PANDA: PAN ALIGNMENT-AWARE DISTILLATION VIA ADAPTIVE FITTING FOR PANSHARPENING

Anonymous authors
Paper under double-blind review

ABSTRACT

PAN-sharpening aims to reconstruct high-resolution multispectral (HRMS) images by combining the spatial detail of high-resolution panchromatic (PAN) images with the spectral information of low-resolution multispectral (MS) images. In practice, however, PAN and MS observations often exhibit residual spatial displacement due to differences in sensor geometry, acquisition timing, and spatial resolution. Directly fusing such misaligned inputs can entangle complementary image content with geometric discrepancy, leading to inaccurate spatial-spectral reconstruction. To address this, we propose PANDA (PAN Alignment-Aware Distillation via Adaptive Fitting), an align-first PAN-sharpening framework with reliability-aware adaptive supervision. PANDA first corrects PAN toward the original MS coordinate frame before joint feature extraction. PAN aligner is jointly optimized with a reconstruction network using HRMS reconstruction and self-supervised shift consistency, enabling task-relevant correction without displacement annotations. The trained model then serves as a frozen reference for student fitting, where different reliability cues regulate different forms of supervision. Reconstruction-based cues adaptively balance GT fitting and reference imitation, while alignment consistency modulates geometry-sensitive supervision and aligner refinement. At inference, only the student aligner and reconstruction network are retained. Extensive experiments across multiple satellite datasets demonstrate state-of-the-art performance at both reduced and full resolutions, showing the effectiveness of aligning before fusion and fitting according to reliability.

1 INTRODUCTION

Satellite imagery supports a wide range of applications, including environmental monitoring and urban analysis, which require both fine spatial detail and rich spectral information Li et al. (2020); Wellmann et al. (2020). Due to inherent sensor limitations, however, a single observation cannot simultaneously provide both at high resolution. Modern satellite systems therefore acquire a high-resolution panchromatic (PAN) image together with a low-resolution multispectral (MS) image (Thomas et al., 2008). PAN-sharpening aims to fuse these complementary observations to reconstruct a high-resolution multispectral (HRMS) image, preserving the spatial structures of PAN while maintaining the spectral information of MS (Vivone et al., 2014; Do et al., 2025; Lee et al., 2021; Meng et al., 2023; Zhang et al., 2019; Masi et al., 2016). Since PAN and MS themselves serve as the spatial and spectral references for reconstruction (Ciotola et al., 2022; Otazu et al., 2005), accurately modeling their relationship is central to faithful PAN-sharpening.

As shown in Fig. 2, in practice, however, PAN and MS observations are not often perfectly aligned. Differences in sensor geometry, acquisition timing, and spatial resolution can introduce residual spatial displacement, causing corresponding structures to appear at different locations across the two modalities (Lee et al., 2021; Do et al., 2025). Nevertheless, many PAN-sharpening methods directly concatenate or jointly process PAN and MS features without explicitly resolving this discrepancy (Masi et al., 2016; Yang et al., 2017; Zhang et al., 2019). As a result, complementary image content and geometric mismatch are entangled from the beginning of feature extraction, often leading to blurred structures, double edges, or spectral distortion (Lee et al., 2021; Do et al., 2025). Content-adaptive filtering and non-local aggregation improve feature modeling but do not explicitly correct cross-modal displacement (Jin et al., 2022; Duan et al., 2024). Alternatively, feature alignment can map MS information toward PAN structures (Lee et al., 2021), yet doing so relocates

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.

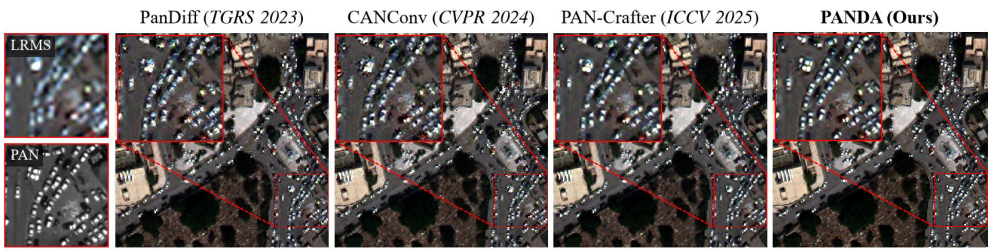


Figure 1: Visual comparison of PAN-sharpening results on the full-resolution WorldView-3 (WV3) dataset. The leftmost column shows corresponding LRMS and PAN crops, followed by reconstruction results from previous models, and PANDA. The red-boxed regions are enlarged in the upper-left insets for detailed comparison. In the highlighted region, PANDA produces clearer vehicle boundaries and more distinct separation between adjacent vehicles than the compared methods.

the spectral reference from its original coordinate frame. More recent alignment-aware approaches compensate for modality discrepancy inside the reconstruction backbone, after the two modalities have already been embedded into shared features (Do et al., 2025). These observations motivate a simple principle: *cross-modal correspondence should be established before joint feature extraction, while preserving the original MS coordinate frame*.

Reliable correspondence alone, however, does not fully determine how the aligned observations should be learned. In particular, reconstruction accuracy and alignment consistency reflect different aspects of the learned model. Accurate reconstruction does not necessarily imply a stable alignment response to spatial perturbations, and conversely, consistent alignment behavior does not guarantee accurate reconstruction. These two properties therefore provide complementary reliability cues. Reconstruction supervision should be guided by prediction quality, whereas geometry-sensitive supervision should depend on alignment consistency. This distinction suggests that PAN-sharpening requires not only *alignment before fusion*, but also *reliability-specific supervision*.

Based on these observations, we propose PANDA (**PAN** **A**lignment-Aware **D**istillation via **A**daptive Fitting), an align-first PAN-sharpening framework with reliability-aware adaptive supervision. As shown in Fig. 1, PANDA reconstructs clearer fine-scale structures. Our newly proposed PANDA first learns to correct PAN toward the corresponding MS coordinate frame before the two modalities are jointly processed. Our proposed PAN aligner is optimized together with the reconstruction network so that the estimated correction is directly guided by HRMS reconstruction. Because native displacement annotations are often unavailable, we additionally impose self-supervised shift consistency: when a known translation is applied to PAN while MS remains fixed, the predicted correction is encouraged to counteract the added displacement.This enables label-free task-relevant alignment.

