FMGRPA: FLOW MATCHING WITH GROUP-RELATIVE PREFERENCE ALIGNMENT FOR PANSHARPENING

Anonymous authors
Paper under double-blind review

ABSTRACT

Pansharpening aims to fuse a low resolution multispectral (LRMS) image and a high resolution panchromatic (PAN) image to recover a high resolution multispectral (HRMS) image with spectral fidelity and spatial detail. Existing methods usually reconstruct paired HRMS images, but a single objective constrains only overall reference error, making it difficult to distinguish spectral shifts from insufficient spatial detail or express the desired spectral and spatial fusion preference among results with comparable reconstruction errors. To this end, we propose FMGRPA, a two stage pansharpening framework combining LRMS anchored conditional flow matching with group relative preference alignment. Stage I takes upsampled LRMS as the flow source and learns a PAN guided conditional vector field toward the HRMS reference, yielding a base prediction initialized from observed multispectral information. Stage II constructs a task specific candidate group around this prediction for spectral enhancement, spatial enhancement, and balanced fusion, then jointly evaluates the candidates by spectral relation consistency, spatial relation consistency, and reference fidelity. Softmax normalization within each group converts candidate rewards into relative supervision weights that guide refinement toward a better spectral and spatial balance. Furthermore, spatial preference aggregation combines global and local evidence, while temporal preference aggregation integrates relative evaluations across flow time conditions to improve preference stability. Extensive experiments demonstrate competitive performance and validate explicit group relative fusion preferences for generative pansharpening.

1 INTRODUCTION

Satellite imaging systems are subject to physical imaging limitations. Multispectral sensors preserve spectral responses but provide limited spatial detail, whereas panchromatic sensors capture fine structures with limited spectral information Vivone et al. (2015); Ghamisi et al. (2019); Meng et al. (2019). Pansharpening fuses a low resolution multispectral (LRMS) image with a high resolution panchromatic (PAN) image to estimate a high resolution multispectral (HRMS) image. Its central challenge is to inject PAN guided high frequency details without corrupting the spectral information observed in LRMS. Therefore, pansharpening is not merely spatial resolution enhancement, but the joint modeling of spectral fidelity and spatial detail reconstruction.

Traditional pansharpening methods are based on mathematical models. Component substitution, multiresolution analysis, and variational optimization inject spatial details through transformations or degradation models Laben & Brower (2000); Haydn et al. (1982); Nunez et al. (1999); Fu et al. (2019). Their fixed transformations have limited ability to adapt when image statistics vary across scenes and sensors. Deep learning learns spectral and spatial fusion from paired data. CNN based target adaptation, deep unfolding, and adaptive feature fusion improve detail injection and spectral preservation Masi et al. (2016); Scarpa et al. (2018); Yang et al. (2022); Wang et al. (2024); Zhang et al. (2022; 2023); Zhao et al. (2023). Transformer, frequency based, and spatial spectral interaction models further develop cross modal fusion Zhou et al. (2022); Meng et al. (2024); Xing et al. (2023); Zhang et al. (2024); He et al. (2024); Huang et al. (2025); Li et al. (2025); Luo et al. (2025). Recent work also studies staged adaptation and multimodal fusion architectures Wu et al. (2025); Do et al. (2025); Wang et al. (2026b); Zheng et al. (2026). Generative models provide another formulation for pansharpening. Diffusion models use reverse denoising for conditional image generation Xing

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

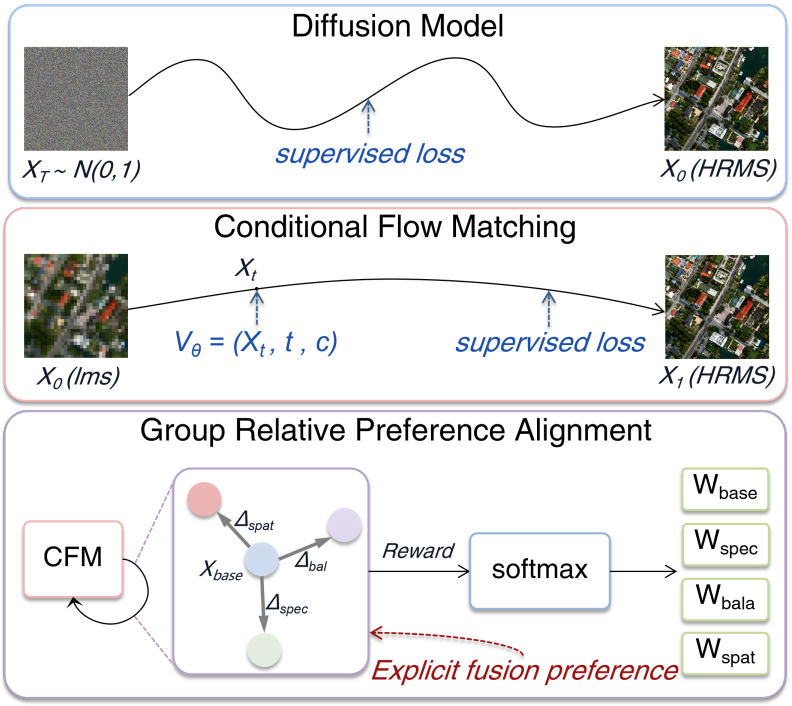

Figure 1: Motivation of FMGRPA. CFM defines an LRMS anchored reconstruction path under PAN guidance. GRPA compares structured spectral and spatial alternatives around the base prediction to make fusion preferences explicit.

et al. (2024); Meng et al. (2025); Hou et al. (2026); Cao et al. (2024); Zhong et al. (2024); Ho et al. (2020); Song et al. (2021); Karras et al. (2022). Flow matching learns a time conditioned vector field along a linear interpolation path Lipman et al. (2023); Tong et al. (2024); Liu et al. (2023). For pansharpening, the observed LRMS image, PAN guidance, and paired HRMS reference naturally provide the source image, condition, and target image.

