PAN-Crafter: Learning Modality-Consistent Alignment for PAN-Sharpening

Jeonghyeok Do[1] Sungpyo Kim[1] Geunhyuk Youk[1] Jaehyup Lee[2†] Munchurl Kim[1†]
[1]KAIST [2]Kyungpook National University
{ehwjdgur0913, ksp04204, rmsgurkjg, mkimee}@kaist.ac.kr jaehyuplee@knu.ac.kr
https://kaist-viclab.github.io/PAN-Crafter_site

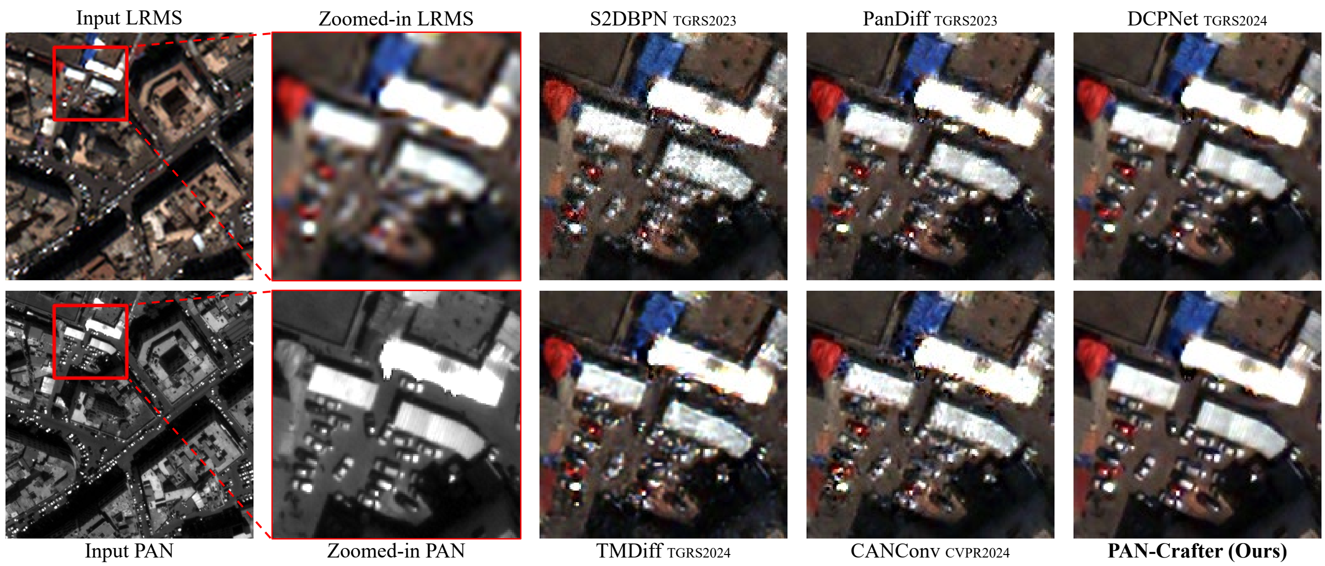

Figure 1. Comparison of PAN-sharpening (PS) results on the WV3 dataset at full-resolution for very recent methods and our PAN-Crafter. The *top-left* image shows the input low-resolution multi-spectral (LRMS) image with a zoomed-in region for better visualization. The *bottom-left* image represents the corresponding panchromatic (PAN) image. The remaining results shows the zoomed-in PAN-sharped patches that are their corresponding restored high-resolution multi-spectral (HRMS) ones. Our proposed PAN-Crafter produces pan-sharpened images with minimal artifacts, especially near buildings and cars, while other approaches often yield blurred or distorted results.

Abstract

PAN-sharpening aims to fuse high-resolution panchromatic (PAN) images with low-resolution multi-spectral (MS) images to generate high-resolution multi-spectral (HRMS) outputs. However, cross-modality misalignment—caused by sensor placement, acquisition timing, and resolution disparity—induces a fundamental challenge. Conventional deep learning methods assume perfect pixel-wise alignment and rely on per-pixel reconstruction losses, leading to spectral distortion, double edges, and blurring when misalignment is present. To address this, we propose PAN-Crafter, a modality-consistent alignment framework that explicitly mitigates the misalignment gap between PAN and MS modalities. At its core, Modality-Adaptive Reconstruction (MARs) enables a single network to jointly reconstruct HRMS and PAN images, leveraging PAN's high-frequency details as auxiliary self-supervision. Additionally, we introduce Cross-Modality Alignment-Aware Attention (CM3A), a novel mechanism that bidirectionally aligns MS texture to PAN structure and vice versa, enabling adaptive feature refinement across modalities. Extensive experiments on multiple benchmark datasets demonstrate that our PAN-Crafter outperforms the most recent state-of-the-art method in all metrics, even with 50.11× faster inference time and 0.63× the memory size. Furthermore, it demonstrates strong generalization performance on unseen satellite datasets, showing its robustness across different conditions.