After this alignment-learning stage, the trained PAN aligner and reconstruction network are frozen and used as a reference branch to guide a following compact student model. Instead of uniformly imitating the reference, PANDA adapts the student supervision according to two complementary reliability cues. Reconstruction error determines how reconstruction supervision is applied: ground-truth fitting is strengthened where the reference prediction is inaccurate, while reference imitation is emphasized only where the reference outperforms the student. In parallel, alignment-consistency error measures the stability of the reference aligner under controlled PAN translations and modulates geometry-sensitive supervision, including edge-based target fitting and refinement of the student aligner. In this way, reconstruction quality and alignment consistency guide the forms of supervision for which they are most relevant. At inference, only the student PAN aligner and compact reconstruction network are retained. Our main contributions are summarized as follows:

- We introduce an **align-first PAN-sharpening framework** that corrects PAN toward the corresponding MS coordinate frame before joint feature extraction. Reconstruction-guided optimization and self-supervised shift consistency jointly learn task-relevant correction without requiring displacement annotations.

- We propose **reliability-aware alignment distillation** that explicitly separates reconstruction reliability from geometric reliability. Reconstruction error selectively balances target

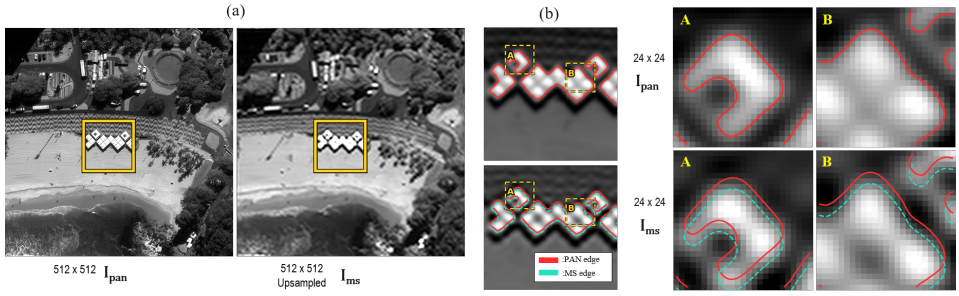


Figure 2: PAN–MS misalignment in full-resolution WorldView-3 imagery. (a) PAN (512×512) and MS bicubic-upsampled by $4\times$ to the PAN grid and averaged across its eight spectral bands. (b) Corresponding 112×112-px crops. Red curves are Otsu-level iso-contours extracted solely from PAN after blurring to MS resolution, overlaid at identical coordinates on both images to reveal misalignment. Right: 24×24-px details of A and B.

and reference supervision, while alignment consistency independently controls geometry-sensitive learning and aligner refinement.

- Extensive experiments on satellite datasets demonstrate that PANDA achieves state-of-the-art reduced- and full-resolution performance with a compact student model.

2 Related Work

2.1 PAN-Sharpening

Traditional PAN-sharpening methods are commonly categorized into component substitution (CS), multi-resolution analysis (MRA), and variational optimization (VO) (Meng et al., 2019). CS and MRA methods inject PAN-derived spatial information through component replacement (Choi et al., 2010) or multi-scale detail decomposition (Aiazzi et al., 2002), whereas VO methods formulate PAN-sharpening as an inverse problem with explicit observation models and priors (Fu et al., 2019). Deep learning has enabled end-to-end spatial–spectral reconstruction, ranging from direct and residual CNN-based models (Masi et al., 2016; Yang et al., 2017) to feature-fusion architectures (Wu et al., 2021). Transformer-based approaches further model long-range and cross-modal interactions (Bandara & Patel, 2022; Hou et al., 2024), while diffusion-based methods formulate PAN-sharpening as conditional iterative reconstruction (Meng et al., 2023; Xing et al., 2024; Zhong et al., 2024). Despite these advances, residual spatial misalignment between PAN and MS remains a fundamental challenge for faithful spatial–spectral reconstruction (Lee et al., 2021; Do et al., 2025).

PAN-MS Misalignment. Several approaches have been developed to mitigate spatial discrepancies between PAN and MS observations. Explicit feature-alignment methods (Lee et al., 2021) align MS features toward PAN structures and employ shift-invariant spectral supervision, effectively treating PAN as the geometric reference. Spatially adaptive convolutional approaches (Jin et al., 2022; Duan et al., 2024) improve local matching through adaptive filtering and non-local self-similarity, but primarily compensate for discrepancy during feature processing rather than explicitly correcting the input geometry. Recently, alignment-aware attention (Do et al., 2025) performs multi-scale cross-modal alignment inside the reconstruction backbone after initial feature embedding.

PANDA differs in both *where* and *how* alignment is introduced. Rather than aligning representations after cross-modal feature extraction or relocating MS information toward PAN, PANDA corrects PAN toward the MS coordinate frame before joint embedding. This align-first design prevents residual geometric discrepancy from being entangled with complementary spatial-spectral content during early feature extraction. Moreover, the learned alignment response is not used only for geometric correction, but also provides a reliability cue for subsequent geometry-sensitive supervision.

2.2 Adaptive Supervision and Knowledge Distillation

Adaptive supervision has been explored to account for spatially varying reconstruction difficulty and observation reliability. Unsupervised learning based PAN-sharpening adjusts spatial and spectral constraints according to local PAN-MS correspondence (Choi et al., 2019), while full-resolution

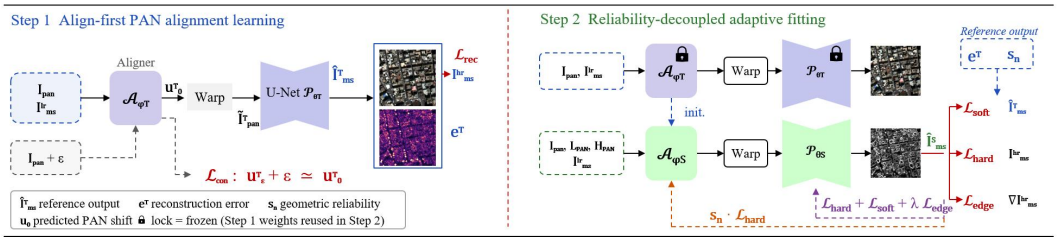


Figure 3: Overview of PANDA. After align-first learning, the frozen reference guides student fitting through two reliability cues: reconstruction error balances target/reference supervision, while alignment consistency regulates edge supervision and aligner refinement. Only the student branch is retained at inference.

learning derives self-supervised spatial and spectral objectives directly from the observed PAN and MS inputs (Ciotola et al., 2022). Target-adaptive approaches further optimize pretrained networks for individual PAN-MS observations at test time (Scarpa et al., 2018). These methods demonstrate the benefit of adapting supervision to the characteristics of the observed data, but generally do not distinguish between reconstruction reliability and geometric reliability.