However, most learning based and generative methods still rely on predefined reconstruction losses against the reference image to jointly constrain spectral and spatial quality. Pixel level reconstruction objectives drive predictions toward the HRMS ground truth, but measure only the overall error against the reference without explicitly distinguishing spectral shifts from missing spatial details. When two candidates have similar reconstruction errors, yet one better preserves LRMS spectral relations while the other contains richer high frequency structures from PAN, a single reconstruction objective cannot directly express the fusion preference required by pansharpening. In other words, existing methods usually encode the tradeoff between spectral preservation and spatial detail injection implicitly in a fixed loss, but lack a task specific relative feedback mechanism around the current prediction.To address this limitation, we propose FMGRPA, a two stage framework that separates CFM reconstruction from preference alignment. The first stage learns an LRMS anchored base prediction. Inspired by relative comparison within a group in GRPO Shao et al. (2024), the second stage constructs controlled candidates around this prediction and evaluates their spectral relation consistency, spatial relation consistency, and reference fidelity. Spatial and temporal aggregation combines the resulting group relative weights over image regions and selected time conditions.

For the first stage, we use conditional flow matching (CFM) to model pansharpening as a conditional flow from LRMS to the reference image under PAN guidance Lipman et al. (2023); Tong et al. (2024). This formulation follows the source, condition, and target structure of pansharpening. It starts from the observed LRMS image, takes PAN as spatial guidance, and learns from the paired HRMS reference. During training, the model receives a state sampled along the path between LRMS and the reference image together with the PAN and LRMS condition, and predicts the HRMS image. Starting from LRMS keeps the observed spectral content on the reconstruction path before PAN guided details are introduced. The trained CFM model is queried at selected time conditions to provide the base predictions for GRPA fine tuning.

The second stage introduces group relative preference alignment (GRPA) for the pretrained CFM model. Starting from the CFM output, GRPA constructs a candidate group containing the base prediction, a spectral variant guided by LRMS, a spatial variant guided by PAN, and a balanced variant. The spectral variant adjusts the low frequency content toward LRMS, whereas the spatial variant incorporates PAN high frequency details. The balanced variant combines both signals. A hybrid reward evaluates each candidate using spectral relation consistency, spatial relation consistency, and reference fidelity. Softmax normalization converts the group reward scores into relative weights for the candidate loss. The final objective combines this loss with ground truth reconstruction supervision. Spatial aggregation estimates weights over the full image and local regions, and temporal aggregation averages weights obtained at multiple selected time conditions. GRPA therefore refines the CFM prediction using spectral and spatial preferences expressed through group comparison.

In summary, this work makes the following contributions.

- We propose FMGRPA, a two stage generative framework for pansharpening that combines conditional flow reconstruction starting from LRMS with group relative preference alignment, thereby decoupling base image generation from spectral and spatial preference refinement.
- We propose task specific structured candidate construction and group relative weighting. By comparing spectral, spatial, and balanced candidates, the method converts spectral relation consistency, spatial relation consistency, and reference fidelity into explicit relative supervision signals.
- We design spatial and temporal preference aggregation to combine global and local image evidence with candidate evaluations at multiple flow time conditions, providing more stable and fine grained fusion preferences.

2 Related Work

2.1 Pansharpening

Existing pansharpening methods can be divided into traditional methods and deep learning methods. Traditional methods mainly include component substitution and multiresolution analysis. Component substitution replaces an intensity component with PAN information, while multiresolution analysis injects high frequency details at different image scales Laben & Brower (2000); Nunez et al. (1999). These methods depend on manually designed transformations and prior knowledge, which can limit their adaptation to different sensors and scenes.

PNN introduced CNNs to pansharpening Masi et al. (2016). Subsequent CNN based methods Yang et al. (2017); Xu et al. (2021) further developed deep feature learning for spectral and spatial fusion. Recent methods employ target adaptation, deep unfolding, adaptive feature fusion, and cross modal interaction to integrate PAN details with LRMS spectral information Scarpa et al. (2018); Yang et al. (2022); Zhang et al. (2022; 2023); Zhao et al. (2023); Zhou et al. (2022); Wang et al. (2024); Do et al. (2025); Wang et al. (2026b). Frequency based methods explore spatial and frequency domain information for global and local feature interaction Meng et al. (2024); Xing et al. (2023); Zhang et al. (2024); He et al. (2024); Huang et al. (2025); Li et al. (2025); Luo et al. (2025); Zheng et al. (2026). Other studies consider limited supervision, target adaptation, and unknown degradations Ciotola et al. (2022); Lin et al. (2024); Xu et al. (2023); Xing et al. (2024); Zhang et al. (2025).

Generative models have also been introduced into pansharpening. Recent diffusion methods Meng et al. (2025); Cao et al. (2024); Zhong et al. (2024) reconstruct HRMS images through conditional denoising. These methods build on diffusion models that generate images through a sequence of noisy states Ho et al. (2020); Song et al. (2021); Karras et al. (2022). More recently, flow matching has provided another formulation based on a time conditioned vector field between image states.

2.2 Flow Matching

Flow matching learns a time conditioned vector field along a prescribed path between source and target states Lipman et al. (2023); Liu et al. (2023). During training, intermediate states are sampled

