Accelerated Diffusion via High-Low Frequency Decomposition for Pan-Sharpening

Ge Meng[1, 2], Jingjia Huang[1, 2], Jingyan Tu[1, 2], Yingying Wang[1, 3], Yunlong Lin[1, 2], Xiaotong Tu[1, 2], Yue Huang[1, 2, 3], Xinghao Ding[1, 2, 3]*

[1] Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, China
[2]School of Informatics, Xiamen University, China
[3]Institute of Artificial Intelligence, Xiamen University, China
{mengg, huangjj, tujingyan, wangyingying7, linyl}@stu.xmu.edu.cn, {xttu, yhuang2010, dxh}@xmu.edu.cn

Abstract

Pan-sharpening aims to preserve the spectral information of the multi-spectral (MS) image while leveraging the high-frequency details from the guided high-resolution panchromatic (PAN) image to enhance its spatial resolution. The key challenge is how to preserve the spectral information from the MS image and the spatial details from the PAN image as much as possible. Diffusion models have achieved favorable results in image restoration and synthesis tasks but suffer from excessive computational resource and time consumption. In this paper, we design a novel and computationally efficient diffusion-based pan-sharpening network that achieves accelerated diffusion while reducing task complexity by decoupling the high and low-frequency components of the fused image. Specifically, leveraging the information-preserving characteristic of the wavelet transformation, we introduce a Wavelet-based Low-frequency Diffusion Model (WLDM). WLDM generates the low-frequency coefficient of high-resolution MS (HRMS) image from the low-resolution MS (LRMS) image. This approach significantly reduces computational resources and complexity compared to the direct restoration of the HRMS image. Furthermore, we have devised a High-frequency Information Restoration Module (HIRM) to restore the high-frequency information in the HRMS image through the interaction of high-frequency coefficients from the PAN image in three directions. Extensive experiments on three different datasets demonstrate that our method outperforms existing approaches in both quantitative metrics, qualitative metrics, and inference efficiency.

Introduction

Compared to traditional RGB images, multispectral (MS) images contain richer spectral information, which leads to the widespread utilization in downstream tasks such as environmental monitoring, land classification, and so on. However, constrained by the physical limitations of sensors, satellites typically capture low-resolution multispectral (LRMS) images along with corresponding high-resolution panchromatic (PAN) images. The purpose of pan-sharpening is to fuse the complementary information from the LRMS and PAN images to generate high-resolution multispectral (HRMS) images.

*Corresponding Author.

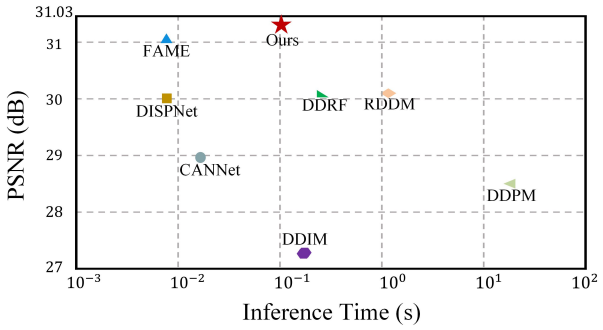

Figure 1: PSNR and inference time of comparison with state-of-the-art methods on the WorldView-III test set. Our method achieves a better balance between efficiency and performance.

Traditional pan-sharpening methods mainly include component substitution (CS) (Carper et al. 1990; Shah, Younan, and King 2008), multi-resolution analysis (MRA) (Mallat 1989; Li and Leung 2008), and variational optimization (VO) methods (Tian et al. 2020, 2021). These methods have proven success in image fusion tasks, but their effectiveness heavily relies on handcrafted features extracted based on prior knowledge. In recent years, due to the powerful representation capabilities of deep neural networks, researchers have designed numerous models based on CNNs to address pan-sharpening (Fu et al. 2021; Wang et al. 2023; Lin et al. 2023). However, these methods often lack optimization for visual fidelity as perceived by humans.

Therefore, following a perceptual-driven approach, some scholars have designed pan-sharpening methods based on Generative Adversarial Networks (GANs) (Xu et al. 2023; Zhou et al. 2022a). These methods introduce a discriminator to constrain both the high-frequency fidelity and spectral fidelity of the fused results. However, the generation process of GANs is often unstable, with no assurance of consistently producing high-quality HRMS images. Recently, diffusion models have demonstrated powerful capabilities in image restoration (Lin et al. 2024b,a) and image synthesis (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020) tasks. Diffusion models rely on the hierarchical struc-

ture of denoising autoencoders to achieve high-quality mapping from randomly sampled Gaussian noise to the target image or latent distribution (Rombach et al. 2022). Diffusion models divide the sampling process into multiple steps, mitigating instability and mode collapse issues compared to previous generative models. However, diffusion models still encounter challenges, such as excessive consumption of computational resource and longer inference time. As illustrated in Figure 1, our proposed diffusion-based method can achieve a favorable tradeoff between high-quality fusion results and relatively low inference times.

In this paper, we propose a novel and efficient diffusion-based pan-sharpening network. Specifically, we introduce a Wavelet-based Low-frequency Diffusion Model (WLDM). Owing to the information-preserving property of wavelet transformation (WT), we apply WT to the HRMS image, extracting both its low-frequency coefficient and high-frequency coefficients in the horizontal, vertical, and diagonal directions. Compared to the original image, the spatial dimension of coefficients in the wavelet domain is significantly reduced. Notably, we observe that the LRMS image is closer to the low-frequency coefficient of the HRMS image. Therefore, we directly utilize the diffusion model to denoise the LRMS image and generate the low-frequency coefficient of the HRMS, reducing the computational resources consumed in the inverse diffusion process. Furthermore, we design a High-frequency Information Restoration Module (HIRM) to generate the high-frequency information in the HRMS image. HIRM integrates the high-frequency information from three directions of the PAN image, while the spectral information from the MS image is incorporated into the network to ensure the fidelity of the generated high-frequency coefficients. Our contributions can be summarized as follows:

- We propose a novel and efficient pan-sharpening network that leverages both the powerful generative capability of the diffusion model and the strengths of wavelet transformation.
- We introduce a Wavelet-based Low-frequency Diffusion Model (WLDM), which reduces computational resource and time consumption during the diffusion sampling process by generating low-frequency coefficient of the HRMS image.
- We design a High-frequency Information Restoration Module (HIRM) to enhance local details in the HRMS image through the interaction of high-frequency information from different directions in the PAN image.
- Extensive experiments on three satellite datasets demonstrate that our method outperforms existing approaches in both quantitative and qualitative metrics. Furthermore, the inference time of our method is significantly reduced compared to other diffusion-based methods.