Knowledge Distillation. Knowledge distillation (KD) transfers predictive or representational knowledge from a teacher to a student (Hinton et al., 2015) and has been widely applied to image restoration. Beyond conventional prediction matching, restoration-oriented KD has explored transferring structural and feature-level knowledge for image and depth super-resolution (Zhang et al., 2021; Sun et al., 2021). Reliability-aware learning further recognizes that different spatial regions should not contribute equally to optimization; for example, uncertainty-aware super-resolution objectives modulate supervision according to prediction uncertainty (Ning et al., 2021). In PAN-sharpening, U-Know-DiffPAN (Kim et al., 2025) uses teacher uncertainty to identify difficult regions and adapt teacher–student learning accordingly.

PANDA takes a different perspective by separating **reconstruction reliability** from **geometric reliability**. Reconstruction error determines whether learning should rely more strongly on the ground-truth target or on the reference prediction, whereas alignment consistency independently controls geometry-sensitive supervision and aligner refinement. Thus, rather than using a single reliability estimate to modulate distillation uniformly, PANDA assigns different reliability cues to the supervision routes for which they are relevant.

3 METHOD

3.1 OVERALL PIPELINE

As illustrated in Fig. 3, PANDA consists of two stages: *align-first PAN correction* and *reliability-decoupled adaptive fitting*. Given a PAN image $\mathbf{I}_{\text{pan}}$ and an upsampled MS image $\mathbf{I}_{\text{ms}}^{\text{lr}}$, the first stage establishes spatial correspondence before joint feature extraction. Specifically, proposed PAN aligner corrects PAN toward the original MS coordinate frame, and is jointly optimized with a reconstruction network using HRMS reconstruction and self-supervised consistency under controlled PAN translations. This allows PAN spatial information to be fused with MS spectral information.

The trained aligner and reconstruction network are then frozen and used as a reference branch for student fitting. Rather than treating the reference as uniformly reliable, PANDA derives two complementary cues with different roles: *reconstruction reliability*, which characterizes prediction quality, and *geometric reliability*, which is estimated from alignment consistency under controlled shifts. The former determines how target and reference predictions supervise reconstruction, whereas the latter controls geometry-sensitive supervision and aligner refinement. Thus, different forms of supervision are governed by the reliability signal relevant to their learning objective. At inference, only the final student aligner and reconstruction network are retained.

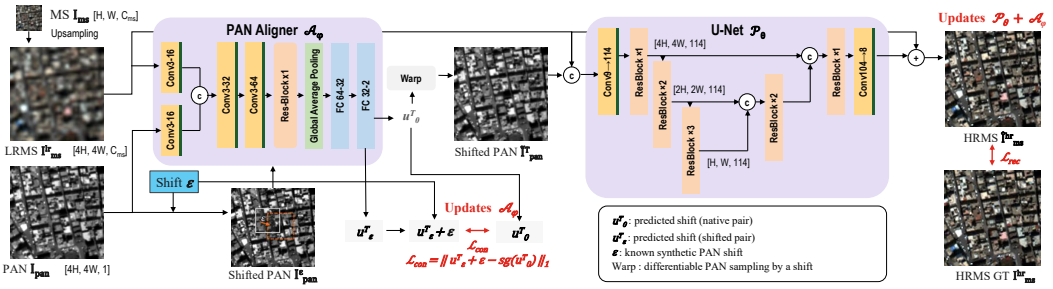


Figure 4: Align-first PAN alignment learning. The PAN aligner corrects PAN toward the MS coordinate frame before reconstruction, guided by reconstruction supervision and self-supervised shift consistency.

3.2 Align-First PAN Alignment Learning

PANDA first establishes PAN-MS correspondence before the two modalities are jointly embedded for reconstruction. As illustrated in Fig. 4, PANDA learns this correction directly from the PAN-sharpening objective while constraining the aligner's response to known spatial perturbations.

PAN Aligner for MS-Referenced Correction. The PAN aligner $\mathcal{A}_{\phi_T}$ takes $\mathbf{I}_{\text{pan}}$ and $\mathbf{I}^{\text{lr}}_{\text{ms}}$ as inputs and predicts a sample-wise 2-D translation $\mathbf{u}^{\text{T}}_0$. Separate convolutional branches first encode the two modalities, followed by feature concatenation, convolutional and residual processing, global average pooling, and fully connected regression. The predicted correction is applied to PAN through differentiable warping,

$$\tilde{\mathbf{I}}^{\text{T}}_{\text{pan}} = \text{Warp}(\mathbf{I}_{\text{pan}}, \mathbf{u}^{\text{T}}_0). \tag{1}$$

Importantly, only PAN is repositioned, while MS remains in its original coordinate frame. This preserves the observed MS geometry as the spectral reference during subsequent reconstruction.

Self-Supervised Shift Consistency. Native PAN-MS displacement is not directly observable because displacement annotations are often unavailable. Moreover, reconstruction supervision alone does not explicitly constrain how the aligner should respond when the relative PAN-MS geometry changes. We therefore introduce controlled PAN translations and enforce consistency of the predicted correction. Specifically, we apply a known synthetic translation $\boldsymbol{\epsilon}$ to PAN while keeping MS fixed. Given the shifted PAN, the same aligner predicts $\mathbf{u}^{\text{T}}_{\epsilon}$. Since the additional displacement is known, the predicted correction should counteract it while preserving the correction associated with the native pair: $\mathbf{u}^{\text{T}}_{\epsilon} + \boldsymbol{\epsilon} \simeq \mathbf{u}^{\text{T}}_0$. We define

$$\mathcal{L}_{\text{con}} = \text{mean}_n \left\| \mathbf{u}^{\text{T}}_{\epsilon,n} + \boldsymbol{\epsilon}_n - \text{sg}\left(\mathbf{u}^{\text{T}}_{0,n}\right) \right\|_1, \tag{2}$$

where n indexes training samples and $\text{sg}(\cdot)$ denotes stop-gradient. The synthetically shifted PAN is used only for consistency supervision and is not passed to the reconstruction network. This objective constrains the *relative response* of the aligner to controlled changes in PAN-MS geometry, without requiring native displacement labels.

