Complex-Valued Residual Diffusion with GRPO for Pansharpening

Zhiyuan Wang , Dong Li*, Kaixin Fu , Chunhui Luo , Xueyang Fu*
University of Science and Technology of China
{ustcwangzhiyuan, dongli6, fkxhh, luochunhui}@mail.ustc.edu.cn, xyfu@ustc.edu.cn

Abstract

Pansharpening aims to fuse a high-resolution panchromatic (PAN) image with a low-resolution multispectral (MS) image to generate a high-resolution multispectral output that preserves both fine spatial details and faithful spectral responses. However, enhancing spatial textures without introducing spectral distortion remains challenging due to the inherent spectral–spatial trade-off. To address this issue, we propose a two-stage pansharpening framework that tackles the problem from both modeling and optimization perspectives. In the first stage, we formulate pansharpening as spectral-prior conditioned residual diffusion, where a stable spectral base constrains the generation process and allows the diffusion model to focus on PAN-guided high-frequency details, leading to improved training stability and reduced spectral drift. To better capture the coupled spatial and frequency characteristics of PAN–MS fusion, we adopt a complex-valued denoising network to enhance spectral–spatial interaction modeling. In the second stage, to bridge the gap between distortion-oriented training objectives and practical quality preferences, we introduce Group Relative Policy Optimization (GRPO) to fine-tune the diffusion model using multi-objective preference signals, explicitly balancing spectral fidelity, texture sharpness, and perceptual quality. Extensive experiments on standard benchmarks demonstrate that the proposed method achieves a more favorable trade-off between fidelity and perceptual quality compared to competitive end-to-end and diffusion-based approaches.

1 Introduction

Pansharpening aims to fuse a high-resolution panchromatic (PAN) image with a low-resolution multispectral (LRMS) image to generate a high-resolution multispectral (HRMS) product that retains both fine spatial details and accurate spectral responses. The intrinsic difficulty stems from the spectral–spatial tension: LRMS largely determines the low-frequency spectral content, whereas PAN provides abundant high-frequency structural cues that should be consistently injected into all spectral bands without introducing spectral distortions. Consequently, despite substantial advances with convolutional and transformer-based architectures [Masi *et al.*, 2016; Yang *et al.*, 2017; Xu *et al.*, 2021; Wu *et al.*, 2024], it remains challenging to achieve a robust balance between spatial sharpness and spectral consistency across diverse scenes and sensors.

Diffusion models have recently shown strong capabilities in image restoration and conditional generation [Ho *et al.*, 2020; Song *et al.*, 2020; Saharia *et al.*, 2022; Kawar *et al.*, 2022], motivating their adoption for pansharpening. However, existing paradigms reveal a practical trade-off. Fig. 1 provides direct evidence: end-to-end CNN/Transformer approaches often achieve higher distortion fidelity (e.g., higher PSNR) but may yield less pleasing perceptual results, while diffusion-based methods tend to improve perceptual quality yet frequently sacrifice pixel-level fidelity. The contrasting behavior between PSNR (fidelity-oriented) and LPIPS (perceptual similarity) [Zhang *et al.*, 2018] indicates that optimizing pansharpening with a single training objective is often insufficient to satisfy both fidelity- and perception-oriented criteria simultaneously.

Applying diffusion directly to full HRMS reconstruction further encounters two bottlenecks. First, diffusion training is typically driven by pixel-wise objectives (e.g., L1/L2 on noise or residual), which are not well aligned with how pansharpening quality is judged in practice—a combination of spectral consistency, structure/texture fidelity, and perceptual realism [Wald *et al.*, 1997; Alparone *et al.*, 2004]. Such losses may average out high-frequency details and remain insensitive to perceptual preferences. Second, learning diffusion on full HRMS signals can be unstable and sample-inefficient, since the model must simultaneously learn low-frequency multi-band spectra and high-frequency PAN-guided details, potentially leading to slow convergence and spectral artifacts. Meanwhile, recent evidence suggests that explicit spatial–frequency interaction is important for robust PAN–MS fusion [Tan *et al.*, 2024; He *et al.*, 2024], calling for denoisers that can model both domains more directly.

To address these challenges, we propose a two-stage framework that separates stable reconstruction from prefer-

*Corresponding authors.

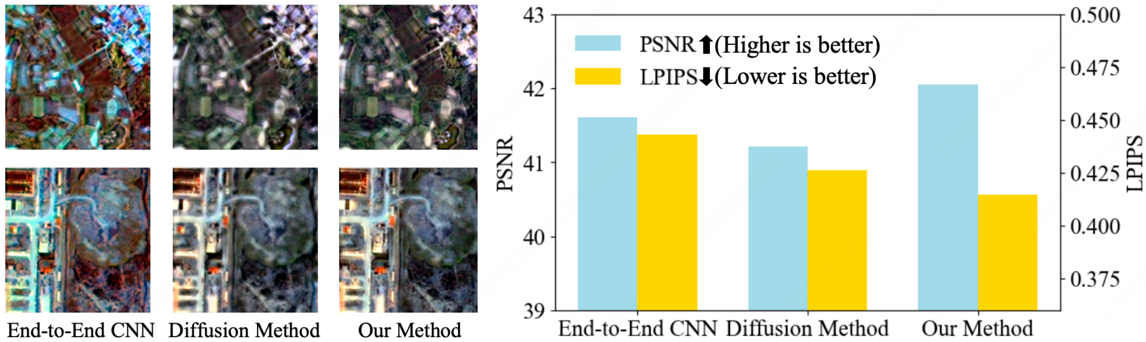

Figure 1: Motivation and trade-off between fidelity and perception in pansharpening. End-to-end CNN/Transformer methods often achieve higher PSNR (higher is better) but inferior perceptual quality (higher LPIPS), while diffusion-based methods tend to reduce LPIPS (lower is better) at the cost of PSNR. Our method achieves a more balanced trade-off across two examples.

ence alignment. In Stage I, we construct a strong spectral base via a frozen pre-enhancement network and formulate pansharpening as residual diffusion conditioned on this spectral prior. Instead of denoising the full HRMS, the diffusion model learns only the residual between the prior and the HRMS target, which simplifies learning, stabilizes training, and reduces the tendency to hallucinate or drift the spectral base, while still focusing on injecting PAN-guided high-frequency structures through the residual. For the denoising backbone, we instantiate a Complex U-Net built upon complex-valued convolutions [Ronneberger *et al.*, 2015; Trabelsi *et al.*, 2018] and a spatial–frequency coupled block design. By jointly manipulating real and imaginary components, the model can capture amplitude–phase couplings and richer spectral–spatial interactions [Luo *et al.*, 2025]. Moreover, PAN/LRMS cues (together with the spectral prior) are injected in a persistent multi-resolution manner, ensuring that structural guidance remains available throughout the iterative denoising trajectory.