Related Work

Component substitution (CS), multi-resolution analysis (MRA), and variational optimization (VO) are three typical traditional pan-sharpening methods. The CS methods (Kwarteng and Chavez 1989; Carper et al. 1990; Gillespie, Kahle, and Walker 1987) mainly project the MS image into a certain space and subsequently enhance the MS spatial resolution by substituting the spatial information components with the PAN image. The MRA methods use multi-resolution decomposition methods such as laplacian pyramid (LP) (Vivone et al. 2014) and decimated wavelet transform (DWT) (Mallat 1989) to extract the spatial information of the PAN image, and then inject it into the MS image, resulting in less spectral distortion. The VO methods (Ballester et al. 2006; Jiang et al. 2015) formulate energy functions based on prior knowledge and solve objective functions to obtain the final HRMS image. With the rapid development of deep learning in recent years, many scholars have been utilizing CNNs (Xia et al. 2024; Zhou et al. 2023; Wang et al. 2023, 2024b; He et al. 2024a) due to their powerful representation capabilities. Besides, (Zhou et al. 2022b) exploits a customized transformer architecture and information-lossless invertible neural module for long-range dependencies modeling and effective feature fusion. However, these methods often overlook the visual fidelity as perceived by humans in the fusion results.

In recent years, due to the powerful generative capabilities exhibited by GAN networks (Goodfellow et al. 2017), some scholars have approached pan-sharpening as an image generation problem and explored the utilization of GAN to address it. PSGAN (Liu, Wang, and Liu 2018) is the first work to introduce GAN to address pan-sharpening task. UC-GAN (Zhou et al. 2022a) introduces a novel hybrid loss based on the cycle-consistency and adversarial scheme to improve the performance. Pan-GAN (Ma et al. 2020) uses the generator to establish adversarial games separately with the spectral discriminator and the spatial discriminator. But the generation process of GAN-based methods lack rigorous mathematical theory, often resulting in unstable outcomes.

Diffusion models (Austin et al. 2021; Gu et al. 2022; Kingma et al. 2021) transform complex data distribution into simple noise distribution and recover data from noise. Diffusion-based generative models have produced promising results in tasks such as super-resolution (Luo et al. 2023), deblurring (Ren et al. 2022), and inpainting (Lugmayr et al. 2022) with improvements in denoising diffusion probability models. RDDM (Liu et al. 2023) introduces residuals to guide the reverse diffusion process. DDRF (Cao et al. 2023) incorporates coarse-grained style information along with fine-grained high-frequency and low-frequency information into the diffusion U-Net. However, DDRF designs complex and cumbersome network structures to supplement more spatial and spectral information for U-Net, resulting in a significant increase in the consumption of computational resources and inference time during the sampling process.

Methodology

Figure 2 shows the overall architecture of our proposed pan-sharpening network, which mainly consists of two parts, Wavelet-based Low-frequency Diffusion Model (WLDM) and High-frequency Information Restoration Module (HIRM). The details will be illustrated as below.

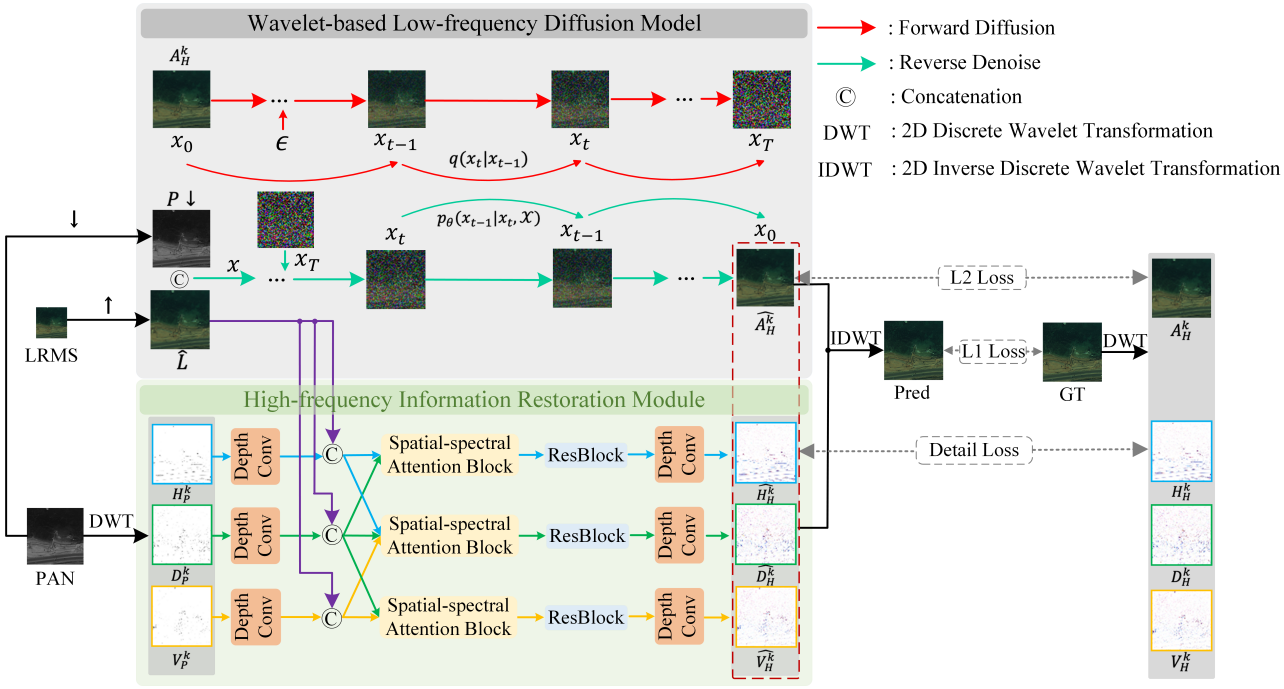

Figure 2: The overall framework of our network. It consists of two key parts: Wavelet-based Low-frequency Diffusion Model (WLDM) and High-frequency Information Restoration Module (HIRM).