1. Introduction

Remote sensing imagery is crucial for a wide range of applications, including environmental monitoring, defense intelligence, and urban planning [11, 23, 32, 49, 53, 56, 59].

†Co-corresponding authors (equal advising).

Many of these tasks require high-resolution images that preserve both fine spatial details and rich spectral information. However, due to inherent limitations in sensor technologies, a single imaging system cannot simultaneously achieve high spatial resolution and high spectral fidelity. To overcome this constraint, modern Earth observation satellites employ dual-sensor systems, consisting of a high-resolution panchromatic (PAN) sensor and a low-resolution multi-spectral (LRMS) sensor.

PAN-sharpening [9, 26, 31, 42, 45, 65] aims to fuse high-resolution PAN images with low-resolution MS images to produce high-resolution multi-spectral (HRMS) outputs. The goal is to retain the spectral fidelity of MS images while preserving the spatial details of PAN images. However, a fundamental challenge in this process is cross-modality misalignment, which arises from differences in sensor placement, acquisition timing, and resolution disparity. As illustrated in Fig. 2, PAN images typically have four times ($4H \times 4W$) the spatial resolution ($H \times W$) of MS images, necessitating up-sampling on the MS images before fusion. However, this up-sampling step introduces interpolation artifacts and spatial shifts, further amplifying alignment discrepancies. Most existing PAN-sharpening methods [13, 30, 37, 47, 50, 62, 65] assume perfect pixel-wise alignment and rely on per-pixel reconstruction losses, such as ℓ_1 and ℓ_2, leading to spectral distortion, double edges, and blurring when misalignment is present. To address these issues, several approaches [13, 20, 22] integrate spatial-adaptive convolutional layers to mitigate misalignment. Despite these efforts, existing approaches lack the capability to dynamically adapt to varying levels of misalignment across different datasets. Fixed-scale alignment mechanisms fail to capture complex spatial shifts [22], while self-similarity-based feature aggregation does not explicitly correct geometric discrepancies between PAN and MS images [13, 20]. Effective PAN-sharpening requires a solution that not only aligns textures and structures across modalities but also ensures modality-consistent feature refinement at multiple spatial scales.

To overcome these limitations, we propose PAN-Crafter, a modality-consistent alignment framework designed to handle cross-modality misalignment during the fusion process. Unlike existing methods that assume strict pixel-wise alignment, PAN-Crafter enables robust learning from misaligned PAN-MS pairs by jointly reconstructing HRMS and PAN images, ensuring structural consistency through feature alignment. Our key innovation is Modality-Adaptive Reconstruction (MARs), which allows a single network to dynamically generate both HRMS and PAN images based on a modality selection mechanism, MARs mode. Given LRMS and PAN as inputs, the network reconstructs HRMS in MS mode and PAN in PAN mode, using PAN's sharpness as an auxiliary self-supervision signal to enhance spatial fidelity. Furthermore, we introduce Cross-Modality Alignment-Aware Attention (CM3A), a novel mechanism that explicitly (i) aligns MS image textures to PAN image structures during HRMS reconstructions and (ii) matches PAN image structures to MS image textures during PAN back-reconstructions. This bidirectional interaction enables adaptive compensation for misalignment while preserving both spatial and spectral integrity. Our main contributions are as follows:

(a) PAN image $\mathbf{I}_{\text{pan}}$ (b) LRMS image $\mathbf{I}_{\text{ms}}^{\text{lr}}$ (c) Overlay of $\mathbf{I}_{\text{pan}}$ and $\mathbf{I}_{\text{ms}}^{\text{lr}}$

Figure 2. Example of PAN, LRMS, and their overlayed visualization. (a) High-resolution PAN image $\mathbf{I}_{\text{pan}}$, (b) Low-resolution multi-spectral image $\mathbf{I}_{\text{ms}}^{\text{lr}}$, (c) Overlay of PAN and LRMS images to highlight differences. The red insets provide zoomed-in views for better visualization.

- We propose Modality-Adaptive Reconstruction (MARs), a unified reconstruction framework that enables robust learning from misaligned PAN-MS image pairs by dynamically generating both HRMS and PAN images;
- We introduce Cross-Modality Alignment-Aware Attention (CM3A), a novel alignment mechanism that adaptively refines textures and structures between PAN and MS images, improving spatial-spectral consistency;
- We achieve state-of-the-art (SOTA) performance across multiple benchmark datasets and show strong robustness on unseen satellite datasets, demonstrating the effectiveness of our PAN-Crafter in handling real-world cross-modality misalignment.

2. Related Work

2.1. PAN-Sharpening

Traditional approaches. Traditional PAN-sharpening (PS) methods are broadly classified into component substitution (CS) [6, 7, 38, 40], multi-resolution analysis (MRA) [1, 2, 34, 39], and variational optimization (VO) [3, 5, 14, 51].