In the second stage, we introduce Group Relative Policy Optimization (GRPO) [Shao *et al.*, 2024] to fine-tune the conditional diffusion model toward task-relevant preferences. For each input, we generate a small group of candidate HRMS outputs via DDIM sampling and evaluate them with a composite reward that explicitly captures (i) spectral fidelity, (ii) PAN-aligned texture/edge consistency, and (iii) perceptual quality. We then compute group-relative advantages and optimize a softmax-weighted objective, yielding a stable preference-aligned learning signal [Schulman *et al.*, 2017]. In contrast to purely likelihood- or pixel-driven training, GRPO directly aligns diffusion updates with task-relevant quality preferences, improving the spectral–spatial trade-off in a controlled manner. To the best of our knowledge, this is the first work that introduces GRPO into the pansharpening domain.

The contributions of this paper are summarized as follows:

- We propose a spectral-prior conditioned residual diffusion framework for pansharpening and employ a **complex-valued denoising network** to strengthen spectral–spatial interaction modeling.
- We introduce Group Relative Policy Optimization into pansharpening for the first time to align diffusion training with multi-objective quality preferences.
- Extensive experiments on standard benchmarks demonstrate that the proposed method achieves a more balanced trade-off between spectral fidelity and perceptual quality than existing end-to-end and diffusion-based approaches.

2 Related Work

2.1 Pansharpening

Pansharpening aims to fuse a high-resolution panchromatic (PAN) image with a low-resolution multispectral (MS) image to obtain a high-resolution MS product. Traditional approaches are typically categorized into component-substitution [Gillespie *et al.*, 1987; Haydn *et al.*, 1982; Laben and Brower, 2000] and multiresolution-analysis [Vivone *et al.*, 2015; Aiazzi *et al.*, 2006] families, which are efficient but often suffer from spectral distortion under complex cross-sensor characteristics. Recently, deep learning has become the dominant paradigm by learning nonlinear mappings from paired data. Early CNN-based methods demonstrated that end-to-end training can significantly reduce artifacts and improve spectral fidelity compared with handcrafted fusion rules [Masi *et al.*, 2016]. More recently, hypercomplex and complex-valued representation learning has been explored to better capture spectral–spatial and spatial–frequency interactions between PAN and MS; for example, quaternion-based modeling strengthens spatial–spectral interaction [Li *et al.*, 2025a], while PanComplex builds a complex-valued pansharpening framework and reports strong performance with compact computation [Luo *et al.*, 2025].

2.2 Diffusion Models for Image Generation and Restoration

Denoising diffusion models have emerged as powerful generative priors by learning a gradual denoising process [Ho *et al.*, 2020]. To accelerate sampling, implicit formulations such as DDIM reduce the number of denoising steps while maintaining generation quality [Song *et al.*, 2020; Lu *et al.*, 2025]. Beyond synthesis, diffusion has been widely studied for inverse problems and image restoration, where the learned prior provides robustness under noise and ill-posed degradation [Saharia *et al.*, 2022; Kawar *et al.*, 2022; Chen *et al.*, 2026]. These advances motivate diffusion-based formulations for remote sensing fusion, as they can model complex conditional distributions and produce visually plausible details under strong spectral constraints.

2.3 Complex-Valued Learning

Complex-valued neural networks (CVNNs) have been investigated as a principled way to jointly model amplitude and phase, offering richer representations for frequency-related patterns [Trabelsi *et al.*, 2018]. In remote sensing fusion, complex-valued backbones are particularly appealing because the PAN image carries abundant high-frequency spatial cues while MS provides spectral content. PanComplex further validates the effectiveness of complex-valued convolution for pansharpening by explicitly leveraging complex-domain interactions [Luo *et al.*, 2025].

2.4 Preference Optimization with Perceptual Rewards

Policy-gradient methods optimize objectives defined by task rewards, with REINFORCE providing an unbiased gradient estimator [Williams, 1992] and PPO improving stability through clipped updates [Schulman *et al.*, 2017]. In parallel, preference optimization has recently gained popularity for aligning generative models to non-differentiable objectives without requiring explicit reward modeling in the deployment loop [Rafailov *et al.*, 2023]. For low-level vision, perceptual criteria are frequently used to complement pixel-wise losses: perceptual losses based on deep features [Johnson *et al.*, 2016] and GAN-based super-resolution [Ledig *et al.*, 2017] highlight the importance of human-aligned quality. Modern perceptual metrics such as LPIPS [Zhang *et al.*, 2018], DISTS [Ding *et al.*, 2020], and MS-SSIM [Wang *et al.*, 2003] are therefore commonly adopted as objective surrogates for visual fidelity and structural consistency. These lines of work suggest that integrating preference optimization with perceptual and domain-specific rewards is a promising direction for high-quality pansharpening.

3 Method

3.1 Framework

Pansharpening takes a low-resolution multispectral image $M \in R^{\frac{H}{s} \times \frac{W}{s} \times C}$ and a high-resolution panchromatic image $P \in R^{H \times W}$ as input, and predicts a high-resolution multispectral image $\hat{Y} \in R^{H \times W \times C}$. The intrinsic difficulty lies in the spectral–spatial tension: LRMS mainly determines low-frequency spectral content, whereas PAN provides the high-frequency structural cues that should be injected into *all* spectral bands without introducing spectral distortion. This tension explains why a single objective often leads to an undesirable compromise between fidelity and perceptual sharpness.