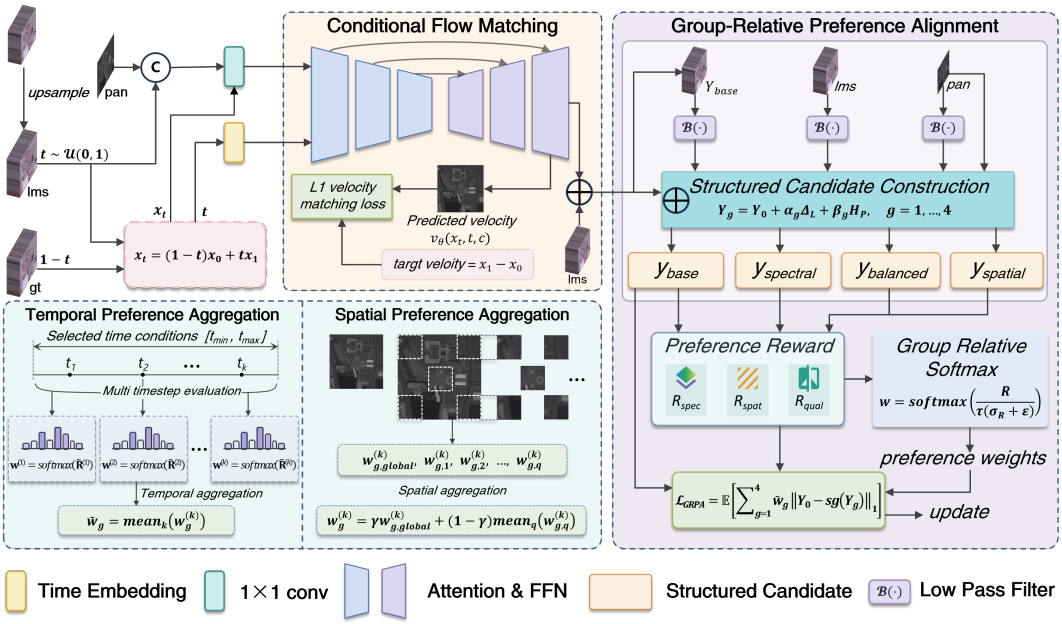

Figure 2: Overview of FMGRPA. The first stage learns a conditional flow from the upsampled LRMS image to the HRMS reference under PAN guidance. The second stage performs GRPA by constructing structured candidates, evaluating candidate preferences, and aggregating preference weights over spatial regions and time conditions.

from the path to learn the corresponding vector field. This formulation has been studied in image generation and other generative tasks. In pansharpening, OTFM applies flow matching to HRMS reconstruction with LRMS and PAN as conditional inputs Cao et al. (2025). FMGRPA follows this formulation and uses CFM as the reconstruction stage before post training.

2.3 Preference Alignment and Post Training

Preference alignment adapts a pretrained model using preference data. DPO learns directly from pairwise preferences Rafailov et al. (2023), while GRPO estimates relative advantages within a group without a separate critic model Shao et al. (2024). Preference optimization has also been extended to diffusion models Wallace et al. (2024); Black et al. (2024) and flow matching models Wang et al. (2026a). These studies show that pairwise and group comparisons can provide post training signals for generative models. FMGRPA applies group relative preference alignment to pansharpening.

3 Method

3.1 Framework

Given a low resolution multispectral (LRMS) image $\mathbf{M}$ and a panchromatic (PAN) image $\mathbf{P}$, pansharpening aims to estimate the high resolution multispectral (HRMS) reference $\mathbf{X}$. We bicubically upsample $\mathbf{M}$ to obtain $\mathbf{L}$ and concatenate it with $\mathbf{P}$ as the conditioning input $\mathcal{C} = [\mathbf{L}, \mathbf{P}]$.

The upsampled LRMS image provides low frequency spectral information, whereas the PAN image supplies high frequency spatial details. Their complementary roles require spectral information to be preserved while details guided by PAN are incorporated. FMGRPA addresses this issue from both modeling and optimization perspectives through a two stage design, as illustrated in Figure 2. In Stage I, CFM learns a conditional flow from the upsampled LRMS image to the HRMS reference under PAN guidance. The model estimates the HRMS image from a sampled state under the LRMS and PAN conditions. Using the upsampled LRMS image as the source retains the observed spectral information before preference alignment.

The CFM model uses a backbone with an encoder and a decoder, together with neighborhood attention blocks. It processes the image state conditioned on time in the main branch and injects the PAN and LRMS condition at multiple scales, making spectral and spatial guidance available throughout reconstruction. A sinusoidal embedding represents the time variable. The model predicts a residual with respect to the upsampled LRMS image to obtain the final HRMS image.

Stage II fine tunes the CFM model through GRPA. It forms a candidate group around each base prediction, with the candidates combining LRMS spectral information and details guided by PAN in different ways. The candidate group defines a controlled comparison set for fusion quality. A hybrid reward measures spectral relation consistency, spatial relation consistency, and reference fidelity for each candidate. At several selected time conditions, CFM produces the corresponding base predictions, from which GRPA derives group relative weights. For each time condition, the weights are first aggregated over the full image and local regions to account for image level and local fusion quality. The resulting weights are then averaged across time conditions. The weighted candidate loss is optimized jointly with the ground truth reconstruction loss.

3.2 Conditional Flow Matching for Pansharpening

In Stage I, we model pansharpening as a conditional flow from the upsampled LRMS image to the HRMS reference under PAN guidance. Given a training triplet $(\mathbf{M}, \mathbf{P}, \mathbf{X})$, we upsample $\mathbf{M}$ to obtain $\mathbf{L}$ and use $\mathcal{C} = [\mathbf{L}, \mathbf{P}]$ as the conditional input. $\mathbf{L}$ serves as the source to retain observed spectral information before PAN details are introduced. Sampling $t \sim \mathcal{U}(0, 1)$ gives a state along the flow path

$$\mathbf{X}_t = (1 - t)\mathbf{L} + t\mathbf{X} + \sigma\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

where $\boldsymbol{\epsilon}$ is standard Gaussian noise and σ controls the noise level. Along the linear path, $\mathbf{X}_t$ evolves from $\mathbf{L}$ to $\mathbf{X}$ as t increases, while LRMS and PAN provide spectral information and spatial guidance. The target velocity is $\mathbf{u}_t = \mathbf{X} - \mathbf{L}$.

Given $\mathbf{X}_t$, the network predicts the HRMS image as $f_\theta(\mathbf{X}_t, t, \mathcal{C})$, and its predicted velocity is

$$\mathbf{s}_\theta(\mathbf{X}_t, t, \mathcal{C}) = f_\theta(\mathbf{X}_t, t, \mathcal{C}) - \mathbf{L}. \tag{2}$$

The CFM encoder–decoder uses neighborhood attention blocks and fuses multiscale PAN and LRMS features with those extracted from $\mathbf{X}_t$. A time embedding represents t for time conditioned prediction, while residual learning with $\mathbf{L}$ yields $f_\theta(\mathbf{X}_t, t, \mathcal{C})$, preserving LRMS spectral information while incorporating PAN details.

For each sampled state, the objective matches the predicted and target velocities:

$$\begin{aligned}\mathcal{L}_{\mathrm{CFM}} = \mathbb{E}\Big[&\left\|\mathbf{s}_\theta(\mathbf{X}_t, t, \mathcal{C}) - \mathbf{u}_t\right\|_1 \\ &+ \ell_{\mathrm{SSIM}}\big(f_\theta(\mathbf{X}_t, t, \mathcal{C}), \mathbf{X}\big)\Big],\end{aligned} \tag{3}$$

where θ denotes the model parameters and ℓ_{SSIM} the SSIM loss. The first term matches the flow velocity, and the second preserves the structure of the reconstructed HRMS image, providing the base reconstruction for explicit spectral and spatial preference alignment in Stage II.