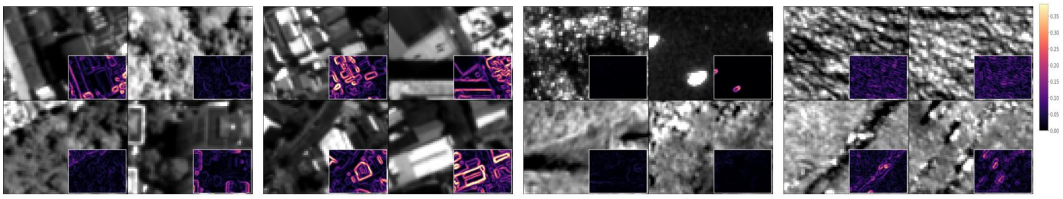
Reconstruction-Guided Alignment. For the native PAN-MS pair, the corrected PAN images $\tilde{\mathbf{I}}^{\text{T}}_{\text{pan}}$ and the corresponding upsampled MS images $\mathbf{I}^{\text{lr}}_{\text{ms}}$ are jointly processed by the reconstruction network $\mathcal{P}_{\theta_T}$. The network predicts an MS residual, yielding

$$\hat{\mathbf{I}}^{\text{T}}_{\text{ms}} = \mathbf{I}^{\text{lr}}_{\text{ms}} + \mathcal{P}_{\theta_T}\left(\tilde{\mathbf{I}}^{\text{T}}_{\text{pan}}, \mathbf{I}^{\text{lr}}_{\text{ms}}\right). \tag{3}$$

We optimize the reconstruction using

$$\mathcal{L}_{\text{rec}} = \text{mean}\left|\hat{\mathbf{I}}^{\text{T}}_{\text{ms}} - \mathbf{I}^{\text{hr}}_{\text{ms}}\right|. \tag{4}$$

Because the differentiable warp lies on the reconstruction path, gradients from $\mathcal{L}_{\text{rec}}$ jointly update the reconstruction network and PAN aligner. The overall alignment-learning objective is



(a) Consistent response / low reconstruction error (q↓, e↓) (b) Consistent response / high reconstruction error (q↓, e↑) (c) Inconsistent response / low reconstruction error (q↑, e↓) (d) Inconsistent response / high reconstruction error(q↑, e↑)

Figure 5: Reliability cues for adaptive supervision. Samples grouped by alignment-consistency error q_n exhibit varying reconstruction errors (insets). Reconstruction error controls target/reference fitting, while q_n separately regulates edge supervision and aligner refinement.

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}. \tag{5}$$

The $\mathcal{L}_{\text{rec}}$ anchors the native correction to spatial correspondence for PAN-sharpening, whereas $\mathcal{L}_{\text{con}}$ regularizes how this correction changes under known relative translations. Thus, the aligner learns task-relevant correspondence without direct displacement supervision.

3.3 Reliability-Decoupled Adaptive Fitting

After alignment learning, we freeze the trained reconstruction network and aligner as a reference branch. PANDA uses two complementary cues during student fitting: reconstruction error for prediction-oriented supervision and alignment consistency for geometry-sensitive supervision.

As shown in Fig. 3, the trained reconstruction network and aligner $(\mathcal{P}_{\theta_{\text{T}}}, \mathcal{A}_{\phi_{\text{T}}})$ are frozen and serve as a reference for the student $(\mathcal{P}_{\theta_{\text{S}}}, \mathcal{A}_{\phi_{\text{S}}})$. The student aligner is initialized from the reference aligner. The student reconstruction network additionally receives low- and high-frequency PAN components, denoted by LPAN and HPAN, respectively. This input setting is independent of the reliability weighting described below.

Reconstruction and Geometric Reliability Cues. Fig. 5 illustrates that reconstruction accuracy and alignment consistency do not necessarily coincide: consistent alignment responses can occur at both low and high reconstruction errors, while inconsistent responses can likewise occur at either error level. This motivates treating them as distinct signals.

Let $\hat{\mathbf{I}}^{\text{T}}_{\text{ms}}$ and $\hat{\mathbf{I}}^{\text{S}}_{\text{ms}}$ denote the reference and student model predictions. We define their channel-averaged absolute reconstruction errors at spatial position p and channel c of sample n as

$$\begin{aligned} e_{\text{T}}(n,p) &= \text{mean}_c \left| \hat{\mathbf{I}}^{\text{T}}_{\text{ms}}(n,p,c) - \mathbf{I}^{\text{hr}}_{\text{ms}}(n,p,c) \right|, \\ e_{\text{S}}(n,p) &= \text{mean}_c \left| \hat{\mathbf{I}}^{\text{S}}_{\text{ms}}(n,p,c) - \mathbf{I}^{\text{hr}}_{\text{ms}}(n,p,c) \right|. \end{aligned} \tag{6}$$

These errors characterize reconstruction quality. Geometric consistency is evaluated separately by probing the frozen reference aligner with a fixed set of controlled translations $\mathcal{E}$:

$$q_n = \text{mean}_{\boldsymbol{\epsilon}\in\mathcal{E}} \left\| \mathbf{u}^{\text{T}}_n(\boldsymbol{\epsilon}) + \boldsymbol{\epsilon} - \mathbf{u}^{\text{T}}_n(\mathbf{0}) \right\|_1. \tag{7}$$

A small q_n indicates that the reference correction changes consistently with known PAN translations. We emphasize that q_n does not measure absolute registration accuracy; instead, it provides a self-supervised cue for the stability of the reference geometry around the observed PAN-MS pair.