We address this challenge with a two-stage design (Fig. 2) that turns the trade-off into a *structured optimization problem*. Stage I focuses on *stability and correctness*: we first construct a strong spectral base and then train diffusion to model only the missing residual, so the generative model does not need to learn low-frequency spectra and high-frequency textures simultaneously. Stage II focuses on *preference alignment*: we further fine-tune the diffusion model via GRPO using a multi-objective reward that directly reflects what pansharpening should optimize in practice—spectral fidelity, PAN-aligned textures, and perceptual realism. This division of labor yields a coherent loop: Stage I ensures a reliable baseline reconstruction anchored by spectral prior, and Stage II improves the spectral–spatial trade-off in a controlled and explainable manner.

3.2 Spectral Prior Construction

A practical bottleneck of applying diffusion directly to HRMS reconstruction is that the model must simultaneously learn (i) low-frequency multi-band spectral content and (ii) high-frequency spatial details guided by PAN. In pansharpening, the low-frequency spectral component is relatively well-determined by LRMS, while the ill-posed part is how to inject PAN textures into each band without breaking spectral consistency. To decouple these factors, we first build a spectral prior $\hat{M}_0$ from LRMS:

$$\hat{M}_0 = F_\phi(M), \tag{1}$$

where F_ϕ is a pre-enhancement network and is **frozen** throughout Stage I and Stage II. We instantiate F_ϕ using the spatial–frequency domain information integration (SFDI) design [Zhou *et al.*, 2022a], which leverages complementary cues across domains to produce an HRMS-like estimate with strong spectral consistency. Freezing F_ϕ is crucial: it turns $\hat{M}_0$ into a stable spectral anchor, preventing diffusion from "re-learning" or drifting the spectral base during training. As a result, the subsequent generative modeling becomes substantially more stable and sample-efficient, and it empirically reduces spectral artifacts caused by over-injecting PAN structures.

3.3 Residual Diffusion Process

Given the spectral prior, we formulate pansharpening as a residual generation problem:

$$r_0 = Y - \hat{M}_0, \qquad \hat{Y} = \hat{M}_0 + \hat{r}_0. \tag{2}$$

This choice is not merely a numerical trick; it reflects the underlying fusion logic. The prior $\hat{M}_0$ already explains most low-frequency spectral information, so the residual r_0 concentrates on (i) PAN-guided spatial details and (ii) subtle band-wise corrections needed for spectral consistency. Consequently, diffusion is trained on a cleaner and more structured target, which reduces low-frequency ambiguity and makes optimization less prone to over-smoothing or spectral drift.

We adopt the standard DDPM forward process [Ho *et al.*, 2020] for residual diffusion:

$$q(r_t \mid r_{t-1}) = \mathcal{N}\Big(r_t; \sqrt{1-\beta_t}\, r_{t-1}, \beta_t \mathbf{I}\Big), \quad t = 1, \ldots, T, \tag{3}$$

with $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, yielding

$$q(r_t \mid r_0) = \mathcal{N}\big(r_t; \sqrt{\bar{\alpha}_t}\, r_0, (1-\bar{\alpha}_t)\mathbf{I}\big). \tag{4}$$

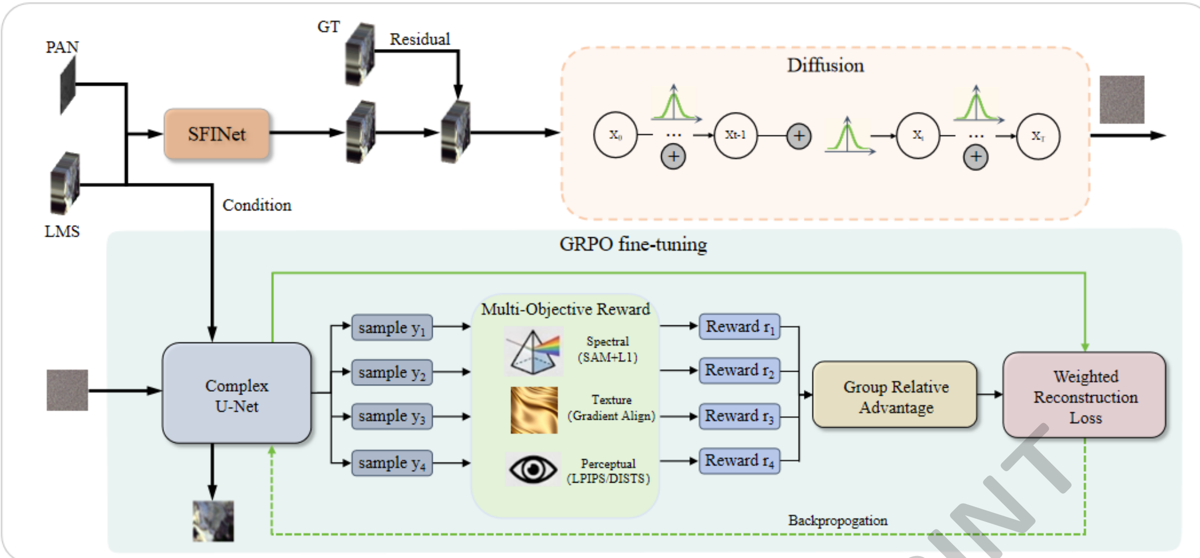

Figure 2: Overview of our two-stage framework. Stage I learns a spectral-prior residual diffusion model with a Complex U-Net denoiser. Stage II performs GRPO fine-tuning: we sample a group of candidate reconstructions, score them with a multi-objective reward, and reweight learning signals to explicitly improve the spectral–spatial trade-off.