After Stage I, CFM takes $\mathbf{L}$ at a selected time step t_e to obtain the base prediction

$$\mathbf{Y}_0 = f_\theta(\mathbf{L}, t_e, \mathcal{C}). \tag{4}$$

The prediction $\mathbf{Y}_0$ then serves as the base reconstruction for constructing and comparing the candidate group in GRPA.

3.3 Group Relative Preference Alignment

CFM first produces a ground truth supervised base prediction. GRPA fine tunes CFM with a candidate group constructed around $\mathbf{Y}_0$. Controlled changes to LRMS low frequency content and PAN high frequency details isolate spectral and spatial fusion alternatives without separate prediction branches. A hybrid reward evaluates the candidates, and softmax normalized group scores yield relative weights for the candidate loss, thereby adding a task specific group relative signal to CFM fine tuning while the ground truth loss remains in the final objective.

3.3.1 CANDIDATE GENERATION

Candidate generation follows the spectral–spatial decomposition used in pansharpening: the low frequency component of the upsampled LRMS image provides spectral information, whereas PAN high frequency details provide spatial structures. Let $\mathcal{B}(\cdot)$ denote a low pass filter and define the LRMS spectral term as

$$\Delta_{\mathrm{L}} = \mathcal{B}(\mathbf{L}) - \mathcal{B}(\mathbf{Y}_0). \tag{5}$$

Let $\mathbf{H}_P$ denote the PAN high frequency detail $\mathbf{P} - \mathcal{B}(\mathbf{P})$, replicated across the LRMS spectral channels. We form four candidate predictions:

$$\mathbf{Y}_g = \mathbf{Y}_0 + \alpha_g \Delta_{\mathrm{L}} + \beta_g \mathbf{H}_P, \qquad g = 1, \ldots, 4, \tag{6}$$

where $g = 1, 2, 3, 4$ indexes the base, spectral, balanced, and spatial candidates, respectively. The base uses $\alpha_1 = \beta_1 = 0$ and $\mathbf{Y}_1 = \mathbf{Y}_0$; the spectral candidate uses only Δ_{L} ($\beta_2 = 0$), the spatial candidate only $\mathbf{H}_P$ ($\alpha_4 = 0$), and the balanced candidate both terms. The nonzero coefficients control their spectral and spatial contributions, and all candidates are clipped to the valid image range. The base serves as the group reference: its score participates in group normalization, while the others provide the spectral, spatial, and joint alternatives used by the candidate loss.

3.3.2 REWARD DESIGN

The candidate group is evaluated by a hybrid reward combining spectral relation consistency, spatial relation consistency, and reference fidelity. Let C be the number of spectral bands, $\mathcal{Q}(\cdot, \cdot)$ the local Q index, and $\mathcal{S}_r(\cdot)$ bicubic downsampling by sensor ratio r. For each candidate $\mathbf{Y}_g$, the spectral relation reward is

$$R_g^{\mathrm{spec}} = -\frac{2}{C(C-1)} \sum_{i<j} \left| \mathcal{Q}(Y_{g,i}, Y_{g,j}) - \mathcal{Q}(M_i, M_j) \right|, \tag{7}$$

which compares candidate interband relations with those observed in LRMS. The spatial relation reward is

$$R_g^{\mathrm{spat}} = -\frac{1}{C} \sum_{i=1}^{C} \left| \mathcal{Q}(Y_{g,i}, \mathbf{P}) - \mathcal{Q}(M_i, \mathcal{S}_r(\mathbf{P})) \right|. \tag{8}$$

It compares each candidate–PAN relation at the PAN scale with the corresponding LRMS–PAN relation at the LRMS scale. We combine the two relation rewards with the reference term as

$$R_g = R_g^{\mathrm{spec}} + 0.8 R_g^{\mathrm{spat}} - 2.5 \left\| \mathbf{Y}_g - \mathbf{X} \right\|_1. \tag{9}$$

The three terms preserve LRMS spectral relations, favor resolution consistent PAN detail injection, and anchor candidate ranking to the HRMS reference, respectively. The resulting score yields relative weights within the candidate group.

We normalize reward scores within each candidate group using softmax:

$$\mathbf{w} = \mathrm{softmax}\left(\frac{\mathbf{R} - \mu_R}{\tau(\sigma_R + \varepsilon)} \right), \tag{10}$$

where $\mathbf{R} = [R_1, \ldots, R_4]$, and μ_R and σ_R are its mean and standard deviation. The vector $\mathbf{w} = [w_1, \ldots, w_4]$ contains the group relative weights; ε provides numerical stability, and τ controls the sharpness of the weight distribution. Normalization removes the influence of the absolute reward scale for each image, while softmax assigns larger weights to candidates with higher scores and retains all four candidates in the fine tuning objective.

3.4 SPATIAL AND TEMPORAL PREFERENCE AGGREGATION

The reward in Eq. (9) scores a candidate group at one time condition, but a full image score compresses local spectral and spatial differences and cannot distinguish spectral distortion or missing PAN details when candidates have similar global fusion quality. A single time condition also reflects only one response of the time conditioned CFM model. We therefore evaluate each candidate group on the full image and local regions at K selected time conditions $\{t^{(k)}\}_{k=1}^{K}$. At each $t^{(k)}$, CFM is queried from $\mathbf{L}$ to construct the candidate group, and Eq. (10) converts its scores into group relative weights. Spatial and temporal aggregation provides the final weights for the GRPA loss, which reweight the candidate group evaluated at the alignment time t_e.

Spatial aggregation. For each time condition, we evaluate the candidate group on the full image and fixed-grid local patches. Candidate images, the PAN image, and the HRMS reference are cropped at the same positions, while the LRMS image is cropped at the corresponding reduced resolution. Let $w_{g,\mathrm{global}}^{(k)}$ and $w_{g,q}^{(k)}$ be the preference weights for candidate g from the full image and local patch q, respectively, where q indexes the grid positions. We combine them as

$$w_g^{(k)} = \gamma w_{g,\mathrm{global}}^{(k)} + (1-\gamma)\,\mathrm{mean}_q\left(w_{g,q}^{(k)}\right), \tag{11}$$

where γ balances full image and local group relative weights. The full image term measures overall fusion quality, while the local term incorporates regional spectral and spatial differences, accounting for LRMS spectral preservation and PAN detail injection at both levels.