Deep learning-based approaches. Recent advancements in PAN-sharpening have been driven by deep learning-based methods [4, 13, 17, 18, 37, 62, 65], primarily leveraging Convolutional Neural Networks (CNNs) [33]. CNN-based models [13, 20, 22, 29, 29, 50, 55, 57, 58, 61, 62] are effective in capturing local spatial-spectral dependencies while maintaining relatively low computational complexity. For instance, S2DBPN [61] introduces a spatial-spectral back-projection framework to iteratively refine high-resolution outputs, while DCPNet [62] formulates a dual-task learning strategy that integrates PAN-sharpening with super-resolution. More recently, diffusion-based models [21, 30, 52, 64] have emerged, leveraging iterative denoising processes to refine reconstructed images. While diffusion models [16] improve generation quality, they suffer from excessive computational costs, limiting their practical deployment in real-world applications.

Cross-modality misalignment handling. Most existing PAN-sharpening methods assume perfect pixel-wise alignment and optimize reconstruction using per-pixel losses. However, real-world PAN-MS pairs often exhibit spatial misalignment, leading to spectral distortion and double-edge artifacts. To address this, several works have explored spatial-adaptive feature alignment strategies. SIPSA [22] explicitly identifies misalignment as a critical challenge in PAN-sharpening and introduces a spatially-adaptive module, but its fixed-scale alignment mechanism limits flexibility when handling diverse misalignment patterns. LAGConv [20] and CANConv [13] adopt non-local spatial-adaptive convolution modules that enhance feature consistency through self-similarity. However, these methods primarily focus on semantic feature aggregation rather than explicit geometric alignment, making them suboptimal for correcting cross-modality shifts.

2.2. Spatial-Adaptive Operation

PAN-sharpening datasets are generally pre-aligned, but local misalignment persists due to sensor inconsistencies, parallax effects, and resolution differences. Addressing these spatial shifts requires spatial-adaptive operations that dynamically adjust processing based on input features, enabling local refinement while maintaining structural consistency. Several adaptive techniques have been proposed to tackle local misalignment by modulating computations according to spatial context. Pixel-adaptive convolution [41] dynamically adjusts convolutional weights based on local pixel intensity, enabling spatially adaptive filtering. Deformable convolutions [8] learn spatially adaptive offsets, shifting receptive fields to enhance feature extraction. However, these methods primarily focus on intra-modality feature refinement and lack explicit mechanisms to enforce cross-modality alignment.

Attention mechanisms. Attention operations [12, 19, 24, 35, 43, 54, 63] are inherently non-local and spatially adaptive, enabling dynamic feature aggregation across spatial regions. Self-attention mechanisms capture long-range dependencies by computing pairwise feature relationships, while cross-attentions extend this by integrating information across different modalities. To incorporate spatial priors, positional embeddings [25, 43] are typically added to query and key representations, encoding relative pixel positions that guide feature interactions. However, in cross-modality settings, fixed positional embeddings fail to handle local misalignment, as real-world distortions vary across datasets and cannot be effectively modeled by static spatial priors.

To address these limitations, we replace fixed positional embeddings with modality-aware feature priors directly derived from PAN and MS representations. Instead of encoding positional information explicitly, our approach leverages cross-modality feature interactions by embedding PAN features into queries ($\mathbf{Q}$) and keys ($\mathbf{K}$) when attending to MS, and vice versa. This design enables the PS networks to dynamically adapt spatial attention to local misalignment patterns, ensuring precise feature alignment without reliance on predefined spatial encodings.

3. Methods

3.1. Overview of the Proposed PAN-Crafter

Given a paired dataset $\mathcal{D} = \{(\mathbf{I}_{\text{pan}}, \mathbf{I}_{\text{ms}}, \mathbf{I}_{\text{ms}}^{\text{hr}})\}$, where $\mathbf{I}_{\text{pan}} \in \mathbb{R}^{4H\times 4W\times 1}$ represents a PAN image, and $\mathbf{I}_{\text{ms}} \in \mathbb{R}^{H\times W\times C_{\text{ms}}}$ denotes an MS image with C_{ms} spectral bands. Since the MS image has a lower spatial resolution, we first up-sample it by a factor of 4, obtaining a LRMS $\mathbf{I}_{\text{ms}}^{\text{lr}} \in \mathbb{R}^{4H\times 4W\times C_{\text{ms}}}$, which serves as an initial estimate of HRMS target $\mathbf{I}_{\text{ms}}^{\text{hr}} \in \mathbb{R}^{4H\times 4W\times C_{\text{ms}}}$.