The denoiser ϵ_θ is trained to predict the injected noise from a noisy residual r_t under conditional guidance, using the standard noise-prediction objective:

$$\mathcal{L}_{\text{diff}}(\theta) = E_{r_0,t,\epsilon}\Big[\big\|\epsilon - \epsilon_\theta(r_t, c, t)\big\|_2^2\Big]. \tag{5}$$

For conditional guidance, we use PAN and LRMS cues (together with the spectral prior) to steer the denoising trajectory; this provides the denoiser with consistent structural information from PAN and band-wise context from LRMS, without requiring additional handcrafted transforms. At inference, we use DDIM sampling [Song *et al.*, 2020] to obtain $\hat{r}_0$ efficiently and reconstruct $\hat{Y} = \hat{M}_0 + \hat{r}_0$.

3.4 Complex U-Net with Spatial–Frequency Interaction

A key question is how to design a denoiser that can repeatedly inject PAN structures during iterative denoising while maintaining cross-band spectral consistency. Conventional real-valued U-Nets mainly rely on spatial convolutions; they are effective for local edges but can struggle to model global frequency patterns (e.g., periodic textures and long-range correlations) and may overfit to PAN gradients, resulting in spectral distortion. Meanwhile, the PanComplex study validates that complex-valued convolutions can be beneficial for pansharpening by better exploiting amplitude–phase information in feature representations. Motivated by this evidence, we design the diffusion denoiser as a **Complex U-Net** (Fig. 3) that combines complex-valued feature learning with explicit spatial–frequency interaction inside each block.

Formally, a complex feature map is represented as $x = x^{(r)} + ix^{(i)}$. Given a complex kernel $W = W^{(r)} + iW^{(i)}$, complex convolution is computed by

$$\begin{aligned} W * x = {} & (W^{(r)} * x^{(r)} - W^{(i)} * x^{(i)}) \\ & + \, i\,(W^{(r)} * x^{(i)} + W^{(i)} * x^{(r)}). \end{aligned} \tag{6}$$

This operation enables the network to jointly manipulate real and imaginary components, offering an expressive mechanism to encode amplitude–phase couplings that are difficult to capture with purely real-valued filters [Trabelsi *et al.*, 2018].

Beyond the representation, the second key design is *where and how* we inject PAN/LRMS cues. As illustrated in Fig. 3, we inject these modalities at every scale of the encoder (and symmetrically propagate them through skip connections), rather than conditioning only once at the input. Let f_ℓ denote the feature at scale ℓ. We downsample PAN and LRMS (and the spectral prior) to the same resolution and concatenate them with the current feature:

$$z_\ell = \text{Concat}\big(f_\ell,\ \text{Down}_\ell(P),\ \text{Down}_\ell(M)\big). \tag{7}$$

This persistent, multi-resolution injection is important for diffusion: since denoising is iterative, high-frequency guidance should remain available throughout the trajectory instead of being diluted after early layers.

The Complex Block (Fig. 3, right) then realizes explicit spatial–frequency interaction. The concatenated feature is split into modality-aware components and processed by two coupled branches: a spatial branch (complex convolutions) that captures local edge-aligned structures, and a frequency branch that transforms features via DFT/IDFT to model global interactions and long-range texture patterns [Xu *et al.*, 2024; Li *et al.*, 2024a; Li *et al.*, 2025b]. We denote the frequency branch as

$$F_\ell^{\text{freq}} = \mathcal{F}^{-1}\Big(\mathcal{G}_{\text{freq}}\big(\mathcal{F}(z_\ell^{\text{ms}}),\ \mathcal{F}(z_\ell^{\text{pan}})\big)\Big), \tag{8}$$

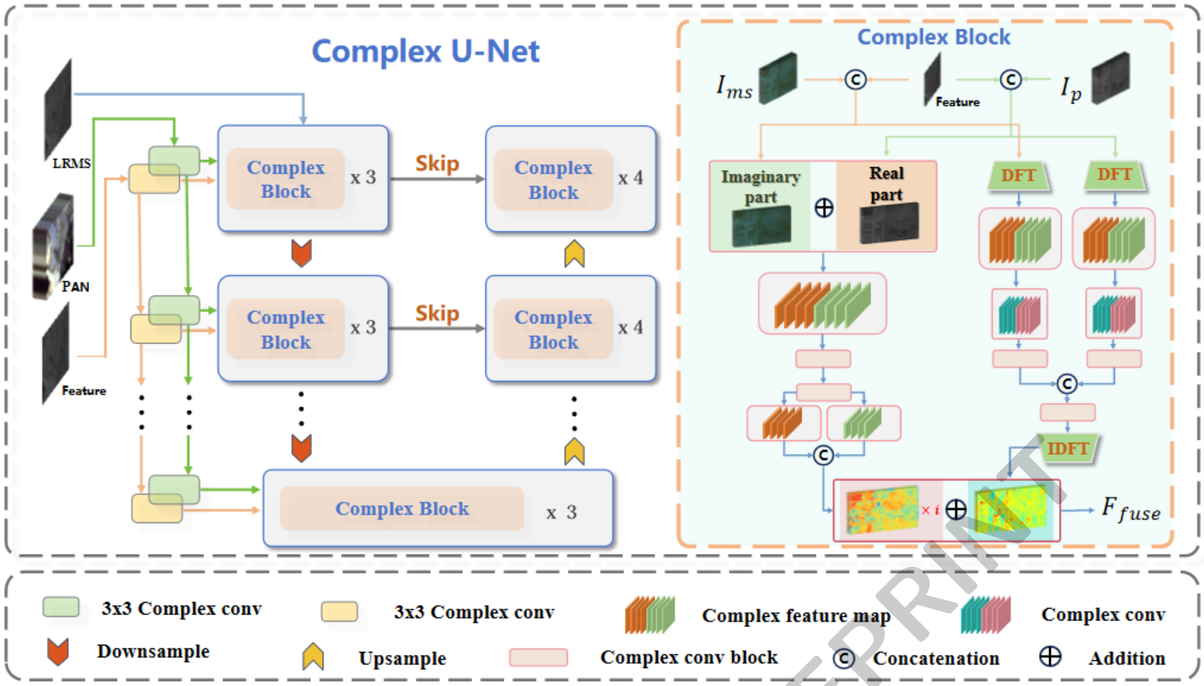

Figure 3: Complex U-Net denoiser and Complex Block. PAN/LRMS cues are injected at each scale. The Complex Block couples a spatial branch and a Fourier branch to enable explicit spatial–frequency interaction, producing features that are simultaneously texture-aware and spectrally coherent.