Problem Formulation

Pan-sharpening aims to establish a stable mapping function $\mathcal{F}(\cdot)$ from the LRMS image L and the PAN image P to the ideal HRMS image H:

$$H = \mathcal{F}(L, P) \tag{1}$$

Let $\widehat{L}$ denote the upsampled LRMS image which has the same spatial dimension as the PAN image. The objective of the pan-sharpening task is to generate a fused image that possesses both the spectral information of $\widehat{L}$ and the spatial information of P. A common paradigm is to inject spatial detail information from P into $\widehat{L}$:

$$\widehat{H} = \widehat{L} + \mathcal{G}(P - P_L) \tag{2}$$

where $\mathcal{G}(\cdot)$ is the injection function, P_L is generated by employing low-pass filtering on P.

Discrete Wavelet Transformation

Given a PAN image $P \in R^{H\times W\times 1}$, a HRMS image $H \in R^{H\times W\times C}$ and a LRMS image $L \in R^{\frac{H}{r}\times\frac{W}{r}\times C}$, we use 2D discrete wavelet transformation (DWT) with Haar wavelets to transform the P and H into four sub-bands respectively:

$$\begin{aligned}[A_P, H_P, D_P, V_P] &= \mathcal{D}(P)\\ [A_H, H_H, D_H, V_H] &= \mathcal{D}(H)\end{aligned} \tag{3}$$

where $\mathcal{D}(\cdot)$ represents DWT, A, H, D, $V \in R^{\frac{H}{2}\times\frac{W}{2}\times 1}$ or $R^{\frac{H}{2}\times\frac{W}{2}\times C}$ represent the low-frequency and high-frequency components in the horizontal, diagonal, and vertical directions respectively. It can be observed that the spatial dimension of the wavelet component processed by the diffusion model after one wavelet transformation is four times smaller than the original image. More generally, we can obtain the components after k times DWT:

$$\begin{aligned}[A_P^k, H_P^k, D_P^k, V_P^k] &= \mathcal{D}\left(A_P^{k-1}\right)\\ [A_H^k, H_H^k, D_H^k, V_H^k] &= \mathcal{D}\left(A_H^{k-1}\right)\end{aligned} \tag{4}$$

To ensure the spatial dimensions of L match those of A_H^k, an upsampling operation is applied to L. The upsampling operation is optional, which depends on the value of k:

$$\widehat{L} = \mathrm{UP}(L) \tag{5}$$

Wavelet-based Low-frequency Diffusion Model

To reduce the computational resource consumption during the diffusion process, we design WLDM to generate the low-frequency coefficients of the HRMS image from the LRMS image. This approach helps to simultaneously decrease the task complexity. The diffusion process is a Markov chain that gradually corrupts data x_0 until it approaches Gaussian noise $x_T \sim \mathcal{N}(0, I)$ at T sampling time steps, which can be formulated as:

$$\begin{aligned}q(x_1, \cdots, x_T \mid x_0) &= \prod_{t=1}^{T} q(x_t \mid x_{t-1})\\ q(x_t \mid x_{t-1}) &= \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right)\end{aligned} \tag{6}$$

where t denotes as diffusion step and β_t are fixed or learned variance schedule. During the forward noising process, any step x_t may be sampled directly from x_0 through the following equation:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \tag{7}$$

where $\epsilon \sim \mathcal{N}(0, I)$, $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, so there is a closed form expression for $q(x_t \mid x_0)$:

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t) I\right) \tag{8}$$

The inverse process iteratively denoises a sampled Gaussian noise to a clean image. Starting from noise $x_T \sim \mathcal{N}(0, I)$, the reverse process from latent x_T to clean data x_0 is defined as below:

$$\begin{aligned} p_\theta(x_0, \cdots, x_{T-1} \mid x_T) &= \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t) \\ p_\theta(x_{t-1} \mid x_t) &= \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I\right) \end{aligned} \tag{9}$$

In our method, x_0 corresponds to the expected low-frequency coefficient A_H^k of the HRMS image H, and x_T represents an upsampled LRMS image $\widehat{L}(L)$. To fully leverage the data synthesis capability of the conditional diffusion model, we incorporate information from the PAN image to learn the conditional diffusion process:

$$\begin{aligned} p_\theta(x_0, \cdots, x_T) &= p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t, \mathcal{X}) \\ \mathcal{X} &= Concat\left(\mathrm{DOW}(P), \widehat{L}\right) \end{aligned} \tag{10}$$

where $\mathcal{X}$ is the condition information, $\mathrm{DOW}(\cdot)$ is the downsampling opertaion used to match the spatial dimensions of the PAN image P with its corresponding LRMS image L, and $p(x_T)$ is the standard normal prior. The mean $\mu_\theta(x_t, \mathcal{X}, t)$ is the target that we aim to estimate by a neural network ϵ_θ with parameters θ:

$$\mu_\theta(x_t, \mathcal{X}, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, \mathcal{X}, t)\right) \tag{11}$$

where $\epsilon_\theta(x_t, \mathcal{X}, t)$ is the noise vector predicted from the network ϵ_θ. The objective function is formulated as:

$$\mathcal{L}_\epsilon = E_{x_0, t, \epsilon_t \sim \mathcal{N}(0,I)} \left[\|\epsilon_t - \epsilon_\theta(x_t, \mathcal{X}, t)\|^2\right] \tag{12}$$

For the denoising process, we ensure the consistency of the reconstructed image content by minimizing the L2 distance between the reconstructed image $\widehat{A_H^k}$ and the reference image A_H^k:

$$\mathcal{L}_{diff} = \lambda_1 \mathcal{L}_\epsilon + \left\| \widehat{A_H^k} - A_H^k \right\|^2 \tag{13}$$

where λ_1 is the weight of loss $\mathcal{L}_\epsilon$, it is empirically set to 0.5 here.