Modality-Adaptive Reconstruction (MARs). Our goal is to learn a PAN-Crafter network $\mathcal{P}_\theta$ that synthesizes a HRMS image $\mathbf{I}_{\text{ms}}^{\text{hr}}$ from the given PAN and LRMS inputs while explicitly handling cross-modality misalignment. To achieve this, we introduce Modality-Adaptive Reconstruction (MARs), a dynamic mechanism that enables the network to selectively reconstruct either HRMS or PAN images depending on the modality selection (MARs mode). By jointly learning to reconstruct both HRMS and PAN images within a shared network, PAN-Crafter effectively incorporates sharp spatial structures while maintaining spectral fidelity in the HRMS prediction. In MS mode, the PS network $\mathcal{P}_\theta$ learns to align structural details from $\mathbf{I}_{\text{pan}}$ to $\mathbf{I}_{\text{ms}}^{\text{lr}}$ using Cross-Modality Alignment-Aware Attention (CM3A), predicting the final HRMS output $\hat{\mathbf{I}}_{\text{ms}}^{\text{hr}}$ as:

$$\hat{\mathbf{I}}_{\text{ms}}^{\text{hr}} = \mathcal{P}_\theta\left(\mathbf{I}_{\text{pan}}, \mathbf{I}_{\text{ms}}^{\text{lr}}; \text{mode} = \mathsf{MS}\right) + \mathbf{I}_{\text{ms}}^{\text{lr}}. \tag{1}$$

Conversely, in PAN mode, $\mathcal{P}_\theta$ predicts a multi-channel version of PAN, defined as:

$$\mathbf{I}_{\text{pan}}^{\text{rep}} = [\mathbf{I}_{\text{pan}} \mid \cdots \mid \mathbf{I}_{\text{pan}}] \in \mathbb{R}^{4H\times 4W\times C_{\text{ms}}}, \tag{2}$$

where $[\,\cdot \mid \cdot\,]$ denotes channel-wise concatenation, ensuring consistency across spectral bands. Since $\mathbf{I}_{\text{pan}}$ is a single-channel image, but $\mathbf{I}_{\text{ms}}^{\text{lr}}$ consists of C_{ms} spectral bands, we formulate the PAN back-reconstruction as:

$$\hat{\mathbf{I}}_{\text{pan}}^{\text{rep}} = \mathcal{P}_\theta\left(\mathbf{I}_{\text{pan}}, \mathbf{I}_{\text{ms}}^{\text{lr}}; \text{mode} = \mathsf{PAN}\right) + \mathbf{I}_{\text{pan}}^{\text{lr,rep}}, \tag{3}$$

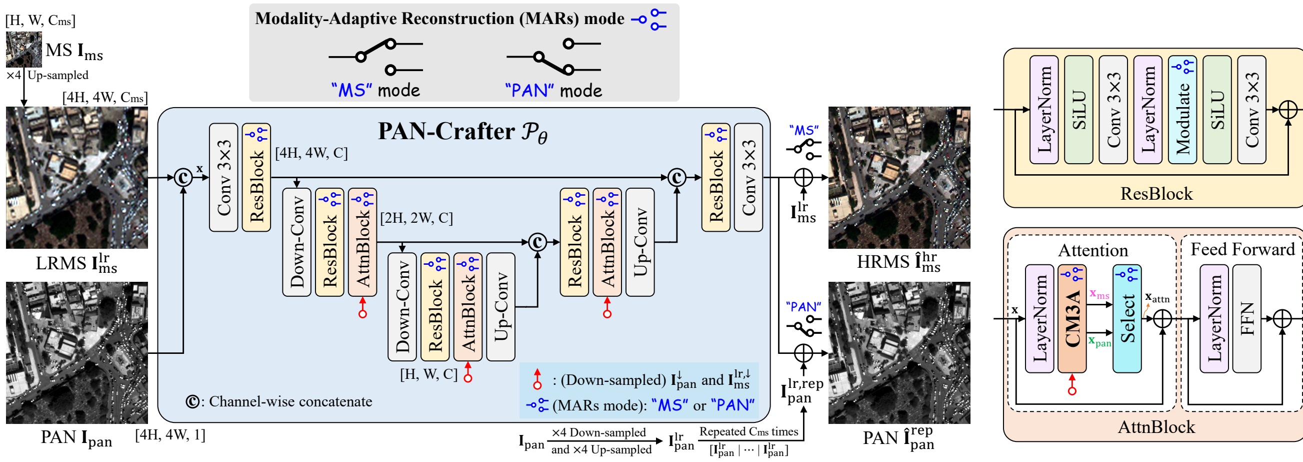

Figure 3. Overview of the proposed PAN-Crafter architecture. The network processes input PAN and LRMS images leveraging Modality-Adaptive Reconstruction (MARs) mode, which enables adaptive generation of HRMS and PAN outputs. By leveraging spatial structures and spectral fidelity of PAN and MS images, MARs ensures high-frequency details are preserved while minimizing spectral distortion. The architecture follows an encoder-decoder design, incorporating residual blocks and cross-modality alignment-aware attention (CM3A) at multiple scales to mitigate modality misalignment while preserving spectral and structural fidelity of MS and PAN images, respectively.

where $\mathbf{I}_{\text{pan}}^{\text{lr}}$ is obtained by first down-sampling the original PAN image $\mathbf{I}_{\text{pan}}$ by a factor of 4, followed by up-sampling it back to the original resolution and $\mathbf{I}_{\text{pan}}^{\text{lr, rep}}$ represents the LR PAN image $\mathbf{I}_{\text{pan}}^{\text{lr}}$ replicated C_{ms} times channel-wise, as illustrated in Fig. 3 (bottom). This formulation aligns with the residual learning strategy in MS mode, ensuring that the network in PAN mode learns high-frequency refinements. At inference time, we fix the MARs mode to MS mode, ensuring that the network always produces $\hat{\mathbf{I}}_{\text{ms}}^{\text{hr}}$ as the final HRMS output.