where $\mathcal{F}(\cdot)$ and $\mathcal{F}^{-1}(\cdot)$ are DFT/IDFT, and z_ℓ^{ms}, z_ℓ^{pan} follow Fig. 3. Finally, we fuse the spatial and frequency outputs to obtain a feature that is simultaneously texture-rich and spectrally coherent:

$$F_\ell^{\text{fuse}} = F_\ell^{\text{spa}} + F_\ell^{\text{freq}}. \tag{9}$$

By repeating this design across scales, the denoiser can (i) inject PAN details consistently at multiple resolutions, and (ii) correct spectral distortions through cross-band interactions that are explicitly supported by complex-valued and frequency-aware processing.

3.5 GRPO Fine-tuning with Multi-Objective Rewards

Stage I provides a stable residual diffusion backbone, but its learning signal is still dominated by pixel-level surrogates (noise prediction), which do not fully reflect how pansharpening quality is judged. In practice, good pansharpening simultaneously requires: (1) spectral consistency with the HRMS target, (2) faithful transfer of PAN textures/edges, and (3) perceptual naturalness. Optimizing only a single surrogate tends to over-emphasize one aspect and hurt the others, which underlies the persistent PSNR–LPIPS trade-off.

We introduce GRPO fine-tuning (Fig. 2) to explicitly align diffusion updates with these multi-faceted preferences. As illustrated in the lower part of Fig. 2, for each training input the denoiser generates K candidate outputs; these candidates are independently scored by three reward functions (spectral, texture, perceptual), and their normalized scores are converted to softmax weights that reweight the reconstruction loss—higher-reward candidates receive larger gradient contributions, steering the model toward outputs that jointly satisfy all quality criteria.

For each training input, we draw a small group of K candidate reconstructions $\{\hat{Y}^{(k)}\}_{k=1}^{K}$ by DDIM sampling with different random seeds. We then assign each candidate a composite reward:

$$R^{(k)} = \omega_{\text{spec}} R_{\text{spec}}^{(k)} + \omega_{\text{tex}} R_{\text{tex}}^{(k)} + \omega_{\text{perc}} R_{\text{perc}}^{(k)}. \tag{10}$$

The spectral term R_{spec} emphasizes band-wise fidelity and spectral consistency; in practice we combine complementary criteria such as SAM and absolute deviation, so the model does not gain PSNR by introducing spectral angle distortions. The texture term R_{tex} enforces structural agreement with PAN; motivated by edge- and noise-aware cues in low-level vision and image forensics [Li *et al.*, 2023; Zhu *et al.*, 2024; Li *et al.*, 2024b], we implement it by measuring edge/gradient alignment between the reconstruction (aggregated across bands) and PAN, encouraging the model to transfer textures while avoiding spurious edges. The perceptual term R_{perc} measures human-aligned similarity (e.g., LPIPS/DISTS) [Zhang *et al.*, 2018; Ding *et al.*, 2020], which is crucial for discouraging over-smoothed diffusion outputs and promoting visually pleasing details. Together, these terms turn the trade-off into an explicit objective rather than an implicit side-effect.

To keep optimization stable and reduce variance, GRPO uses group-relative weighting instead of trusting only the best sample. We normalize rewards within each group and convert

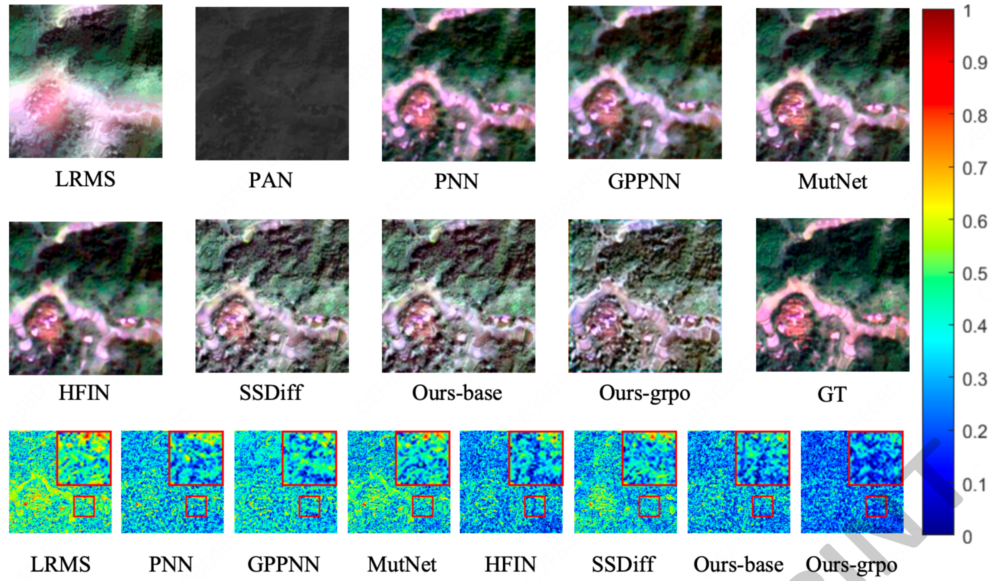

Figure 4: The visual comparisons between other methods and our method on Gaofen-2 satellite

them to softmax weights:

$$w^{(k)} = \frac{\exp(\tilde{R}^{(k)}/\tau)}{\sum_{j=1}^{K}\exp(\tilde{R}^{(j)}/\tau)}, \qquad \tilde{R}^{(k)} = \frac{R^{(k)} - \mu_R}{\sigma_R + \epsilon}. \tag{11}$$

This mechanism emphasizes higher-preference candidates while still leveraging information from all samples, preventing unstable updates caused by occasional outliers. Conceptually, Stage I provides a strong and stable baseline anchored by the spectral prior, and Stage II refines the denoiser toward reconstructions that better satisfy spectral, texture, and perceptual preferences; together they yield a principled improvement of the spectral–spatial trade-off.