High-frequency Information Restoration Module

We further designed HIRM to achieve high-quality restoration of both global and local details in the HRMS image. Figure 2 illustrates the details of HIRM. HIRM implements interaction among the high-frequency coefficients in the horizontal, vertical, and diagonal directions from the PAN image, while supplementing spectral information from the MS image. Two 3×3 convolutional layers and two 1×1 convolutional layers are first used to extract features of the three high-frequency coefficients. The horizontal and vertical high-frequency information supplements the diagonal, and vice versa for the other two directions. Then, an Spatial-spectral Attention Block is introduced to achieve spatial and spectral information interaction for high-frequency coefficients in different directions. We further employ a multi-scale dilated ResBlock to extract spatial features at different scales, enabling the restoration of both global and local details. Finally, three depth-wise separable convolutions are used to fuse the fine-grained multi-scale feature to obtain the reconstructed high-frequency coefficients. The wavelet coefficients generated by HIRM at scale k will be used to generate the low-frequency coefficient at scale $k-1$ through 2D inverse discrete wavelet transform (IDWT):

$$\widehat{A_H^{k-1}} = \mathcal{ID}\left(\left[\widehat{A_H^k}, \widehat{H_H^k}, \widehat{D_H^k}, \widehat{V_H^k}\right]\right) \tag{14}$$

where $\mathcal{ID}(\cdot)$ represents IDWT.

Loss Function

Besides the diffusion loss $\mathcal{L}_{diff}$ mentioned in Section 3.3, which is used to optimize the diffusion model, we have also designed a detail loss to ensure the reconstruction quality of wavelet coefficients:

$$\mathcal{L}_{detail} = \sum_{k=1}^{K} \left\| \left[\widehat{H_H^k}, \widehat{D_H^k}, \widehat{V_H^k}\right] - \left[H_H^k, D_H^k, V_H^k\right] \right\|^2 \tag{15}$$

where k is the number of DWT times. Moreover, we utilize a content loss $\mathcal{L}_{content}$ that use L1 loss to minimize the content difference between the restored image $\widehat{H}$ and the reference HRMS image H:

$$\mathcal{L}_{content} = \frac{1}{n} \sum_{i}^{n} \left| \widehat{H}(i) - H(i) \right|_1 \tag{16}$$

where n is the total number of sampled pixels, $\widehat{H}(i)$ represents the predicted value for pixel i, and $H(i)$ represents its corresponding ground truth. The total loss $\mathcal{L}_{total}$ is defined by combing the diffusion loss, the detail loss, and the content loss as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{diff} + \mathcal{L}_{detail} + \mathcal{L}_{content} \tag{17}$$

Experiments

Baseline Methods

To demonstrate the effectiveness of our method, we compare the performance of our method with three categories of existing state-of-the-art methods, including 1) four traditional pan-sharpening methods Brovey (Gillespie, Kahle, and Walker 1987), GS (Laben and Brower 2000), IHS (Carper et al. 1990), and GFPCA (Liao et al. 2015), 2) three deep learning-based pan-sharpening methods, CANNet (Duan et al. 2024), DISPNet (Wang et al. 2024a) and FAME (He et al. 2024b), and 3) four diffusion-based methods DDPM (Ho, Jain, and Abbeel 2020), DDIM (Song, Meng, and Ermon 2020), RDDM (Liu et al. 2023), and DDRF (Cao et al. 2023). For fairness, we also consider the PAN image as the input condition for other diffusion-based methods.

Method	WorldView-II				WorldView-III				GaoFen2			
	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
Brovey	35.8646	0.9216	0.0403	1.8238	22.5060	0.5466	0.1159	8.2331	37.7974	0.9026	0.0218	1.3720
GS	35.6376	0.9176	0.0423	1.8774	22.5608	0.5470	0.1217	8.2433	37.2260	0.9034	0.0309	1.6736
IHS	35.2962	0.9027	0.0461	2.0278	22.5579	0.5354	0.1266	8.3616	38.1754	0.9100	0.0243	1.5336
GFPCA	34.5581	0.9038	0.0488	2.1411	22.3344	0.4826	0.1294	8.3964	37.9443	0.9204	0.0314	1.5604
CANNet	41.5627	0.9671	0.0230	0.9556	28.9690	0.8918	0.0950	3.6681	47.9369	0.9874	0.0103	0.5122
DISPNet	41.8768	0.9702	0.0221	0.9157	30.0426	0.9153	0.0776	3.2620	47.4529	0.9898	0.0111	0.5532
FAME	42.0262	0.9723	0.0215	0.9172	30.9903	0.9287	0.0697	2.9531	47.6721	0.9898	0.0098	0.5542
DDPM	40.2193	0.9598	0.0272	1.1057	28.4039	0.8785	0.0946	3.9593	46.8791	0.9870	0.0117	0.5907
DDIM	35.6012	0.9076	0.0421	2.0094	27.2983	0.7741	0.1193	5.0762	38.9768	0.9142	0.0312	1.4877
RDDM	40.8172	0.9702	0.0277	1.0795	30.1007	0.9225	0.0841	3.2946	44.6996	0.9855	0.0163	0.7751
DDRF	41.2331	0.9659	0.0240	0.9926	30.0801	0.9105	0.0825	3.2783	45.3347	0.9790	0.0144	0.6799
Ours	**42.5780**	**0.9735**	**0.0203**	**0.8432**	**31.0295**	**0.9298**	**0.0683**	**2.9475**	**48.7757**	**0.9901**	**0.0094**	**0.4632**

Table 1: Quantitative comparison of reference metrics. The best values are bolded. The up or down arrow indicates higher or lower metric corresponding to better images.