MARs loss. To ensure stable training, we duplicate each $(\mathbf{I}_{\text{pan}}, \mathbf{I}_{\text{ms}}, \mathbf{I}_{\text{ms}}^{\text{hr}})$ triplet along the batch dimension, processing one instance in MS mode and the other in PAN mode. This enables $\mathcal{P}_\theta$ to learn modality-specific features while maintaining consistency across both reconstruction tasks. For each mode, we apply the ℓ_1 loss independently to the predicted HRMS image $\hat{\mathbf{I}}_{\text{ms}}^{\text{hr}}$ and the back-reconstructed multi-channel PAN image $\hat{\mathbf{I}}_{\text{pan}}^{\text{rep}}$, ensuring spatial and spectral fidelity. Our MARs loss function is defined as:

$$\mathcal{L}_{\text{MARs}} = \left\|\hat{\mathbf{I}}_{\text{ms}}^{\text{hr}} - \mathbf{I}_{\text{ms}}^{\text{hr}}\right\|_1 + \lambda \left\|\hat{\mathbf{I}}_{\text{pan}}^{\text{rep}} - \mathbf{I}_{\text{pan}}^{\text{rep}}\right\|_1, \tag{4}$$

where λ is a weighting factor that balances the contribution of the PAN back-reconstruction loss. Since PAN images contain rich high-frequency details, tuning λ ensures that $\mathcal{P}_\theta$ effectively incorporates sharp spatial structures while maintaining spectral fidelity in the MS reconstruction.

3.2. PAN-Crafter

The proposed PAN-Crafter $\mathcal{P}_\theta$ is a U-Net-based network designed for robust cross-modality feature alignment and high-quality HRMS reconstruction. As shown in Fig. 3, the network follows an encoder-decoder architecture where each stage consists of a combination of residual blocks (ResBlocks) and cross-modality alignment blocks (AttnBlocks). To effectively handle varying levels of misalignment, we integrate Cross-Modality Alignment-Aware Attention (CM3A) at multiple scales throughout the network. Specifically, low- and mid-resolution stages incorporate both ResBlock and AttnBlock in a cascaded manner, while high-resolution stages use only ResBlock to reduce computational overhead.

Network Architecture. $\mathcal{P}_\theta$ takes as input the channel-wise concatenation of $\mathbf{I}_{\text{pan}}$ and $\mathbf{I}_{\text{ms}}^{\text{lr}}$. A convolutional layer (Conv) first embeds the input into a feature representation $\mathbf{x}$ of channel dimension C. The feature $\mathbf{x}$ then passes through multiple encoder stages that consist of down-sampling convolutional layers (Down-Conv), ResBlocks, and AttnBlocks. Then, $\mathbf{x}$ is progressively decoded using up-sampling layers (Up-Conv), ResBlocks, and AttnBlocks while preserving structural details and spectral integrity.

Residual Block (ResBlock). The ResBlock is designed to refine modality-specific features while preserving spatial structures. As illustrated in Fig. 3, each ResBlock consists of Layer Normalization (LN), SiLU activation, and Conv as:

$$\begin{aligned} &\mathbf{x} \leftarrow \mathsf{Conv}\left(\mathsf{SiLU}\left(\mathsf{LN}\left(\mathbf{x}\right)\right)\right), \\ &\mathbf{x} \leftarrow \mathbf{x} + \mathsf{Conv}\left(\mathsf{SiLU}\left(\mathsf{Modulate}\left(\mathsf{LN}\left(\mathbf{x}\right); \mathsf{mode}\right)\right)\right), \end{aligned} \tag{5}$$

where Modulate is a feature modulation layer. We incorporate a modulation mechanism that adjusts channel-wise feature scaling based on the input modality as:

$$\begin{aligned} \mathsf{Modulate}\left(\mathbf{x}; \mathsf{MS}\right) &: \mathbf{x} \leftarrow (1 + \gamma_{\text{ms}}) \odot \mathbf{x} + \boldsymbol{\beta}_{\text{ms}}, \\ \mathsf{Modulate}\left(\mathbf{x}; \mathsf{PAN}\right) &: \mathbf{x} \leftarrow (1 + \gamma_{\text{pan}}) \odot \mathbf{x} + \boldsymbol{\beta}_{\text{pan}}, \end{aligned} \tag{6}$$

where $\gamma_{\text{ms}}, \boldsymbol{\beta}_{\text{ms}}, \gamma_{\text{pan}}, \boldsymbol{\beta}_{\text{pan}} \in \mathbb{R}^C$ are learnable parameters, and $\odot$ denotes channel-wise multiplication. This ensures modality-aware feature adaptation while maintaining structural consistency.