Temporal aggregation. After spatial aggregation, we average the preference weights across the selected time conditions:

$$\bar{w}_g = \mathrm{mean}_k\left(w_g^{(k)}\right). \tag{12}$$

The final weight $\bar{w}_g$ aggregates the group relative weight for candidate g across the selected time conditions and is calculated from CFM reconstruction results at multiple conditions rather than one. These conditions estimate the weights, while the GRPA loss uses the base prediction and candidates constructed at the alignment time t_e.

The candidates $\mathbf{Y}_g$ are constructed from the base prediction $\mathbf{Y}_0$ at t_e. Their aggregated weights are detached and reweight the candidate loss:

$$\mathcal{L}_{\mathrm{GRPA}} = \mathbb{E}\left[\sum_{g=1}^{4} \bar{w}_g \left\|\mathbf{Y}_0 - \mathrm{sg}(\mathbf{Y}_g)\right\|_1\right] \tag{13}$$

where the expectation is over the training samples and $\mathrm{sg}(\cdot)$ stops gradients through the candidates. For the base candidate, the stopped gradient term is zero because $\mathbf{Y}_1 = \mathbf{Y}_0$, while its score provides the reference for group normalization. The final post training objective is

$$\mathcal{L}_{\mathrm{post}} = \lambda_a \ell_{\mathrm{gt}}\big(\mathbf{Y}_0, \mathbf{X}\big) + \lambda_g \mathcal{L}_{\mathrm{GRPA}}, \tag{14}$$

where λ_a and λ_g balance the ground truth reconstruction loss and the GRPA candidate loss. The ground truth term maintains fidelity to the HRMS reference, while the GRPA term introduces task specific group relative weighting estimated over image regions and time conditions.

4 Experiments

4.1 Datasets and Benchmarks

We conduct experiments on three widely used pansharpening datasets: WorldView-2 (WV2), WorldView-3 (WV3), and GaoFen-2 (GF2). WV2 contains 760 training image pairs and 80 testing image pairs, WV3 contains 2150 training image pairs and 200 testing image pairs, and GF2 contains 2712 training image pairs and 200 testing image pairs. We evaluate the competing methods under the standard reduced-resolution and full-resolution protocols. The reduced-resolution protocol provides reference high-resolution multispectral (HRMS) images for supervised evaluation, while the full-resolution protocol evaluates real-scale PAN and MS observations without HRMS references.

The baseline set covers both traditional and learning based pansharpening methods, including GS, Brovey, IHS, PNN, MSDCNN, GPPNN, MutNet, INNformer, HFIN, and FAME-Net. For data preparation, MS images are cropped into 32×32 patches and the corresponding PAN images are cropped into 128×128 patches. The LRMS input is bicubically upsampled to the PAN resolution before fusion. WV2 and GF2 samples are normalized by 255, while WV3 samples are normalized by 2047 according to their radiometric ranges.

For reduced-resolution evaluation, we report PSNR (↑), SSIM (↑), SAM (↓), and ERGAS (↓). For full-resolution evaluation, we report D_λ (↓), D_s (↓), and QNR (↑), which do not require reference images. Reduced-resolution results provide the main supervised comparison, while full-resolution results examine the behavior of each method on real scale observations.More descriptions of the implementation details and dataset are provided in the supplementary material.

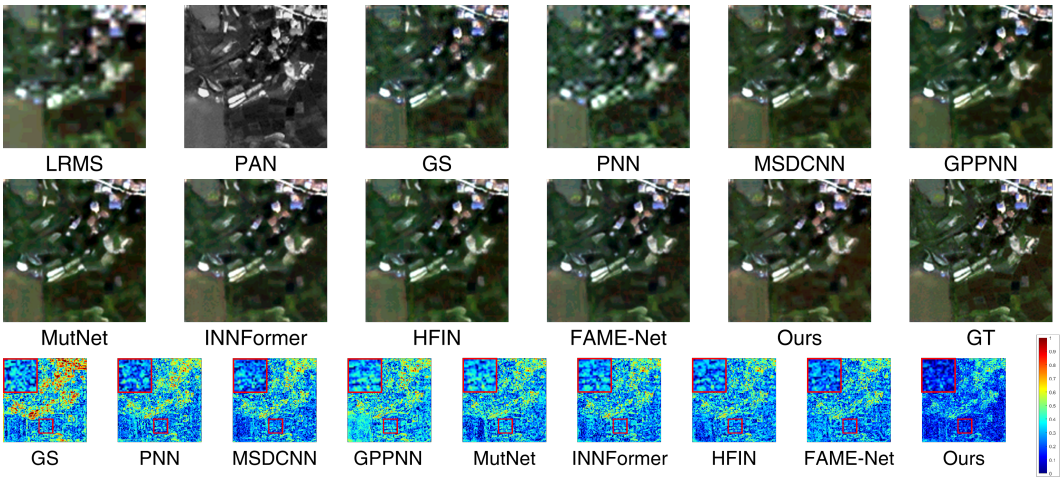

Figure 3: The visual comparisons between other pansharpening methods and our method on WorldView-II satellite.

Table 1: Experimental results of all the competing methods on the three benchmark datasets. The best and second best results are shown in bold and underlined, respectively.

	WV2				WV3				GF2			
Method	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
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	40.7550	0.9624	0.0259	1.0646	29.9418	0.9121	0.0824	3.3206	43.1208	0.9704	0.0172	0.8528
MSDCNN	41.3355	0.9664	0.0242	0.9940	30.3038	0.9184	0.0782	3.1884	45.6874	0.9827	0.0135	0.6389
GPPNN	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	41.6773	0.9705	0.0224	0.9519	30.4907	0.9223	0.0749	3.1125	47.3042	0.9892	0.0102	0.5481
INNformer	41.6903	0.9704	0.0227	0.9514	30.5365	0.9225	0.0747	3.0997	47.3528	0.9893	0.0102	0.5479
HFIN	42.2319	0.9714	**0.0215**	0.8807	30.6147	0.9203	0.0742	3.0786	48.8783	**0.9898**	0.0093	**0.4591**
FAME-Net	42.0262	0.9723	**0.0215**	0.9172	30.9903	0.9287	**0.0697**	2.9531	47.6721	**0.9898**	0.0098	0.5242
FMGRPA	**42.6738**	**0.9739**	**0.0215**	**0.8754**	**31.2389**	**0.9393**	0.0702	**2.9425**	**49.1965**	0.9896	**0.0089**	0.4594

4.2 Comparison with State of the Art Methods

We compare FMGRPA with ten representative pansharpening methods—three traditional and seven learning-based baselines—under reduced- and full-resolution protocols.