Experiment Settings

Implementation Details. We implement our network on the PC with a single NVIDIA TITAN RTX 3090 GPU, and build our network in Pytorch framework. For our Wavelet-based Low-frequency Diffusion Model, the commonly used U-Net architecture (Ronneberger, Fischer, and Brox 2015) is adopted as the noise estimator network. The Adam optimizer is adopted for optimization. The initial learning rate is set to 1×10^{-4}. The batch size is set to 8. During the forward and reverse processes, the time step T is set to 200 for the training phase, and the implicit sampling step S is set to 10 for both the training and inference phases.

Datasets. The paired training samples are unavailable in practice. To create the training dataset, we employ the Wald protocol (Wald, Ranchin, and Mangolini 1997) for generating the necessary paired samples. For instance, given an original high-resolution MS image $H \in R^{H \times W \times C}$ and its corresponding PAN image $\bar{P} \in R^{rH \times rW \times c}$, both are downscaled by a ratio r to produce image pairs $M \in R^{\frac{H}{r} \times \frac{W}{r} \times C}$ and $P \in R^{H \times W \times c}$ in the training set. Our experiments involve three satellite image datasets, namely WorldView-II, WorldView-III, and GaoFen2, each comprising several hundred PAN and LRMS image pairs. The PAN images are cropped into patches with the size of 128×128, and the corresponding LRMS patches are with the size of 32×32.

Metrics. We use the peak signal-to-noise ratio (PSNR), the structural similarity (SSIM), SAM, and the relative dimensionless global error in synthesis (ERGAS) as quantitative metrics to evaluate the quality of the fused HRMS image. Additionally, to compare the models' generalization ability, we test them on 200 full-resolution real datasets without down-sampling. Since this dataset does not contain ground truth, we use three no-reference image quality evaluation metrics to assess model performance, including the spectral distortion index D_λ, the spatial distortion index D_S, and the quality without reference QNR. Besides, the average inference time and GPU memory costs are used to compare the efficiency of different methods.

Performance Comparison

Table 1 presents the performance of our proposed method and the baseline methods on three datasets, with the best results bolded. Our method outperforms existing pan-sharpening methods in all metrics. Specifically, compared to the second-best results, our method achieves a PSNR improvement of 0.55 dB, 0.03 dB, and 0.83 dB on the WorldView-II, WorldView-III and GaoFen2 datasets respectively. Moreover, our method has shown significant improvements over other metrics as well. Additionally, we show the comparison of the visual results in Figure 3 and Figure 4. The top two rows compare the fusion results with SOTA methods, and the bottom row are the MSE residues between the pan-sharpened results and the ground truth. Traditional methods tend to lose a significant amount of information during the process of feature dimension reduction, resulting in severe spatial and spectral distortion in the fusion results. By zooming in on the local regions, it is apparent that varying degrees of artifacts present in the results of current deep learning-based methods. Compared with other diffusion-based methods, it is easy to observe that our method shows clearer details, minimal spectral distortion, and a closer alignment to the ground truth. The poor performance of other diffusion-based methods can be attributed to the direct generation of the HRMS image with rich details, which significantly increases the difficulty of the diffusion process.

To evaluate the generalization ability of our network, we apply a pre-trained model trained on unseen full-resolution real dataset. Table 2 and Figure 5 show the results. Our method generates minimal spectral distortion and exhibits fewer artifacts and blurriness.

We further compare the memory consumption and inference time with other methods, and the results are shown in Table 3. For fairness, we set the sampling steps to 1000 for DDPM and set the sampling steps to 10 for other diffusion-based methods. Table 3 shows that our method significantly outperforms the other diffusion-based methods in terms of computational efficiency.

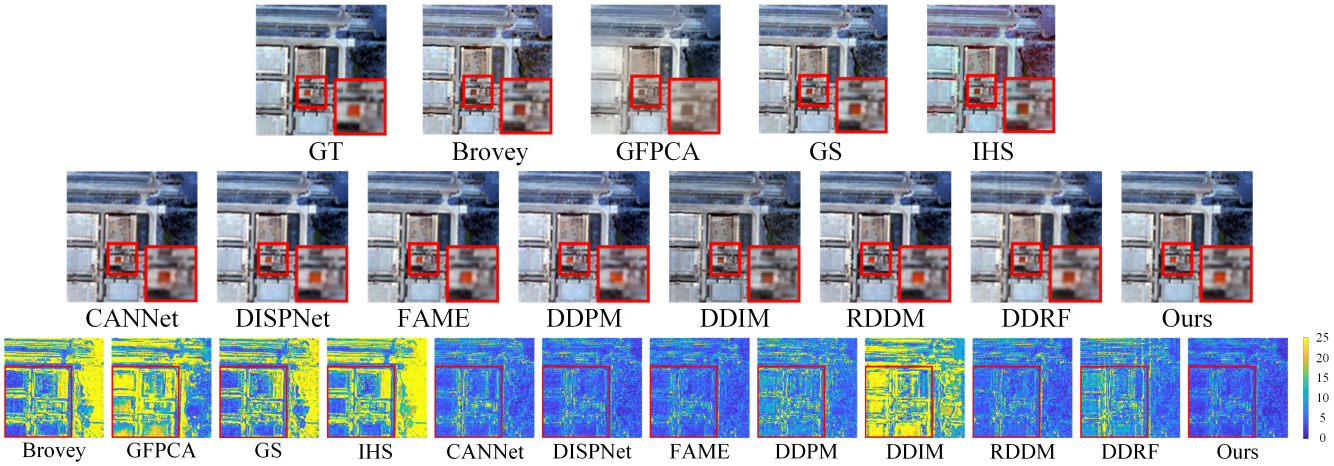

Figure 3: The visual comparisons between other pan-sharpening methods and our method on WorldView-II satellite.