4 Experiments

4.1 Datasets and Benchmarks

We conduct experiments on three widely used pansharpening benchmarks: WorldView-II (WV2), WorldView-III (WV3), and Gaofen-2 (GF2). Following common practice, we construct training/validation/test splits under the Wald protocol [Wald *et al.*, 1997], where LRMS and PAN pairs are generated in a reduced-resolution (RR) manner to provide ground-truth HRMS for supervised evaluation. For full-resolution (FR) evaluation without ground-truth HRMS, we report no-reference fusion metrics including D_λ, D_s, and QNR.

Baselines. To validate both reconstruction fidelity and perceptual quality, we compare against a compact yet representative set of baselines that are also used in our qualitative figure (Fig. 4). Specifically, we include classical end-to-end pansharpening networks PNN [Masi *et al.*, 2016], GPPNN [Xu *et al.*, 2021], and MutNet [Zhou *et al.*, 2022b], as well as recent methods that emphasize stronger structure guidance such as HFIN [Tan *et al.*, 2024]. We also include a diffusion-based baseline SSDiff [Zhong *et al.*, 2024] as a representative diffusion paradigm for pansharpening. In addition, we report three traditional component-substitution baselines (Brovey [Gillespie *et al.*, 1987], IHS [Haydn *et al.*, 1982], and GS [Laben and Brower, 2000]) to contextualize the gain brought by learning-based fusion. Unless otherwise specified, all results are averaged over the corresponding test sets.

4.2 Implementation Details

All models are implemented in PyTorch and trained on a single NVIDIA GeForce RTX 4090 GPU. The multispectral images are pre-cropped into 128×128 patches with a scale factor of 4. We use the AdamW optimizer with a learning rate of 1×10^{-4} and batch size of 4. The diffusion process uses $T{=}500$ timesteps with a cosine noise schedule; 25-step DDIM is adopted at inference. For GRPO, $K{=}4$ candidates are sampled per input with reward weights $\omega_{\text{spec}}{=}0.5$, $\omega_{\text{tex}}{=}0.3$, $\omega_{\text{perc}}{=}0.2$.

4.3 Comparison with State-of-the-art Methods

Evaluation on reduced-resolution scenes. Table 1 reports quantitative results on WV2/WV3/GF2 under the Wald protocol. We highlight the best results in **bold** and the second-best in underline. Across all three benchmarks, our approach achieves a consistent improvement pattern: it increases distortion-fidelity metrics (PSNR/SSIM) and reduces spectral distortion (SAM/ERGAS). Together with the qualitative results and the ablation study on perceptual metrics, this indicates that the proposed two-stage training improves the fidelity–perception trade-off.

A closer look reveals a clear division among baselines. End-to-end methods such as PNN and GPPNN typically deliver competitive PSNR/SSIM but may struggle to preserve subtle PAN-aligned textures without introducing spectral bias. Conversely, diffusion-based methods (SSDiff) tend to yield visually more pleasing outputs and reduce perceptual errors, yet they often sacrifice pixel-level fidelity, which is reflected by a noticeable PSNR drop. Our method bridges

Methods	Params (M)	WorldView-II				WorldView-III				Gaofen-2			
		PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
Brovey	–	35.8646	0.9216	0.0403	1.8238	21.8212	0.5457	0.1208	8.9730	36.9060	0.8882	0.0318	1.7398
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
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
PNN	0.0689	40.7550	0.9624	0.0259	1.0646	29.9418	0.9121	0.0824	3.3206	43.1208	0.9704	0.0172	0.8528
GPPNN	0.1198	41.1622	0.9684	0.0244	1.0315	30.1785	0.9175	0.0776	3.2593	44.2145	0.9815	0.0137	0.7361
MutNet	0.1588	41.6773	0.9705	0.0224	0.9519	30.4907	0.9223	0.0749	3.1125	47.3042	0.9892	0.0102	0.5481
HFIN	0.0772	42.2319	0.9714	0.0215	0.8807	30.6147	0.9203	0.0742	3.0786	48.8783	0.9898	0.0093	0.4591
SSDiff	0.3643	42.2071	<u>0.9722</u>	<u>0.0206</u>	0.9015	30.9162	0.9265	0.0714	3.0271	48.9251	<u>0.9931</u>	0.0104	0.4835
Ours w/o GRPO	0.2947	<u>42.2497</u>	0.9693	0.0221	<u>0.8796</u>	<u>31.1480</u>	<u>0.9290</u>	<u>0.0692</u>	<u>2.9842</u>	<u>49.0749</u>	0.9902	<u>0.0086</u>	<u>0.4516</u>
Ours	0.2947	**42.4372**	**0.9746**	**0.0198**	**0.8541**	**31.2341**	**0.9347**	**0.0638**	**2.8354**	**49.1687**	**0.9967**	**0.0081**	**0.4351**

Table 1: Quantitative comparisons on reduced-resolution benchmarks under the Wald protocol.

this gap by separating responsibilities across stages: Stage I provides a strong fidelity-oriented foundation, while Stage II (GRPO fine-tuning) explicitly aligns optimization with multi-objective quality preferences. As evidenced by the comparison between *Ours* and *Ours without GRPO*, GRPO consistently improves perceptual metrics while keeping fidelity metrics at a strong level, validating the motivation introduced in Sec. 1.

	PNN	GPPNN	MutNet	HFIN	SSDiff	Ours without GRPO	Ours
$D_\lambda\downarrow$	0.0746	0.0782	0.0694	0.0710	0.0683	<u>0.0659</u>	**0.0594**
$D_s\downarrow$	0.1164	0.1253	0.1118	0.1098	<u>0.1067</u>	0.1118	**0.1049**
QNR↑	0.8191	0.8073	0.8247	0.8261	0.8258	<u>0.8283</u>	**0.8438**

Table 2: No-reference metrics on full-resolution data.