Qualitative comparison. Figure 3 presents qualitative comparisons on WV2: the top two rows show RGB results, and the last row shows mean squared error maps relative to the ground truth. FMGRPA produces the visually closest results to the reference, with smaller discrepancies in the displayed regions.

Evaluation on reduced-resolution scenes. Table 1 reports results on WV2, WV3, and GF2; bold and underlining denote the best and second-best values. FMGRPA attains the best or tied-best result on nine of the twelve metric–dataset pairs and ranks second on the other three. On WV2, it achieves the highest PSNR (42.6738) and SSIM (0.9739), exceeding the second-best values by 0.4419 dB and 0.0016; its SAM (0.0215) matches the best result, and its ERGAS (0.8754) is the lowest. On WV3, FMGRPA improves PSNR from 30.9903 to 31.2389 and SSIM from 0.9287 to 0.9393 over the second-best method, obtains the lowest ERGAS (2.9425), and ranks second in SAM (0.0702 versus 0.0697). On GF2, it yields the highest PSNR (49.1965) and lowest SAM (0.0089); its SSIM (0.9896) and ERGAS (0.4594) are within 0.0002 and 0.0003 of the best values. These results show consistent spectral and spatial fusion quality across the three sensors.

Evaluation on full-resolution scenes. For full-resolution GF2, where HRMS references are unavailable, we evaluate original-scale PAN and MS pairs using D_λ, D_s, and QNR without reference images. Table 2 shows that FMGRPA obtains the lowest D_λ (0.0519) and D_s (0.1097), together

with the highest QNR (0.8440) among the reported methods. These results indicate a favorable spectral and spatial balance on real-scale observations.

Table 2: Full-resolution comparison on GF2 without reference images. The best and second best results are shown in bold and underlined, respectively.

Metric	PNN	MSDCNN	GPPNN	MutNet	INNformer	HFIN	FAME-Net	FMGRPA
$D_\lambda \downarrow$	0.0746	0.0734	0.0782	0.0694	0.0697	0.0710	<u>0.0674</u>	**0.0519**
$D_s \downarrow$	0.1164	0.1151	0.1253	0.1118	0.1128	<u>0.1098</u>	0.1121	**0.1097**
QNR↑	0.8191	0.8215	0.8073	0.8247	0.8253	0.8261	<u>0.8291</u>	**0.8440**

4.3 Ablation Experiments

We conduct GF2 ablations of the post-training GRPA, spatial aggregation, temporal aggregation, and structured candidate group. Table 3 compares the Stage I CFM reconstruction with GRPA variants to isolate the effects of GRPA and the aggregation modules.

Table 3: Ablation results on GF2. The best and second best results are shown in bold and underlined, respectively.

Metric	Stage I CFM	CFM + GRPA	w/o temporal aggregation	w/o spatial aggregation	FMGRPA
PSNR↑	48.8446	48.9451	<u>49.0718</u>	49.0607	**49.1965**
SSIM↑	0.9878	0.9885	<u>0.9891</u>	0.9889	**0.9896**
SAM↓	0.0098	<u>0.0092</u>	0.0094	<u>0.0092</u>	**0.0089**
ERGAS↓	0.4862	0.4681	<u>0.4630</u>	0.4657	**0.4594**

GRPA improves the Stage I CFM reconstruction on all four metrics. Spatial and temporal aggregation further improve PSNR, SSIM, and ERGAS, and their combination achieves the best results on all four metrics. This demonstrates the effect of aggregating preference information over image regions and time conditions.

Table 4 examines the structured candidate group by removing one candidate from the full group at a time while retaining the remaining components.

Table 4: Candidate ablation on GF2. All variants retain Stage I CFM, hybrid reward, and spatial and temporal aggregation; each removes one member from the full structured candidate group. Best and second-best results are bold and underlined, respectively.

Metric	w/o Y_{base}	w/o Y_{spectral}	w/o Y_{balanced}	w/o Y_{spatial}	ours(full)
PSNR↑	<u>49.0512</u>	48.8168	48.9717	48.7796	**49.1965**
SSIM↑	<u>0.9892</u>	0.9891	0.9889	0.9881	**0.9896**
SAM↓	<u>0.0093</u>	0.0102	0.0098	0.0095	**0.0089**
ERGAS↓	<u>0.4653</u>	0.4772	0.4810	0.4784	**0.4594**

Each removal degrades final performance. The complete group obtains the best values on all four metrics, demonstrating that the base, spectral, spatial, and balanced candidates jointly support group relative preference alignment. Additional ablations are provided in the supplementary materials.

5 Conclusion

We present FMGRPA, a two-stage pansharpening framework combining CFM and GRPA to explicitly model spectral and spatial preferences. CFM preserves spectral information before incorporating PAN-guided details, while GRPA constructs spectral, spatial, and balanced candidates for preference refinement. A hybrid reward and group relative Softmax produce weights for spatial and temporal preference aggregation across regions and time. FMGRPA achieves competitive quantitative results and closer WV2 agreement with the reference.

AI USE STATEMENT

In this work, we used generative AI tools to assist with research ideation or execution. We have not used generative AI tools to generate synthetic datasets or formulate mathematical proofs, and other required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to search for and identify related work and to edit the manuscript for clarity and readability. We did not use generative AI tools to draft portions of the paper. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/14f75513f0f1ca01de1e826b52e6b840-Paper-Conference.pdf.

Zihan Cao, Shiqi Cao, Liang-Jian Deng, Xiao Wu, Junming Hou, and Gemine Vivone. Diffusion model with disentangled modulations for sharpening multispectral and hyperspectral images. *Information Fusion*, 104:102158, Apr 2024. ISSN 1566-2535. doi: 10.1016/j.inffus.2023.102158. URL http://dx.doi.org/10.1016/j.inffus.2023.102158.