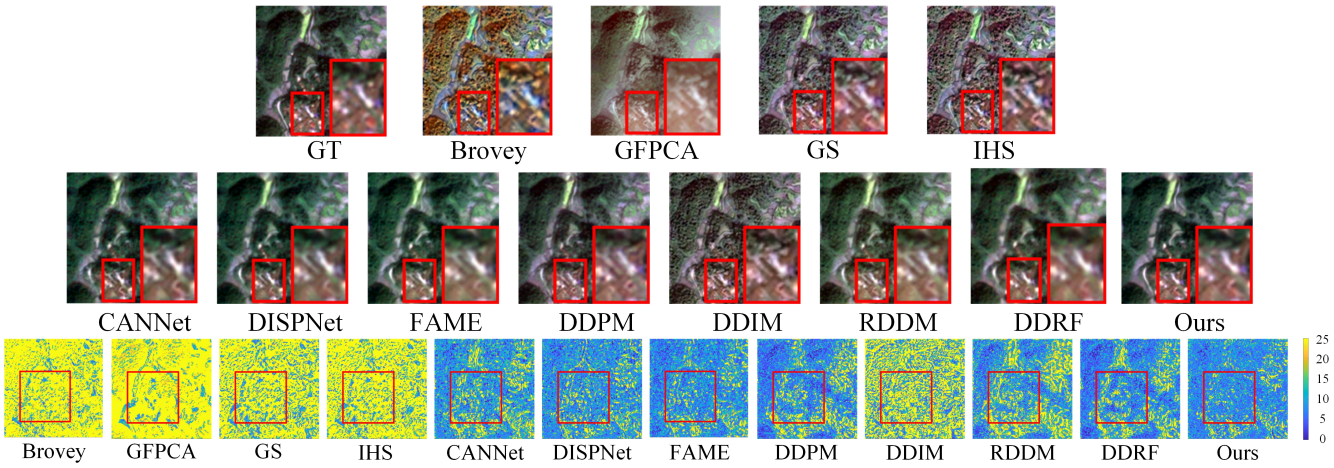

Figure 4: The visual comparisons between other pan-sharpening methods and our method on GaoFen2 satellite.

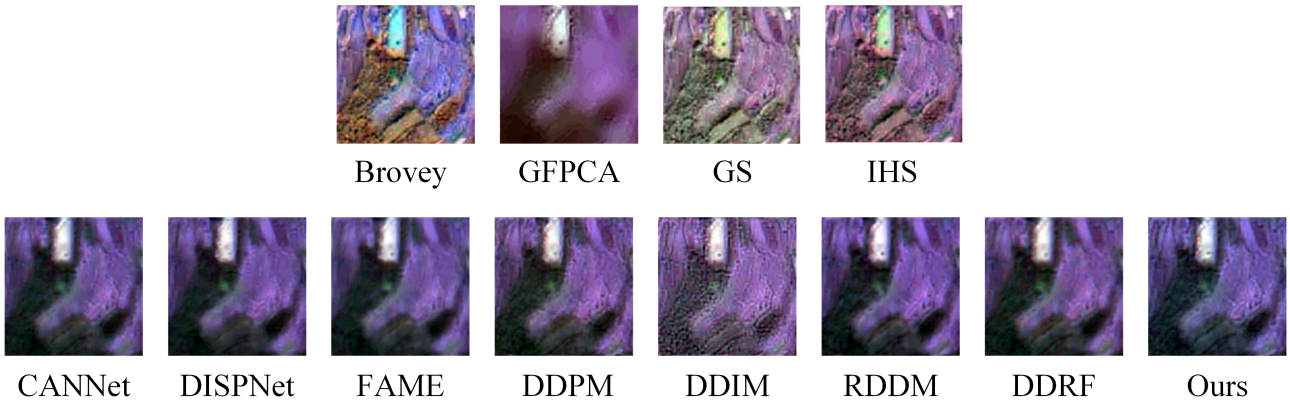

Figure 5: The visual comparisons between other pan-sharpening methods and our method on full-resolution dataset.

Ablation Studies

In this section, we further design a series of ablation experiments to independently demonstrate the effectiveness of the proposed HIRM and loss function, as well as to analyze the influence of wavelet transformation scale on the fusion result.

Metrics	GS	Brovey	IHS	GFPCA	CANNet	DISPNet	FAME	DDPM	DDIM	RDDM	DDRF	Ours
$D_\lambda\downarrow$	0.0696	0.1378	0.0770	0.0914	0.0861	0.0671	0.0674	0.0628	0.0714	0.0685	0.0804	**0.0611**
$D_S\downarrow$	0.2456	0.2605	0.2985	0.1635	0.1144	0.1826	0.1121	0.1136	0.2407	0.1055	0.1100	**0.1018**
QNR↑	0.7025	0.6390	0.6485	0.7615	0.7884	0.7638	0.8291	0.8319	0.6896	0.8345	0.8200	**0.8462**

Table 2: Evaluation on the full-resolution scenes from GaoFen2 dataset. The best values are bolded. The up or down arrow indicates higher or lower metric corresponding to better.

Method	Time(S)↓	Mem.(G)↓
CANNet	0.180	0.013
DISPNet	**0.010**	**0.008**
FAME	**0.010**	0.023
DDPM	0.289	0.602
DDIM	0.308	0.765
RDDM	1.171	1.196
DDRF	0.440	0.532
Ours	0.118	0.467

Table 3: The average time (seconds) and GPU memory (G) costs of different methods consumed on the WorldView-III test set.

Config	PSNR↑	SSIM↑	SAM↓	ERGAS↓
w/o HIRM	39.7350	0.8885	0.0190	1.4182
w/ HIRM	**48.7757**	**0.9901**	**0.0094**	**0.4632**

Table 4: Ablation studies of the HIRM on GaoFen2 dataset. The best values are bolded. 'w/o' denotes without, 'w/' denotes with.