Cross-Modality Attention Block (AttnBlock). The AttnBlock is designed to facilitate modality-aware feature interaction while preserving structural consistency. As illustrated in Fig. 3, the block consists of two key components: an attention layer and a feed-forward layer. The attention layer employs the CM3A to dynamically align features between the PAN and MS modalities as:

$$\mathbf{x}_{\text{ms}},\ \mathbf{x}_{\text{pan}} = \mathsf{CM3A}\left(\mathsf{LN}\left(\mathbf{x}\right);\text{mode}\right). \tag{7}$$

Subsequently, the selection layer (Select) integrates complementary information as:

$$\begin{aligned}\mathbf{x}_{\text{attn}} &= \boldsymbol{\alpha}^1_{\text{ms}} \odot \mathbf{x}_{\text{ms}} + \boldsymbol{\alpha}^2_{\text{ms}} \odot \mathbf{x}_{\text{pan}}\ (\mathsf{MS\ mode}),\\ \mathbf{x}_{\text{attn}} &= \boldsymbol{\alpha}^1_{\text{pan}} \odot \mathbf{x}_{\text{ms}} + \boldsymbol{\alpha}^2_{\text{pan}} \odot \mathbf{x}_{\text{pan}}\ (\mathsf{PAN\ mode}),\end{aligned} \tag{8}$$

where $\boldsymbol{\alpha}^1_{\text{ms}}, \boldsymbol{\alpha}^2_{\text{ms}}, \boldsymbol{\alpha}^1_{\text{pan}}, \boldsymbol{\alpha}^2_{\text{pan}} \in \mathbb{R}^C$ are learnable parameters. The final attended feature is integrated via a residual connection as $\mathbf{x} \leftarrow \mathbf{x} + \mathbf{x}_{\text{attn}}$. Following the attention operation, the feed-forward network (FFN) refines the attended features as $\mathbf{x} \leftarrow \mathbf{x} + \mathsf{FFN}\left(\mathsf{LN}\left(\mathbf{x}\right)\right)$.

3.3. Cross-Modality Alignment-Aware Attention

Local attention mechanism. To effectively handle cross-modality misalignment, it is not necessary to estimate the global displacement across the entire image. Since $(\mathbf{I}_{\text{pan}}, \mathbf{I}_{\text{ms}}, \mathbf{I}^{\text{hr}}_{\text{ms}})$ triplets are generally pre-aligned to a certain degree, our CM3A adopts Pan *et al.* [35] and operates within a local attention window rather than global attention. As shown in Fig. 4, for a given query position (i, j), we compute attention scores only within a $k \times k$ receptive field centered around the query, ensuring computational efficiency while capturing local misalignment, where $k = 2k' + 1$ is the receptive field size. Given a query feature $\mathbf{Q} \in \mathbb{R}^{H\times W\times C}$, and key-value pairs $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{H\times W\times C}$, Local Attention function (LocalAttn) [35] computes attention scores within the $k \times k$ local receptive field.

Misalignment-guided feature interaction. Our CM3A dynamically aligns features based on the selected MARs mode by integrating both self-attention and alignment-aware attention mechanisms. As illustrated in Fig. 4, the attention process differs depending on whether $\mathcal{P}_\theta$ operates in MS mode or PAN mode. In MS mode, $\mathcal{P}_\theta$ aims to predict $\mathbf{I}^{\text{hr}}_{\text{ms}}$, ensuring spectral fidelity while incorporating structural details from PAN images. To achieve this, self-attention is first applied to maintain consistency within the MS feature space. Specifically, the query feature $\mathbf{Q}$ is constructed by concatenating the input feature $\mathbf{x}$ with a down-sampled version of the LRMS image $\mathbf{I}^{\text{lr},\downarrow}_{\text{ms}}$, ensuring that both have the same spatial resolution:

$$\mathbf{Q} = \mathsf{Conv}\left(\left[\mathbf{I}^{\text{lr},\downarrow}_{\text{ms}} \mid \mathbf{x}\right]\right). \tag{9}$$

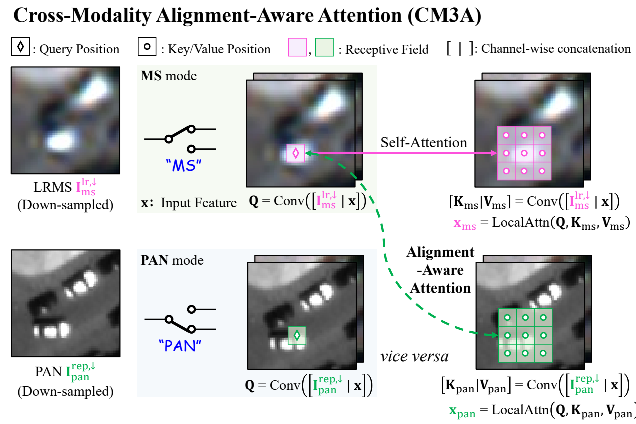

Figure 4. Cross-Modality Alignment-Aware Attention (CM3A) enables bidirectional alignment by transferring MS texture to PAN structure during HRMS reconstruction and PAN structure to MS texture during PAN back-reconstruction. This mechanism not only mitigates cross-modality misalignment but also ensures structural and spectral fidelity in the reconstructed images.