Reconstruction-Reliability Guided Distillation. We first determine how reconstruction supervision should be distributed between the HRMS target and the reference model prediction. Following the reference-student-target relation of (Hu et al., 2024), we define

$$d_{\text{T}} = \frac{e_{\text{T}}}{e_{\text{T}} + \tau_{\text{rec}}}, \qquad a_{\text{T}} = \frac{\max(e_{\text{S}} - e_{\text{T}}, 0)}{e_{\text{S}} + \delta_{\text{rec}}}, \tag{8}$$

where d_{T} measures reference difficulty and a_{T} measures the advantage of the reference over the student. τ_{rec} is a fixed training-set error scale and $\delta_{\mathrm{rec}} > 0$ is a numerical stabilizer. The per-sample hard and soft reconstruction objectives are

$$\begin{aligned} \ell_n^{\mathrm{hard}} &= \mathrm{mean}_p \left[(1 + \lambda_{\mathrm{hard}} d_{\mathrm{T}}) e_{\mathrm{S}} \right], \\ \ell_n^{\mathrm{soft}} &= \mathrm{mean}_p \left[\lambda_{\mathrm{soft}} (1 - d_{\mathrm{T}}) a_{\mathrm{T}} \Delta_{\mathrm{TS}} \right], \end{aligned} \tag{9}$$

where Δ_{TS} denotes the channel-averaged absolute difference between the reference and student predictions Direct target supervision is retained for all spatial locations and is strengthened where the reference itself has high reconstruction error. In contrast, reference guidance is emphasized only when the reference is accurate and provides an advantage over the current student. Thus, reconstruction reliability determines *which reconstruction target* should provide stronger supervision rather than forcing uniform reference guidance.

Geometric-Reliability Guided Spatial Fitting. We next use alignment consistency to regulate geometry-sensitive supervision. The consistency error q_n is converted into a reliability weight

$$s_n = \frac{q_{\mathrm{ref}}}{q_{\mathrm{ref}} + q_n}, \tag{10}$$

where q_{ref} is the median consistency error measured over the training set. Lower consistency error therefore yields a larger s_n. We interpret s_n as a proxy for geometric reliability.

Let ℓ_n^{edge} denote the absolute difference between the signed horizontal and vertical gradients of the student prediction and HRMS target, averaged over spectral bands and valid spatial positions. The geometry-sensitive edge objective is

$$\mathcal{L}_{\mathrm{edge}} = \mathrm{mean}_n \left[s_n \ell_n^{\mathrm{edge}} \right]. \tag{11}$$

The same geometric reliability also controls refinement of the student aligner. Because displacement labels are unavailable, the aligner is updated through the reconstruction objective, but its contribution is reduced when the reference geometry exhibits inconsistent responses. The reconstruction network and aligner are therefore optimized separately as

$$\begin{aligned} \mathcal{L}_{\mathcal{P}} &= \mathrm{mean}_n \left(\ell_n^{\mathrm{hard}} + \ell_n^{\mathrm{soft}} \right) + \lambda_{\mathrm{edge}} \mathcal{L}_{\mathrm{edge}}, \\ \mathcal{L}_{\mathcal{A}} &= \mathrm{mean}_n \left[s_n \ell_n^{\mathrm{hard}} \right]. \end{aligned} \tag{12}$$

Only $\mathcal{L}_{\mathcal{P}}$ updates the reconstruction parameters θ_{S}, whereas $\mathcal{L}_{\mathcal{A}}$ updates the aligner parameters ϕ_{S}. Accordingly, reference guidance and edge supervision do not directly modify the aligner. This decoupled design ensures that reconstruction reliability governs reconstruction knowledge transfer, while geometric reliability governs geometry-sensitive learning.

4 Experiments

4.1 Experimental Setup

We evaluate PANDA on three datasets from PanCollection (Deng et al., 2022): WorldView-3 (WV3), QuickBird (QB), and GaoFen-2 (GF2). WV3 provides eight MS bands, whereas QB and GF2 provide four. Their native PAN/MS spatial resolutions are 0.3/1.2 m, 0.6/2.4 m, and 0.8/3.2 m, respectively. The corresponding training sets contain 9,714, 17,139, and 19,809 samples, with models trained separately for each sensor. Training uses 64×64 PAN and 16×16 MS patches, maintaining a $4\times$ spatial scale ratio. Each dataset includes 20 reduced-resolution (RR) and 20 full-resolution (FR) test samples, with PAN sizes of 256×256 and 512×512, respectively. RR evaluation follows Wald's protocol with available HRMS references (Wald, 2000), whereas FR evaluation uses native-resolution observations without corresponding HRMS references. For RR evaluation, we report PSNR, SSIM (Wang et al., 2004), SAM (Yuhas et al., 1992), ERGAS (Wald, 2000), SCC (Zhou et al., 1998), and Q_4/Q_8 (Garzelli & Nencini, 2009) against HRMS references, using Q_8 for WV3

Table 1: Quantitative comparison on the WV3 dataset. Full-resolution (FR) and reduced-resolution (RR) results are reported together with computational cost. Red and blue denote the best and second-best results, respectively.

WV3 Dataset		Full-Resolution			Reduced-Resolution						Cost	
Methods	Publications	HQNR↑	D_s ↓	D_λ ↓	ERGAS↓	SCC↑	SAM↓	Q8↑	PSNR↑	SSIM↑	Infer (s)↓	FLOPs (G)↓
PanNet(Yang et al., 2017)	ICCV 2017	0.918	0.049	0.035	2.538	0.979	3.402	0.913	36.148	0.966	0.001	5.17
MSDCNN(Yuan et al., 2018)	JSTARS 2018	0.924	0.050	0.028	2.489	0.979	3.300	0.914	36.329	0.967	0.002	14.96
FusionNet(Wu et al., 2021)	ICCV 2021	0.920	0.053	0.029	2.428	0.981	3.188	0.916	36.569	0.968	0.002	5.13
LAGNet(Jin et al., 2022)	AAAI 2022	0.915	0.055	0.033	2.380	0.981	3.153	0.916	36.732	0.970	0.004	8.43
S2DBPN(Zhang et al., 2023)	TGRS 2023	0.946	0.030	0.025	2.245	0.985	3.019	0.917	37.216	0.972	0.005	158.94
PanDiff(Meng et al., 2023)	TGRS 2023	0.952	0.034	0.014	2.276	0.984	3.058	0.913	37.029	0.971	2.955	62.07
DCPNet(Zhang et al., 2024)	TGRS 2024	0.923	0.036	0.043	2.301	0.984	3.083	0.915	37.009	0.972	0.109	105.40
TMDiff(Xing et al., 2024)	TGRS 2024	0.924	0.059	0.018	2.151	0.986	2.885	0.915	37.477	0.973	9.997	1284.42
CANConv(Duan et al., 2024)	CVPR 2024	0.951	0.030	0.020	2.163	0.985	2.927	0.918	37.441	0.973	0.451	52.21
PAN-Crafter(Do et al., 2025)	ICCV 2025	0.950	0.026	0.025	2.305	0.983	2.838	0.915	37.451	0.971	0.017	313.93
PANDA (Ours)	–	0.959	0.026	0.016	2.056	0.988	2.787	0.921	37.951	0.976	0.005	57.10