Config	PSNR↑	SSIM↑	SAM↓	ERGAS↓
w/o $\mathcal{L}_{diff}$	46.5993	0.9755	0.0168	0.5945
w/o $\mathcal{L}_{detail}$	47.5214	0.9776	0.0151	0.4709
w/o $\mathcal{L}_{content}$	44.7655	0.9639	0.0188	0.6438
Ours	**48.7757**	**0.9901**	**0.0094**	**0.4632**

Table 5: Ablation studies of the loss function terms on GaoFen2 dataset. The best values are bolded. 'w/o' denotes without.

Wavelet scale	PSNR↑	SSIM↑	SAM↓	ERGAS↓
$K = 1$	**48.7757**	**0.9901**	**0.0094**	**0.4632**
$K = 2$	42.8220	0.9538	0.0199	0.9384
$K = 3$	40.7969	0.9366	0.0253	1.2173

Table 6: Ablation studies of the wavelet transformation scale on GaoFen2 dataset. The best values are bolded.

High-frequency Information Restoration Module (HIRM). To assess the impact of HIRM, we remove it to verify its necessity. We directly use the high-frequency coefficients from the PAN image in three directions to reconstruct the HRMS image. Table 4 shows that removing HIRM will degrade all metrics. This highlights the essential role of HIRM in providing complementary high-frequency information from different directions, which is vital for effective restoration of global and local details in the HRMS image. Meanwhile, the supplementation of spectral information from the MS image enhances the fidelity of high-frequency coefficients. Therefore, HIRM plays a significant role in our network.

Loss Function. We verify the effectiveness of each loss function by removing them individually, where the quantitative results are reported in Table 5. The diffusion loss $\mathcal{L}_{diff}$ is employed to ensure the stability of the generation process and the quality of the low-frequency coefficient. Since low-frequency coefficient contains a significant amount of spectral information, removing $\mathcal{L}_{diff}$ leads to a notable decrease in all metrics. The detail loss $\mathcal{L}_{detail}$ is employed to reconstruct more global and local spatial details. Therefore, removing the detail loss leads to a significant decrease in PSNR and SSIM, 1.25 dB and 0.01, respectively. Similarly, the incorporation of the content loss $\mathcal{L}_{content}$ leads to improvements in all metrics, with PSNR and SSIM increasing by 4.01 dB and 0.02, respectively.

Wavelet Transformation Scale. We validate the impact of performing diffusion processes on the low-frequency coefficient at different wavelet scales K. Table 6 reveals that with an increase in the number of wavelet transform iterations, the richness of input information received by U-Net decreases, resulting in performance degradation. We choose K = 1 as the default setting.

Conclusion

In this paper, we propose a novel and computationally efficient diffusion-based pan-sharpening network. By decoupling the high and low-frequency information, we not only simplify the image fusion process but also enhance the efficiency of diffusion. We design a Wavelet-based Low-frequency Diffusion Model (WLDM) to generate the low-frequency coefficient of the HRMS image from the LRMS image and a High-frequency Information Restoration Module (HIRM) to restore the high-frequency information in the HRMS image through the interaction of high-frequency coefficients from the PAN image. Extensive experiments on three different satellite datasets demonstrate the effectiveness and efficiency of our method.

Acknowledgments

The work was supported in part by the National Natural Science Foundation of China under Grant 82172033, U19B2031, 61971369, 52105126, 82272071, 62271430, and the Fundamental Research Funds for the Central Universities 20720230104.

References

Austin, J.; Johnson, D.; Ho, J.; Tarlow, D.; and Berg, R. 2021. Structured Denoising Diffusion Models in Discrete State-Spaces. *arXiv: Learning,arXiv: Learning*.

Ballester, C.; Caselles, V.; Igual, L.; Verdera, J.; and Rougé, B. 2006. A variational model for P+ XS image fusion. *International Journal of Computer Vision*, 69(1): 43.

Cao, Z.; Cao, S.; Wu, X.; Hou, J.; Ran, R.; and Deng, L.-J. 2023. DDRF: Denoising Diffusion Model for Remote Sensing Image Fusion.

Carper, W.; Lillesand, T.; Kiefer, R.; et al. 1990. The use of intensity-hue-saturation transformations for merging SPOT panchromatic and multispectral image data. *Photogrammetric Engineering and remote sensing*, 56(4): 459–467.

Duan, Y.; Wu, X.; Deng, H.; and Deng, L.-J. 2024. Content-Adaptive Non-Local Convolution for Remote Sensing Pansharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27738–27747.

Fu, X.; Wang, W.; Huang, Y.; Ding, X.; and Paisley, J. 2021. Deep Multiscale Detail Networks for Multiband Spectral Image Sharpening. *IEEE Transactions on Neural Networks and Learning Systems*, 2090–2104.

Gillespie, A. R.; Kahle, A. B.; and Walker, R. E. 1987. Color enhancement of highly correlated images. II. Channel ratio and "chromaticity" transformation techniques. *Remote Sensing of Environment*, 22(3): 343–365.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2017. GANGenerative Adversarial Nets. *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics*, 177–177.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, X.; Cao, K.; Zhang, J.; Yan, K.; Wang, Y.; Li, R.; Xie, C.; Hong, D.; and Zhou, M. 2024a. Pan-mamba: Effective pan-sharpening with state space model. *Information Fusion*, 102779.

He, X.; Yan, K.; Li, R.; Xie, C.; Zhang, J.; and Zhou, M. 2024b. Frequency-Adaptive Pan-Sharpening with Mixture of Experts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2121–2129.

Ho, J.; Jain, A.; and Abbeel, P. 2020. (DDPM)Denoising Diffusion Probabilistic Models. *Neural Information Processing Systems,Neural Information Processing Systems*.

Jiang, Y.; Ding, X.; Zeng, D.; Huang, Y.; and Paisley, J. 2015. Pan-sharpening with a hyper-Laplacian penalty. In *Proceedings of the IEEE International Conference on Computer Vision*, 540–548.

Kingma, D.; Salimans, T.; Poole, B.; and Ho, J. 2021. Variational Diffusion Models. *Cornell University - arXiv,Cornell University - arXiv*.