Qualitative evaluation. Figure 4 presents visual comparisons on GF2 with the same set of methods as in Table 1. The first two rows show RGB renderings for intuitive inspection, while the last row visualizes per-pixel error maps w.r.t. GT using the same color scale. Several consistent phenomena can be observed. (i) *End-to-end baselines* (PNN/GPPNN) often produce locally over-smoothed textures or slightly biased colors in challenging regions (e.g., thin edges, small structures), indicating that optimizing with a single distortion-oriented objective may not fully capture perceptual preferences. (ii) The *diffusion baseline* (SSDiff) improves perceptual appearance in some regions, yet it may weaken fine structures or deviate in pixel-level fidelity, which is consistent with the quantitative trade-off seen in Table 1. (iii) Our reconstruction preserves sharper PAN-aligned structures while maintaining better spectral consistency, and the error maps show smaller discrepancies overall, suggesting that our method improves both structural transfer and spectral preservation.

Moreover, comparing *Ours* against *Ours without GRPO* isolates the effect of Stage II: GRPO fine-tuning yields cleaner local details and reduces perceptual errors in textured areas without introducing noticeable color shifts, supporting our claim that preference-aligned optimization is crucial for narrowing the fidelity–perception gap in diffusion-based pansharpening.

Evaluation on full-resolution scenes. Since real full-resolution scenes lack ground-truth HRMS, we further report no-reference metrics on full-resolution data. Table 2 shows that our method achieves the best D_λ, D_s, and QNR, indicating strong generalization in practical scenarios.

4.4 Ablation Experiments

We conduct ablation studies on WV2 to verify the contribution of each key component in our framework. Specifically, we examine: (i) the spectral prior (replacing it with a trainable prior or removing it), (ii) residual diffusion (directly diffusing HRMS), (iii) the Complex U-Net design (replacing complex-valued blocks with real-valued ones), and (iv) GRPO fine-tuning (removing preference alignment). Table 3 summarizes the results. Removing any component degrades performance, confirming that the proposed design is not a simple combination of tricks: the frozen spectral prior and residual diffusion stabilize training and suppress spectral artifacts, the Complex U-Net strengthens spectral–spatial interaction, and GRPO further improves the fidelity–perception trade-off.

Configuration	PSNR↑	SSIM↑	SAM↓	ERGAS↓	LPIPS↓	DISTS↓
without spectral prior (diffuse Y)	41.8327	0.9603	0.0249	0.9879	0.4528	0.1743
without residual (direct HRMS)	41.9022	0.9647	0.0238	0.9725	0.4456	0.1713
Real U-Net (no complex blocks)	42.2357	<u>0.9713</u>	<u>0.0214</u>	0.9003	0.4233	0.1669
without GRPO (Stage I only)	<u>42.2497</u>	0.9693	0.0221	<u>0.8796</u>	<u>0.4097</u>	<u>0.1610</u>
Ours (full)	**42.4372**	**0.9746**	**0.0198**	**0.8541**	**0.3894**	**0.1558**

Table 3: Ablation results on WV2 (Wald protocol). The best and the second best values are highlighted in bold and underline, respectively.

5 Conclusion

We presented a two-stage pansharpening framework that improves the fidelity–perception trade-off by coupling spectral-prior conditioned residual diffusion with preference-aligned optimization. In Stage I, a frozen spectral prior anchors low-frequency spectral content and lets the diffusion model focus on PAN-guided high-frequency residuals, improving training stability and reducing spectral drift. To capture coupled spatial and frequency characteristics of PAN–MS fusion, we adopt a Complex U-Net denoiser with explicit spatial–frequency interaction. In Stage II, GRPO fine-tuning uses multi-objective rewards for spectral fidelity, PAN-aligned texture transfer, and perceptual quality, providing an alignment signal beyond pixel-wise surrogates. Experiments on standard benchmarks show consistent improvements over strong end-to-end and diffusion-based baselines, and ablations verify the contribution of each component. We believe this work provides a principled and practical direction for integrating diffusion modeling and preference optimization in future remote sensing image fusion tasks.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62422609, and in part by the Fundamental Research Funds for the Central Universities under Grant WK2102026001.

References

[Aiazzi *et al.*, 2006] Bruno Aiazzi, Luciano Alparone, Stefano Baronti, Andrea Garzelli, and Massimo Selva. MTF-tailored multiscale fusion of high-resolution MS and Pan imagery. *Photogrammetric Engineering & Remote Sensing*, 72(5):591–596, 2006.

[Alparone *et al.*, 2004] Luciano Alparone, Stefano Baronti, Andrea Garzelli, and Filippo Nencini. A global quality measurement of pan-sharpened multispectral imagery. *IEEE Geoscience and Remote Sensing Letters*, 1(4):313–317, 2004.

[Chen *et al.*, 2026] Wanxiang Chen, Dongshang Deng, Chaocan Xiang, Zhi Liu, and Bin Xiao. When one-shot federated learning meets diffusion models at the edge: Technological advances and applications. *Journal of Intelligent Computing and Networking*, 2(1):35–54, 2026.

[Ding *et al.*, 2020] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.

[Gillespie *et al.*, 1987] Alan R. Gillespie, Anne B. Kahle, and Robert E. Walker. Color enhancement of highly correlated images. II. channel ratio and "Chromaticity" transform techniques. *Remote Sensing of Environment*, 22(3):343–365, 1987.

[Haydn *et al.*, 1982] Roger Haydn, Gary W. Dalke, James Henkel, and John E. Bare. Application of the IHS color transform to the processing of multisensor data and image enhancement. In *Proceedings of the International Symposium on Remote Sensing of Environment*, 1982.