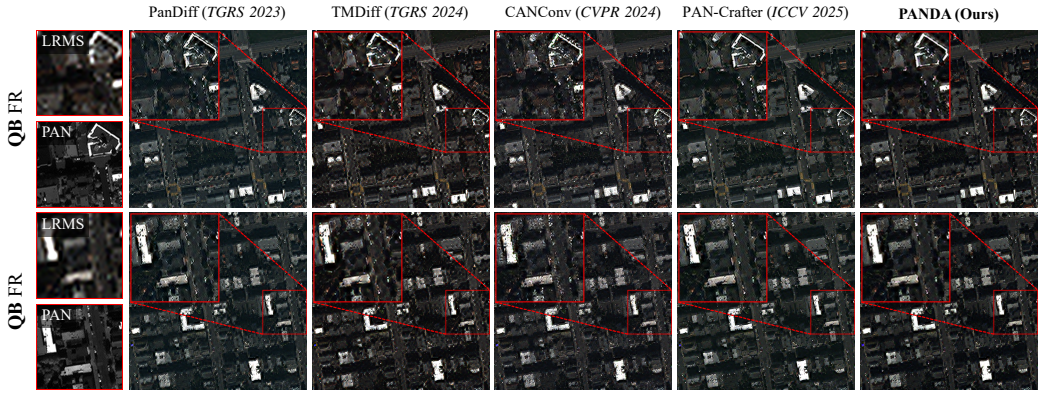


Figure 6: Qualitative comparison on the full-resolution QuickBird (QB) dataset. Each row shows a different scene, with corresponding LRMS and PAN crops in the leftmost column. Red-boxed regions are enlarged in the upper-left insets. PANDA produces clearer, more continuous building outlines with fewer local distortions in the highlighted regions.

and Q_4 for QB/GF2. For FR evaluation, where HRMS references are unavailable, we use HQNR, spatial distortion D_s, and spectral distortion D_λ (Vivone et al., 2020). These metrics are computed using the PanCollection implementations (Deng et al., 2022).

Implementation Details. We implement PANDA in PyTorch (Paszke et al., 2017) on a single NVIDIA RTX 5090 GPU. Each stage is trained for 50,000 updates using AdamW (Loshchilov & Hutter, 2017) with a batch size of 48. Peak learning rates are 10^{-4} for the reconstruction networks and 10^{-5} and 3×10^{-6} for the reference and student aligners, respectively, with cosine decay (Loshchilov & Hutter, 2016). We set $\lambda_{\text{hard}} = 1$, $\lambda_{\text{soft}} = 0.1$, and $\lambda_{\text{edge}} = 0.002$.

4.2 Experimental Results

Quantitative Comparison. Table 1 compares PANDA with existing PAN-sharpening methods on WV3 under FR and RR settings. PANDA achieves the highest HQNR of 0.959, ties with PAN-Crafter for the lowest spatial distortion D_s of 0.026, and attains the second-lowest spectral distortion D_λ of 0.016. Under RR evaluation, PANDA achieves the best results across all six reported metrics, including a PSNR of 37.951 dB and a SAM of 2.787, demonstrating improved spatial–spectral reconstruction fidelity. These gains are accompanied by lower computational cost: PANDA requires 57.10 GFLOPs and 0.005 s per inference, corresponding to approximately 81.8% fewer FLOPs and a $3.4\times$ inference speedup over PAN-Crafter based on the reported measurements.

Table 2 presents results on QB and GF2. On QB, PANDA achieves the highest HQNR and the best results across all reported RR metrics, sharing the highest SCC with DCPNet, while PAN-Crafter achieves a lower D_s. On GF2, PANDA achieves the highest HQNR of 0.965 and the lowest D_s of 0.016. It also matches the highest SCC, while ranking second to PAN-Crafter in ERGAS, SAM, and PSNR. Overall, PANDA achieves the highest HQNR across all three datasets, alongside leading RR performance on WV3 and QB and competitive RR fidelity on GF2.

Table 2: Quantitative comparison on the QB and GF2 datasets. Full-resolution (FR) and reduced-resolution (RR) results are reported for each sensor. Red and blue denote the best and second-best results, respectively.

