15] Q. Meng, W. Shi, S. Li, and L. Zhang, "PanDiff: A novel pansharpening method based on denoising diffusion probabilistic model," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5611317.

[16] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 4195–4205.

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

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

[19] M. Zhou, K. Yan, J. Huang, Z. Yang, X. Fu, and F. Zhao, "Mutual information-driven pan-sharpening," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1798–1808.

[20] H. Zhou, Q. Liu, and Y. Wang, "PanFormer: A transformer based model for pan-sharpening," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2022, pp. 1–6.

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

[22] Z. Cao, S. Cao, X. Wu, J. Hou, R. Ran, and L.-J. Deng, "DDRF: Denoising diffusion model for remote sensing image fusion," 2023, *arXiv:2304.04774*.

[23] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 11895–11907.

[24] J. Chen et al., "PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis," 2023, *arXiv:2310.00426*.

[25] K. Cheng et al., "Effective diffusion transformer architecture for image super-resolution," in *Proc. AAAI Conf. Artif. Intell.*, 2025, vol. 39, no. 3, pp. 2455–2463.

[26] L. Wang, Q. Yang, C. Wang, W. Wang, J. Pan, and Z. Su, "Learning a coarse-to-fine diffusion transformer for image restoration," 2023, *arXiv:2308.08730*.

[27] H. Yan, S. Xiong, L. Wang, L. Jian, and G. Vivone, "ATFusion: An alternate cross-attention transformer network for infrared and visible image fusion," 2024, *arXiv:2401.11675*.

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

[29] W. Liao et al., "Two-stage fusion of thermal hyperspectral and visible RGB image by PCA and guided filter," in *Proc. 7th Workshop Hyperspectral Image Signal Process., Evol. Remote Sens. (WHISPERS)*, Jun. 2015, pp. 1–4.

[30] R. Haydn, "Application of the ihs color transform to the processing of multisensor data and image enhancement," in *Proc. Int. Symp. Remote Sens. Arid Semi-Arid Lands*, Cairo, Egypt, 1982, pp. 599–616.

[31] J. Cai and B. Huang, "Super-resolution-guided progressive pansharpening based on a deep convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5206–5220, Jun. 2021.

[32] X. He et al., "Multiscale dual-domain guidance network for pan-sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5403213.

[33] G. Yang et al., "PanFlowNet: A flow-based deep network for pan-sharpening," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 16811–16821.

[34] X. He et al., "Pan-Mamba: Effective pan-sharpening with state space model," *Inf. Fusion*, vol. 115, Mar. 2025, Art. no. 102779.

[35] Y. Xing, L. Qu, S. Zhang, K. Zhang, Y. Zhang, and L. Bruzzone, "CrossDiff: Exploring self-supervised representation of pansharpening via cross-predictive diffusion model," *IEEE Trans. Image Process.*, vol. 33, pp. 5496–5509, 2024.

[36] R. H. Yuhas, A. Goetz, and J. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm," in *Proc. JPL*, 1992, pp. 147–149.