The query attends to key-value pairs derived from the same modality, enabling self-attention to refine MS-specific features. $\mathbf{K}_{\text{ms}}$ and $\mathbf{V}_{\text{ms}}$ are constructed as:

$$\begin{aligned}\left[\mathbf{K}_{\text{ms}} \mid \mathbf{V}_{\text{ms}}\right] &= \mathsf{Conv}\left(\left[\mathbf{I}^{\text{lr},\downarrow}_{\text{ms}} \mid \mathbf{x}\right]\right),\\ \mathbf{x}_{\text{ms}} &= \mathsf{LocalAttn}(\mathbf{Q}, \mathbf{K}_{\text{ms}}, \mathbf{V}_{\text{ms}}).\end{aligned} \tag{10}$$

To further enhance the MS feature representation with PAN's structural information, alignment-aware attention allows $\mathbf{Q}$ to attend to $\mathbf{K}$-$\mathbf{V}$ pairs derived from a down-sampled PAN image $\mathbf{I}^{\text{rep},\downarrow}_{\text{pan}}$:

$$\begin{aligned}\left[\mathbf{K}_{\text{pan}} \mid \mathbf{V}_{\text{pan}}\right] &= \mathsf{Conv}\left(\left[\mathbf{I}^{\text{rep},\downarrow}_{\text{pan}} \mid \mathbf{x}\right]\right),\\ \mathbf{x}_{\text{pan}} &= \mathsf{LocalAttn}(\mathbf{Q}, \mathbf{K}_{\text{pan}}, \mathbf{V}_{\text{pan}}),\end{aligned} \tag{11}$$

This process enables $\mathcal{P}_\theta$ to extract high-frequency details from PAN images while mitigating cross-modality misalignment. In PAN mode, $\mathcal{P}_\theta$ back-reconstructs $\mathbf{I}^{\text{rep}}_{\text{pan}}$, prioritizing the preservation of sharp spatial details while leveraging spatial information from MS images. Here, $\mathbf{Q}$ is constructed differently to reflect the modality shift. Instead of using $\mathbf{I}^{\text{lr},\downarrow}_{\text{ms}}$, the query is formed by concatenating the input feature $\mathbf{x}$ with $\mathbf{I}^{\text{rep},\downarrow}_{\text{pan}}$ as:

$$\mathbf{Q} = \mathsf{Conv}\left(\left[\mathbf{I}^{\text{rep},\downarrow}_{\text{pan}} \mid \mathbf{x}\right]\right). \tag{12}$$

The subsequent self-attention and alignment-aware attention operations mirror those in MS mode but in reverse, ensuring structural consistency in PAN back-reconstruction. Since PAN back-reconstruction serves as an auxiliary task, it reinforces spatial sharpness in HRMS prediction. By jointly leveraging self-attention for modality-consistent refinement and alignment-aware attention for cross-modality adaptation, CM3A effectively mitigates misalignment while