	QB						GF2					
	Full-Resolution		Reduced-Resolution				Full-Resolution		Reduced-Resolution			
Methods	HQNR↑	D_s ↓	ERGAS↓	SCC↑	SAM↓	PSNR↑	HQNR↑	D_s ↓	ERGAS↓	SCC↑	SAM↓	PSNR↑
PanNet(Yang et al., 2017)	0.851	0.092	4.856	0.966	5.273	35.563	0.929	0.052	1.038	0.975	1.050	39.197
MSDCNN(Yuan et al., 2018)	0.888	0.058	4.074	0.977	4.828	37.040	0.898	0.079	0.862	0.983	0.946	40.730
FusionNet(Wu et al., 2021)	0.853	0.079	4.183	0.975	4.892	36.821	0.865	0.105	0.960	0.980	0.971	39.866
LAGNet(Jin et al., 2022)	0.892	0.035	3.845	0.980	4.682	37.565	0.895	0.078	0.816	0.985	0.886	41.147
S2DBPN(Zhang et al., 2023)	0.908	0.036	3.956	0.980	4.849	37.314	0.935	0.046	0.686	0.990	0.772	42.686
PanDiff(Meng et al., 2023)	0.919	0.055	3.723	0.982	4.611	37.842	0.936	0.045	0.674	0.990	0.767	42.827
DCPNet(Zhang et al., 2024)	0.880	0.073	3.618	0.983	4.420	38.079	0.953	0.024	0.724	0.988	0.806	42.312
TMDiff(Xing et al., 2024)	0.901	0.068	3.804	0.981	4.627	37.642	0.942	0.030	0.754	0.988	0.764	41.896
CANConv(Duan et al., 2024)	0.893	0.070	3.740	0.982	4.554	37.795	0.919	0.063	0.653	0.991	0.722	43.166
PAN-Crafter(Do et al., 2025)	0.923	0.025	4.425	0.972	4.626	36.756	0.938	0.035	0.536	0.994	0.611	44.712
PANDA (Ours)	0.925	0.032	3.561	0.983	4.383	38.323	0.965	0.016	0.551	0.994	0.625	44.542

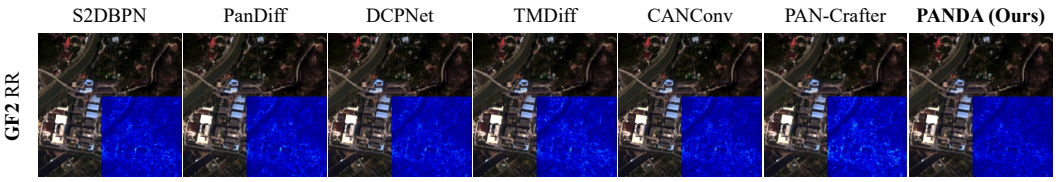


Figure 7: Qualitative comparison on the reduced-resolution GaoFen-2 (GF2) dataset. Insets show reconstruction errors against the HRMS ground truth. PANDA preserves fine spatial details with relatively low residual errors.

Table 3: Ablation study on the WV3. Components are added in turn, following the two steps of PANDA: align-first PAN alignment learning (Step I) and reliability-decoupled adaptive fitting (Step II).

Configuration		Full-Resolution		Reduced-Resolution				Cost	
Component	Step	HQNR↑	D_s ↓	ERGAS↓	SCC↑	SAM↓	PSNR↑	Infer (s)↓	FLOPs (G)↓
Reconstruction baseline (PAN + MS)	I	0.951	0.027	2.054	0.988	2.781	37.949	0.005	56.60
+ LPAN/HPAN frequency inputs	I	0.951	0.031	**2.044**	0.988	**2.774**	**37.996**	0.005	56.60
+ PAN aligner, trained from scratch	I	0.952	0.031	2.050	0.988	2.776	37.974	0.005	57.10
+ self-supervised shift consistency	I	0.954	0.029	2.050	0.988	2.781	37.967	0.006	57.10
+ reconstruction-reliability guided distillation	II	0.958	**0.021**	2.056	0.988	2.792	37.941	0.005	57.10
+ geometric-reliability guided spatial fitting (Ours)	II	**0.959**	0.026	2.056	0.988	2.787	37.951	0.005	57.10

Qualitative Comparison. We compare PANDA with representative methods on QB at FR (Fig. 6) and GF2 at RR (Fig. 7). In the enlarged QB regions, PANDA produces clearer, more continuous building outlines, whereas several competing methods exhibit blurred boundaries or locally distorted structures. The GF2 example further shows well-defined spatial details and relatively subdued residuals in the error-map inset, indicating competitive reconstruction fidelity under RR evaluation.

Ablation Study. Table 3 evaluates PANDA's components on WV3. LPAN/HPAN inputs improve RR accuracy without changing HQNR. PAN alignment with shift consistency raises HQNR to 0.954, and reconstruction-reliability guided distillation further improves it to 0.958 while reducing D_s to 0.021. Geometric-reliability guided spatial fitting achieves the highest HQNR of 0.959, although D_s increases to 0.026. Overall, PANDA improves HQNR with comparable RR accuracy.

5 Conclusion

We presented PANDA, an align-first PAN-sharpening framework with reliability-aware supervision. PANDA corrects PAN toward the original MS coordinate frame before joint feature extraction through reconstruction guidance and self-supervised shift consistency, without displacement annotations. A frozen reference guides a compact student: reconstruction error balances target and reference supervision, while alignment consistency regulates geometry-sensitive learning and aligner refinement. Experiments demonstrate improved full-resolution performance, competitive reduced-resolution fidelity, and efficient inference.

AI USE STATEMENT

In this work, we did not use generative AI tools for any tasks requiring mandatory disclosure. Additionally, we used generative AI tools to aid and polish the writing of this paper, including improving grammar, clarity, and readability. All AI-assisted writing was reviewed and revised by the authors to ensure that it accurately reflects the intended technical content. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Bruno Aiazzi, Luciano Alparone, Stefano Baronti, and Andrea Garzelli. Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis. *IEEE Transactions on geoscience and remote sensing*, 40(10):2300–2312, 2002.

Wele Gedara Chaminda Bandara and Vishal M Patel. Hypertransformer: A textural and spectral feature fusion transformer for pansharpening. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1757–1767. IEEE, 2022.

Jae-Seok Choi, Yongwoo Kim, and Munchurl Kim. S3: A spectral-spatial structure loss for pansharpening networks. *IEEE Geoscience and Remote Sensing Letters*, 17(5):829–833, 2019.

Jaewan Choi, Kiyun Yu, and Yongil Kim. A new adaptive component-substitution-based satellite image fusion by using partial replacement. *IEEE transactions on geoscience and remote sensing*, 49(1):295–309, 2010.

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.

Liang-Jian Deng, Gemine Vivone, Mercedes E Paoletti, Giuseppe Scarpa, Jiang He, Yongjun Zhang, Jocelyn Chanussot, and Antonio Plaza. Machine learning in pansharpening: A benchmark, from shallow to deep networks. *IEEE Geoscience and Remote Sensing Magazine*, 10(3):279–315, 2022.