Zihan Cao, Yu Zhong, and Liang-Jian Deng. Taming flow matching with unbalanced optimal transport into fast pansharpening. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2803–2813, October 2025.

Matteo Ciotola, Sergio Vitale, Antonio Mazza, Giovanni Poggi, and Giuseppe Scarpa. Pansharpening by convolutional neural networks in the full resolution framework. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022. ISSN 1558-0644. doi: 10.1109/tgrs.2022.3163887. URL http://dx.doi.org/10.1109/TGRS.2022.3163887.

Jeonghyeok Do, Sungpyo Kim, Geunhyuk Youk, Jaehyup Lee, and Munchurl Kim. Pan-crafter: Learning modality-consistent alignment for pan-sharpening. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4242–4252, October 2025.

Xueyang Fu, Zihuang Lin, Yue Huang, and Xinghao Ding. A variational pan-sharpening with local gradient constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10265–10274, June 2019.

Pedram Ghamisi, Behnood Rasti, Naoto Yokoya, Qunming Wang, Bernhard Höfle, Lorenzo Bruzzone, Francesca Bovolo, Mingmin Chi, Katharina Anders, Richard Gloaguen, Peter M. Atkinson, and Jón Atli Benediktsson. Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 7(1):6–39, Mar 2019. ISSN 2373-7468. doi: 10.1109/mgrs.2018.2890023. URL http://dx.doi.org/10.1109/MGRS.2018.2890023.

R. Haydn, G. W. Dalke, J. Henkel, and J. 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*, volume 1, pp. 599–616, Cairo, Egypt, 1982. Environmental Research Institute of Michigan.

Xuanhua He, Keyu Yan, Rui Li, Chengjun Xie, Jie Zhang, and Man Zhou. Frequency-adaptive pan-sharpening with mixture of experts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(3):2121–2129, Mar 2024. ISSN 2159-5399. doi: 10.1609/aaai.v38i3.27984. URL http://dx.doi.org/10.1609/aaai.v38i3.27984.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Junming Hou, Ran Ran, Sixing Chen, Zihao Chen, Xiaofeng Cong, Junling Li, and Liang-Jian Deng. Nodiff: Neural operator diffusion for multispectral image fusion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(6):4753–4761, Mar 2026. ISSN 2159-5399. doi: 10.1609/aaai.v40i6.42477. URL http://dx.doi.org/10.1609/aaai.v40i6.42477.

Jie Huang, Rui Huang, Jinghao Xu, Siran Peng, Yule Duan, and Liang-Jian Deng. Wavelet-assisted multi-frequency attention network for pansharpening. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(4):3662–3670, Apr 2025. ISSN 2159-5399. doi: 10.1609/aaai.v39i4.32381. URL http://dx.doi.org/10.1609/aaai.v39i4.32381.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems 35*, NeurIPS 2022, pp. 26565–26577. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022. doi: 10.52202/068431-1926. URL http://dx.doi.org/10.52202/068431-1926.

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, January 2000. URL https://patents.google.com/patent/US6011875A/en.

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 (ICCV)*, pp. 10908–10918, October 2025.

Huangxing Lin, Yuhang Dong, Xinghao Ding, Tianpeng Liu, and Yongxiang Liu. Unsupervised pan-sharpening via mutually guided detail restoration. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(4):3386–3394, Mar 2024. ISSN 2159-5399. doi: 10.1609/aaai.v38i4.28125. URL http://dx.doi.org/10.1609/aaai.v38i4.28125.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tUjvLIXarl.

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-2025, pp. 1702–1710. International Joint Conferences on Artificial Intelligence Organization, Sept 2025. doi: 10.24963/ijcai.2025/190. URL http://dx.doi.org/10.24963/ijcai.2025/190.

Giuseppe Masi, Davide Cozzolino, Luisa Verdoliva, and Giuseppe Scarpa. Pansharpening by convolutional neural networks. *Remote Sensing*, 8(7):594, July 2016. ISSN 2072-4292. doi: 10.3390/rs8070594. URL http://dx.doi.org/10.3390/rs8070594.

Ge Meng, Jingjia Huang, Yingying Wang, Zhenqi Fu, Xinghao Ding, and Yue Huang. Progressive high-frequency reconstruction for pan-sharpening with implicit neural representation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5):4189–4197, Mar 2024. ISSN 2159-5399. doi: 10.1609/aaai.v38i5.28214. URL http://dx.doi.org/10.1609/aaai.v38i5.28214.

Ge Meng, Jingjia Huang, Jingyan Tu, Yingying Wang, Yunlong Lin, Xiaotong Tu, Yue Huang, and Xinghao Ding. Accelerated diffusion via high-low frequency decomposition for pan-sharpening. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(6):6117–6125, Apr 2025. ISSN 2159-5399. doi: 10.1609/aaai.v39i6.32654. URL http://dx.doi.org/10.1609/aaai.v39i6.32654.

Xiangchao Meng, Huanfeng Shen, Huifang Li, Liangpei Zhang, and Randi Fu. Review of the pansharpening methods for remote sensing images based on the idea of meta-analysis: Practical discussion and challenges. *Information Fusion*, 46:102–113, Mar 2019. ISSN 1566-2535. doi: 10.1016/j.inffus.2018.05.006. URL http://dx.doi.org/10.1016/j.inffus.2018.05.006.

J. Nunez, X. Otazu, O. Fors, A. Prades, V. Pala, and R. Arbiol. Multiresolution-based image fusion with additive wavelet decomposition. *IEEE Transactions on Geoscience and Remote Sensing*, 37 (3):1204–1211, May 1999. ISSN 0196-2892. doi: 10.1109/36.763274. URL http://dx.doi.org/10.1109/36.763274.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36*, NeurIPS 2023, pp. 53728–53741. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-2338. URL http://dx.doi.org/10.52202/075280-2338.

Giuseppe Scarpa, Sergio Vitale, and Davide Cozzolino. Target-adaptive cnn-based pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 56(9):5443–5457, Sept 2018. ISSN 1558-0644. doi: 10.1109/tgrs.2018.2817393. URL http://dx.doi.org/10.1109/TGRS.2018.2817393.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=CD9Snc73AW.