Kwarteng, P.; and Chavez, A. 1989. Extracting spectral contrast in Landsat Thematic Mapper image data using selective principal component analysis. *Photogramm. Eng. Remote Sens*, 55(1): 339–348.

Laben, C. A.; and Brower, B. V. 2000. Process for enhancing the spatial resolution of multispectral imagery using pansharpening. US Patent 6,011,875.

Li, Z.; and Leung, H. 2008. Fusion of multispectral and panchromatic images using a restoration-based method. *IEEE transactions on geoscience and remote sensing*, 47(5): 1482–1491.

Liao, W.; Huang, X.; Van Coillie, F.; Thoonen, G.; Pižurica, A.; Scheunders, P.; and Philips, W. 2015. Two-stage fusion of thermal hyperspectral and visible RGB image by PCA and guided filter. In *2015 7th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 1–4. Ieee.

Lin, Y.; Fu, Z.; Meng, G.; Wang, Y.; Dong, Y.; Fan, L.; Yu, H.; and Ding, X. 2023. Domain-irrelevant Feature Learning for Generalizable Pan-sharpening. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3287–3296.

Lin, Y.; Fu, Z.; Wen, K.; Ye, T.; Chen, S.; Meng, G.; Wang, Y.; Huang, Y.; Tu, X.; and Ding, X. 2024a. Unsupervised Low-light Image Enhancement with Lookup Tables and Diffusion Priors. *arXiv preprint arXiv:2409.18899*.

Lin, Y.; Ye, T.; Chen, S.; Fu, Z.; Wang, Y.; Chai, W.; Xing, Z.; Zhu, L.; and Ding, X. 2024b. AGLLDiff: Guiding Diffusion Models Towards Unsupervised Training-free Real-world Low-light Image Enhancement. *arXiv preprint arXiv:2407.14900*.

Liu, J.; Wang, Q.; Fan, H.; Wang, Y.; Tang, Y.; and Qu, L. 2023. Residual denoising diffusion models. *arXiv preprint arXiv:2308.13712*.

Liu, X.; Wang, Y.; and Liu, Q. 2018. Psgan: A Generative Adversarial Network for Remote Sensing Image Pan-Sharpening. In *2018 25th IEEE International Conference on Image Processing (ICIP)*.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. RePaint: Inpainting using Denoising Diffusion Probabilistic Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Luo, Z.; Gustafsson, F.; Zhao, Z.; Sjölund, J.; and Schön, T. 2023. (SR3)Image Restoration with Mean-Reverting Stochastic Differential Equations.

Ma, J.; Yu, W.; Chen, C.; Liang, P.; Guo, X.; and Jiang, J. 2020. Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion. *Information Fusion*, 110–120.

Mallat, S. G. 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7): 674–693.

Ren, M.; Delbracio, M.; Talebi, H.; Gerig, G.; and Milanfar, P. 2022. Multiscale Structure Guided Diffusion for Image Deblurring. *Cornell University - arXiv,Cornell University - arXiv*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Lecture Notes in Computer Science,Lecture Notes in Computer Science*.

Shah, V. P.; Younan, N. H.; and King, R. L. 2008. An Efficient Pan-Sharpening Method via a Combined Adaptive PCA Approach and Contourlets. *IEEE Transactions on Geoscience and Remote Sensing*, 1323–1335.

Song, J.; Meng, C.; and Ermon, S. 2020. (DDIM)Denoising Diffusion Implicit Models. *arXiv: Learning,arXiv: Learning*.

Tian, X.; Chen, Y.; Yang, C.; Gao, X.; and Ma, J. 2020. A variational pansharpening method based on gradient sparse representation. *IEEE Signal Processing Letters*, 27: 1180–1184.

Tian, X.; Chen, Y.; Yang, C.; and Ma, J. 2021. Variational pansharpening by exploiting cartoon-texture similarities. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–16.

Vivone, G.; Alparone, L.; Chanussot, J.; Dalla Mura, M.; Garzelli, A.; Licciardi, G. A.; Restaino, R.; and Wald, L. 2014. A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5): 2565–2586.

Wald, L.; Ranchin, T.; and Mangolini, M. 1997. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images.

Wang, H.; Gong, M.; Mei, X.; Zhang, H.; and Ma, J. 2024a. Deep Unfolded Network with Intrinsic Supervision for Pan-Sharpening. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5419–5426.

Wang, Y.; He, X.; Dong, Y.; Lin, Y.; Huang, Y.; and Ding, X. 2024b. Cross-Modality Interaction Network for Pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*.

Wang, Y.; Lin, Y.; Meng, G.; Fu, Z.; Dong, Y.; Fan, L.; Yu, H.; Ding, X.; and Huang, Y. 2023. Learning High-frequency Feature Enhancement and Alignment for Pan-sharpening. In *Proceedings of the 31st ACM International Conference on Multimedia*, 358–367.

Xia, J.; Yang, Z.; Li, S.; Zhang, S.; Fu, Y.; Gündüz, D.; and Li, X. 2024. Blind Super-Resolution Via Meta-Learning and Markov Chain Monte Carlo Simulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xu, Q.; Li, Y.; Nie, J.; Liu, Q.; and Guo, M. 2023. UPanGAN: Unsupervised pansharpening based on the spectral and spatial loss constrained Generative Adversarial Network. *Information Fusion*, 31–46.

Zhou, H.; Liu, Q.; Weng, D.; and Wang, Y. 2022a. Unsupervised Cycle-Consistent Generative Adversarial Networks for Pan Sharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 1–14.

Zhou, M.; Huang, J.; Fang, Y.; Fu, X.; and Liu, A. 2022b. Pan-Sharpening with Customized Transformer and Invertible Neural Network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 3553–3561.

Zhou, M.; Yan, K.; Fu, X.; Liu, A.; and Xie, C. 2023. PAN-Guided Band-Aware Multi-Spectral Feature Enhancement for Pan-Sharpening. *IEEE Transactions on Computational Imaging*, 9: 238–249.