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

Yule Duan, Xiao Wu, Haoyu Deng, and Liang-Jian Deng. Content-adaptive non-local convolution for remote sensing pansharpening. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27738–27747. IEEE, 2024.

Xueyang Fu, Zihuang Lin, Yue Huang, and Xinghao Ding. A variational pan-sharpening with local gradient constraints. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10257–10266. IEEE, 2019.

Andrea Garzelli and Filippo Nencini. Hypercomplex quality assessment of multi/hyperspectral images. *IEEE Geoscience and Remote Sensing Letters*, 6(4):662–665, 2009.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Junming Hou, Zihan Cao, Naishan Zheng, Xuan Li, Xiaoyu Chen, Xinyang Liu, Xiaofeng Cong, Danfeng Hong, and Man Zhou. Linearly-evolved transformer for pan-sharpening. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 1486–1494, 2024.

Chengming Hu, Haolun Wu, Xuan Li, Chen Ma, Xi Chen, Boyu Wang, Jun Yan, and Xue Liu. Less or more from teacher: Exploiting trilateral geometry for knowledge distillation. In *International Conference on Learning Representations*, volume 2024, pp. 24503–24518, 2024.

Zi-Rong Jin, Tian-Jing Zhang, Tai-Xiang Jiang, Gemine Vivone, and Liang-Jian Deng. Lagconv: Local-context adaptive convolution kernels with global harmonic bias for pansharpening. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 1113–1121, 2022.

Sungpyo Kim, Jeonghyeok Do, Jaehyup Lee, and Munchurl Kim. U-know-diffpan: An uncertainty-aware knowledge distillation diffusion framework with details enhancement for pan-sharpening. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23069–23079. IEEE, 2025.

Jaehyup Lee, Soomin Seo, and Munchurl Kim. Sipsa-net: Shift-invariant pan sharpening with moving object alignment for satellite imagery. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10161–10169. IEEE, 2021.

Jun Li, Yanqiu Pei, Shaohua Zhao, Rulin Xiao, Xiao Sang, and Chengye Zhang. A review of remote sensing for environmental monitoring in china. *Remote Sensing*, 12(7):1130, 2020.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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

Qingyan Meng, Wenxu Shi, Sijia Li, and Linlin Zhang. Pandiff: A novel pansharpening method based on denoising diffusion probabilistic model. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–17, 2023.

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, 2019.

Qian Ning, Weisheng Dong, Xin Li, Jinjian Wu, and Guangming Shi. Uncertainty-driven loss for single image super-resolution. *Advances in Neural Information Processing Systems*, 34:16398–16409, 2021.

Xavier Otazu, María González-Audícana, Octavi Fors, and Jorge Núñez. Introduction of sensor spectral response into image fusion methods. application to wavelet-based methods. *IEEE Transactions on Geoscience and Remote Sensing*, 43(10):2376–2385, 2005.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, Adam Lerer, et al. Automatic differentiation in pytorch. 2017.

Giuseppe Scarpa, Sergio Vitale, and Davide Cozzolino. Target-adaptive cnn-based pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 56(9):5443–5457, 2018.

Baoli Sun, Xinchen Ye, Baopu Li, Haojie Li, Zhihui Wang, and Rui Xu. Learning scene structure guidance via cross-task knowledge transfer for single depth super-resolution. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7788–7797. IEEE, 2021.

Claire Thomas, Thierry Ranchin, Lucien Wald, and Jocelyn Chanussot. Synthesis of multispectral images to high spatial resolution: A critical review of fusion methods based on remote sensing physics. *IEEE Transactions on Geoscience and Remote Sensing*, 46(5):1301–1312, 2008.

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, 2014.

Gemine Vivone, Mauro Dalla Mura, Andrea Garzelli, Rocco Restaino, Giuseppe Scarpa, Magnus O Ulfarsson, Luciano Alparone, and Jocelyn Chanussot. A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods. *IEEE Geoscience and Remote Sensing Magazine*, 9(1):53–81, 2020.

Lucien Wald. Quality of high resolution synthesised images: Is there a simple criterion? In *Third conference" Fusion of Earth data: merging point measurements, raster maps and remotely sensed images"*, pp. 99–103. SEE/URISCA, 2000.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Thilo Wellmann, Angela Lausch, Erik Andersson, Sonja Knapp, Chiara Cortinovis, Jessica Jache, Sebastian Scheuer, Peleg Kremer, André Mascarenhas, Roland Kraemer, et al. Remote sensing in urban planning: Contributions towards ecologically sound policies? *Landscape and urban planning*, 204:103921, 2020.

Xiao Wu, Ting-Zhu Huang, Liang-Jian Deng, and Tian-Jing Zhang. Dynamic cross feature fusion for remote sensing pansharpening. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14667–14676. IEEE, 2021.

Yinghui Xing, Litao Qu, Shizhou Zhang, Jiapeng Feng, Xiuwei Zhang, and Yanning Zhang. Empower generalizability for pansharpening through text-modulated diffusion model. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024.

Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, and John Paisley. Pannet: A deep network architecture for pan-sharpening. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 1753–1761. IEEE, 2017.

Qiangqiang Yuan, Yancong Wei, Xiangchao Meng, Huanfeng Shen, and Liangpei Zhang. A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(3):978–989, 2018.

Roberta H Yuhas, Alexander FH Goetz, and Joe W Boardman. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. In *JPL, summaries of the third annual JPL airborne geoscience workshop. volume 1: AVIRIS workshop*, 1992.

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.

Yafei Zhang, Xuji Yang, Huafeng Li, Minghong Xie, and Zhengtao Yu. Dcpnet: A dual-task collaborative promotion network for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024.

Yiman Zhang, Hanting Chen, Xinghao Chen, Yiping Deng, Chunjing Xu, and Yunhe Wang. Data-free knowledge distillation for image super-resolution. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7848–7857. IEEE, 2021.

Yongjun Zhang, Chi Liu, Mingwei Sun, and Yangjun Ou. Pan-sharpening using an efficient bidirectional pyramid network. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8):5549–5563, 2019.

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.

Jie Zhou, Daniel L Civco, and John A Silander. A wavelet transform method to merge landsat tm and spot panchromatic data. *International journal of remote sensing*, 19(4):743–757, 1998.

A Appendix