[He *et al.*, 2024] Xuanhua He, Keyu Yan, Rui Li, Chengjun Xie, Jie Zhang, and Man Zhou. Frequency-adaptive pansharpening with mixture of experts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2121–2129, 2024.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Johnson *et al.*, 2016] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.

[Kawar *et al.*, 2022] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in neural information processing systems*, 35:23593–23606, 2022.

[Laben and Brower, 2000] Craig A. Laben and Bernard V. Brower. Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening. U.S. Patent 6,011,875, 2000.

[Ledig *et al.*, 2017] Christian Ledig, Lucas Theis, Ferenc Huszár, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[Li *et al.*, 2023] Dong Li, Jiaying Zhu, Menglu Wang, Jiawei Liu, Xueyang Fu, and Zheng-Jun Zha. Edge-aware regional message passing controller for image forgery localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8222–8232, 2023.

[Li *et al.*, 2024a] Dong Li, Yidi Liu, Xueyang Fu, Senyan Xu, and Zheng-Jun Zha. Fouriermamba: Fourier learning integration with state space models for image deraining. *arXiv preprint arXiv:2405.19450*, 2024.

[Li *et al.*, 2024b] Dong Li, Jiaying Zhu, Xueyang Fu, Xun Guo, Yidi Liu, Gang Yang, Jiawei Liu, and Zheng-Jun Zha. Noise-assisted prompt learning for image forgery detection and localization. In *European Conference on Computer Vision*, pages 18–36. Springer, 2024.

[Li *et al.*, 2025a] Dong Li, Chunhui Luo, Yuanfei Bao, Gang Yang, Jie Xiao, Xueyang Fu, and Zheng-Jun Zha. Enhanced pansharpening via quaternion spatial-spectral interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10908–10918, 2025.

[Li *et al.*, 2025b] Dong Li, Jiaying Zhu, Yidi Liu, Xin Lu, Xueyang Fu, Jiawei Liu, Aiping Liu, and Zheng-Jun Zha. Learnable frequency decomposition for image forgery detection and localization. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 1359–1367, 2025.

[Lu *et al.*, 2025] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22(4):730–751, 2025.

[Luo *et al.*, 2025] Chunhui Luo, Dong Li, Xiaoliang Ma, Xin Lu, Zhiyuan Wang, Jiangtong Tan, and Xueyang Fu. PanComplex: Leveraging complex-valued neural networks for enhanced pansharpening. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)*, pages 1702–1710. International Joint Conferences on Artificial Intelligence Organization, 2025.

[Masi *et al.*, 2016] Giuseppe Masi, Davide Cozzolino, Luisa Verdoliva, and Giuseppe Scarpa. Pansharpening by convolutional neural networks. *Remote Sensing*, 8(7):594, 2016.

[Rafailov *et al.*, 2023] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[Saharia *et al.*, 2022] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

[Schulman *et al.*, 2017] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[Shao *et al.*, 2024] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[Song *et al.*, 2020] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[Tan *et al.*, 2024] Jiangtong Tan, Jie Huang, Naishan Zheng, Man Zhou, Keyu Yan, Danfeng Hong, and Feng Zhao. Revisiting spatial-frequency information integration from a hierarchical perspective for panchromatic and multispectral image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25922–25931, 2024.

[Trabelsi *et al.*, 2018] Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, João Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J. Pal. Deep complex networks. In *International Conference on Learning Representations*, 2018.

[Vivone *et al.*, 2015] Gemine Vivone, Luciano Alparone, Jocelyn Chanussot, Mauro Dalla Mura, Andrea Garzelli, Giorgio A. Licciardi, Rocco Restaino, and Lucien Wald. A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5):2565–2586, 2015.

[Wald *et al.*, 1997] Lucien Wald, Thierry Ranchin, and Marc Mangolini. Fusion of satellite images of different spatial resolutions: assessing the quality of resulting images. *Photogrammetric Engineering & Remote Sensing*, 63(6):691–699, 1997.

[Wang *et al.*, 2003] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar conference on signals, systems & computers, 2003*, volume 2, pages 1398–1402. IEEE, 2003.

[Williams, 1992] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992.

[Wu *et al.*, 2024] Gang Wu, Junjun Jiang, Kui Jiang, and Xianming Liu. Fully 1×1 convolutional network for lightweight image super-resolution. *Machine Intelligence Research*, 21(6):1062–1076, 2024.

[Xu *et al.*, 2021] Shuang Xu, Jiangshe Zhang, Zixiang Zhao, Kai Sun, Junmin Liu, and Chunxia Zhang. Deep gradient projection networks for pan-sharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1366–1375, 2021.

[Xu *et al.*, 2024] Gang Xu, Qibin Hou, and Ming-Ming Cheng. Dual frequency transformer for efficient SDR-to-HDR translation. *Machine Intelligence Research*, 21(3):538–548, 2024.

[Yang *et al.*, 2017] Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, and John Paisley. Pannet: A deep network architecture for pan-sharpening. In *Proceedings of the IEEE international conference on computer vision*, pages 5449–5457, 2017.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[Zhong *et al.*, 2024] Yu Zhong, Xiao Wu, Zihan Cao, Hong-Xia Dou, and Liang-Jian Deng. SSDiff: Spatial-spectral integrated diffusion model for remote sensing pansharpening. *Advances in Neural Information Processing Systems*, 37:77962–77986, 2024.

[Zhou *et al.*, 2022a] Man Zhou, Jie Huang, Keyu Yan, Hu Yu, Xueyang Fu, Aiping Liu, Xian Wei, and Feng Zhao. Spatial-frequency domain information integration for pan-sharpening, 2022. Preprint.

[Zhou *et al.*, 2022b] Man Zhou, Keyu Yan, Jie Huang, Zihe Yang, Xueyang Fu, and Feng Zhao. Mutual information-driven pan-sharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1798–1808, 2022.

[Zhu *et al.*, 2024] Jiaying Zhu, Dong Li, Xueyang Fu, Gang Yang, Jie Huang, Aiping Liu, and Zheng-Jun Zha. Learning discriminative noise guidance for image forgery detection and localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7739–7747, 2024.