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, May 2015. ISSN 1558-0644. doi: 10.1109/tgrs.2014.2361734. URL http://dx.doi.org/10.1109/TGRS.2014.2361734.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238, June 2024.

Hebaixu Wang, Meiqi Gong, Xiaoguang Mei, Hao Zhang, and Jiayi Ma. Deep unfolded network with intrinsic supervision for pan-sharpening. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5419–5426, Mar 2024. ISSN 2159-5399. doi: 10.1609/aaai.v38i6.28350. URL http://dx.doi.org/10.1609/aaai.v38i6.28350.

Shaomeng Wang, He Wang, Longquan Dai, and Jinhui Tang. Pc-flow: Preference alignment in flow matching via classifier. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(12): 10047–10055, Mar 2026a. ISSN 2159-5399. doi: 10.1609/aaai.v40i12.37971. URL http://dx.doi.org/10.1609/aaai.v40i12.37971.

Yingying Wang, Xuanhua He, Chen Wu, Jialing Huang, Suiyun Zhang, Rui Liu, Xinghao Ding, and Haoxuan Che. Mmmamba: A versatile cross-modal in context fusion framework for pansharpening and zero-shot image enhancement. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(22):18656–18664, Mar 2026b. ISSN 2159-5399. doi: 10.1609/aaai.v40i22.38933. URL http://dx.doi.org/10.1609/aaai.v40i22.38933.

RuoCheng Wu, Zien Zhang, Shangqi Deng, Yule Duan, and Liang-Jian Deng. Panadapter: Two-stage fine-tuning with spatial-spectral priors injecting for pansharpening. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(8):8450–8459, Apr 2025. ISSN 2159-5399. doi: 10.1609/aaai.v39i8.32912. URL http://dx.doi.org/10.1609/aaai.v39i8.32912.

Yinghui Xing, Yan Zhang, Houjun He, Xiuwei Zhang, and Yanning Zhang. Pansharpening via frequency-aware fusion network with explicit similarity constraints. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023. ISSN 1558-0644. doi: 10.1109/tgrs.2023.3281829. URL http://dx.doi.org/10.1109/TGRS.2023.3281829.

Yinghui Xing, Litao Qu, Shizhou Zhang, Kai Zhang, Yanning Zhang, and Lorenzo Bruzzone. Cross-diff: Exploring self-supervised representation of pansharpening via cross-predictive diffusion model. *IEEE Transactions on Image Processing*, 33:5496–5509, 2024. ISSN 1941-0042. doi: 10.1109/tip.2024.3461476. URL http://dx.doi.org/10.1109/TIP.2024.3461476.

Qizhi Xu, Yuan Li, Jinyan Nie, Qingjie Liu, and Mengyao Guo. Upangan: Unsupervised pansharpening based on the spectral and spatial loss constrained generative adversarial network. *Information Fusion*, 91:31–46, Mar 2023. ISSN 1566-2535. doi: 10.1016/j.inffus.2022.10.001. URL http://dx.doi.org/10.1016/j.inffus.2022.10.001.

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 (CVPR)*, pp. 1366–1375, June 2021.

Gang Yang, Man Zhou, Keyu Yan, Aiping Liu, Xueyang Fu, and Fan Wang. Memory-augmented deep conditional unfolding network for pan-sharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1788–1797, June 2022.

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 (ICCV)*, pp. 5449–5457, Oct 2017.

Jie Zhang, Xuanhua He, Keyu Yan, Ke Cao, Rui Li, Chengjun Xie, Man Zhou, and Danfeng Hong. Pan-sharpening with wavelet-enhanced high-frequency information. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. ISSN 1558-0644. doi: 10.1109/tgrs.2024.3367165. URL http://dx.doi.org/10.1109/TGRS.2024.3367165.

Kai Zhang, Anfei Wang, Feng Zhang, Wenxiu Diao, Jiande Sun, and Lorenzo Bruzzone. Spatial and spectral extraction network with adaptive feature fusion for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. ISSN 1558-0644. doi: 10.1109/tgrs.2022.3187025. URL http://dx.doi.org/10.1109/TGRS.2022.3187025.

Kai Zhang, Anfei Wang, Feng Zhang, Wenbo Wan, Jiande Sun, and Lorenzo Bruzzone. Spatial-spectral dual back-projection network for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023. ISSN 1558-0644. doi: 10.1109/tgrs.2023.3266799. URL http://dx.doi.org/10.1109/TGRS.2023.3266799.

Zhiyuan Zhang, Haoxuan Li, Chengjie Ke, Jun Chen, and Xin Tian. Deep variational network for blind pansharpening. *IEEE Transactions on Neural Networks and Learning Systems*, 36(5):9283–9297, May 2025. ISSN 2162-2388. doi: 10.1109/tnnls.2024.3436850. URL http://dx.doi.org/10.1109/TNNLS.2024.3436850.

Xin Zhao, Jiayi Guo, Yueting Zhang, and Yirong Wu. Asymmetric bidirectional fusion network for remote sensing pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023. ISSN 1558-0644. doi: 10.1109/tgrs.2023.3296510. URL http://dx.doi.org/10.1109/TGRS.2023.3296510.

Huangqimei Zheng, Chengyi Pan, Qian Jiang, Wei Zhou, and Xin Jin. Hierarchical dual-domain fusion with frequency-guided spatial modeling for pan-sharpening. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(16):13405–13413, Mar 2026. ISSN 2159-5399. doi: 10.1609/aaai.v40i16.38344. URL http://dx.doi.org/10.1609/aaai.v40i16.38344.

Yu Zhong, Xiao Wu, Zihan Cao, Hong-Xia Dou, and Liang-Jian Deng. Ssdiff: Spatial-spectral integrated diffusion model for remote sensing pansharpening. In *Advances in Neural Information Processing Systems 37*, NeurIPS 2024, pp. 77962–77986. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-2478. URL http://dx.doi.org/10.52202/079017-2478.

Huanyu Zhou, Qingjie Liu, and Yunhong Wang. Panformer: A transformer based model for pansharpening. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, July 2022. doi: 10.1109/icme52920.2022.9859770. URL http://dx.doi.org/10.1109/ICME52920.2022.9859770.