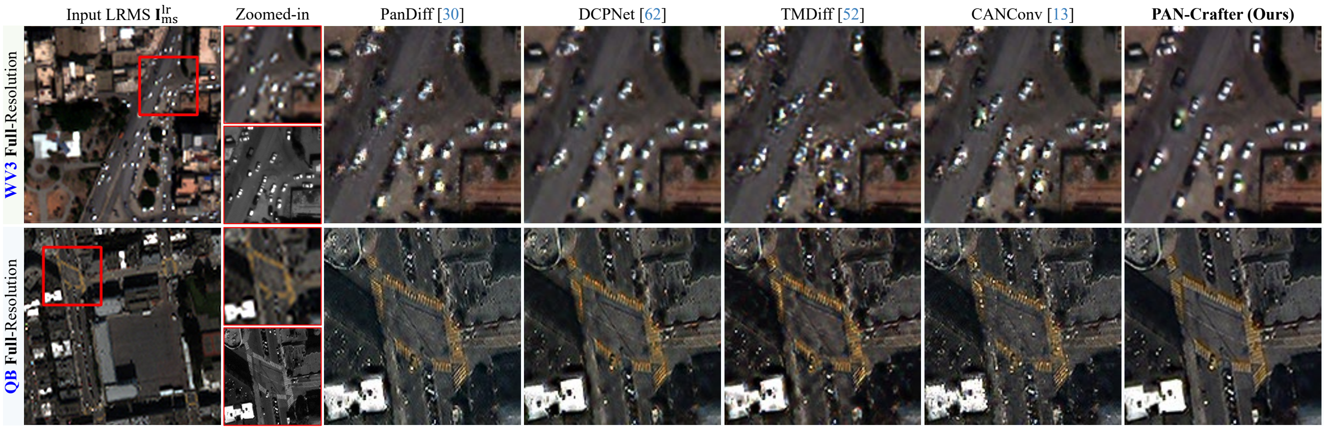

Figure 5. Visual comparison of PAN-Sharpening (PS) results on the WV3 and QB datasets at full-resolution. The leftmost column shows the input LRMS images, with red boxes indicating zoomed-in regions for both LRMS and PAN images. Only our proposed PAN-Crafter is capable of generating high-quality images with clear edges around cars, buildings, and crosswalk lines, whereas previous methods tend to produce blurry and distorted results from misaligned input PAN and MS pair images.

WV3 Dataset		Full-Resolution			Reduced-Resolution						Inference	Memory↓
Methods	Publications	HQNR↑	D_s↓	D_λ↓	ERGAS↓	SCC↑	SAM↓	Q8↑	PSNR↑	SSIM↑	Time↓ (s)	(GB)
PanNet [57]	ICCV 2017	0.918	0.049	0.035	2.538	0.979	3.402	0.913	36.148	0.966	-	-
MSDCNN [58]	JSTARS 2018	0.924	0.050	0.028	2.489	0.979	3.300	0.914	36.329	0.967	-	-
FusionNet [50]	ICCV 2021	0.920	0.053	0.029	2.428	0.981	3.188	0.916	36.569	0.968	-	-
LAGConv [20]	AAAI 2022	0.915	0.055	0.033	2.380	0.981	3.153	0.916	36.732	0.970	**0.004**	3.281
S2DBPN [61]	TGRS 2023	0.946	0.030	0.025	2.245	0.985	3.019	0.917	37.216	0.972	0.005	2.387
PanDiff [30]	TGRS 2023	0.952	0.034	**0.014**	2.276	0.984	3.058	0.913	37.029	0.971	2.955	2.328
DCPNet [62]	TGRS 2024	0.923	0.036	0.043	2.301	0.984	3.083	0.915	37.009	0.972	0.109	7.213
TMDiff [52]	TGRS 2024	0.924	0.059	0.018	2.151	0.986	2.885	0.915	37.477	0.973	9.997	9.910
CANConv [13]	CVPR 2024	0.951	0.030	0.020	2.163	0.985	2.927	0.918	37.441	0.973	0.451	2.713
PAN-Crafter	-	**0.958**	**0.027**	0.016	**2.040**	**0.988**	**2.787**	**0.922**	**37.956**	**0.976**	0.009	**1.711**

Table 1. Quantitative comparison of deep learning-based PS methods on the WV3 dataset. **Red** indicate the best performance in each metric. The inference time and memory usage are measured on a $256 \times 256 \times 8$ HRMS target at reduced-resolution.

preserving spectral fidelity and structural coherence, ultimately enhancing HRMS quality.

4. Experiments

4.1. Datasets

We evaluate PAN-Crafter on four widely used PAN-sharpening datasets from PanCollection [9]: WorldView-3 (WV3), QuickBird (QB), GaoFen-2 (GF2), and WorldView-2 (WV2). For training, we use $64 \times 64 \times 1$ patches for PAN and $16 \times 16 \times C_{\text{ms}}$ patches for MS, where $C_{\text{ms}} = 4$ for QB and GF2, and $C_{\text{ms}} = 8$ for WV3. Each satellite dataset has its own test set, which consists of reduced-resolution and full-resolution images. To further assess generalization, we evaluate PAN-Crafter on WV2, an unseen satellite dataset used exclusively for testing. WV2 serves as a zero-shot benchmark, measuring the model's robustness to sensor variations without fine-tuning. The reduced-resolution test images have a PAN spatial size of 256×256, while the full-resolution test images have a higher PAN spatial size of 512×512.

4.2. Experiment Details

We implement PAN-Crafter in PyTorch [36] and conduct all experiments on a single NVIDIA GeForce RTX 3090 GPU. Each model is trained for 50,000 iterations with a 100-step warmup period. We use AdamW optimizer [28] with an initial learning rate of 1×10^{-4}, a weight decay of 0.01, and a cosine-annealing scheduler [27] to progressively reduce the learning rate. The batch size is set to 48, but for MARs loss computation, the batch is duplicated across MS mode and PAN mode, resulting in an effective batch size of 96. We empirically set the loss weight to $\lambda = 1.0$ and the local attention kernel size to $k = 3$. All feature dimensions are fixed to $C = 128$. Standard data augmentation techniques, including random horizontal/vertical flips, 90-degree rotations, and random cropping, are applied to improve generalization. To ensure reproducibility, we fix the random seed to 2,025 across all experiments. We evaluate PAN-Crafter in terms of ERGAS [46], SCC [15], SAM [60], Q4/Q8 [44], PSNR [48], and SSIM [48] metrics for reduced-resolution datasets, while HQNR [45], D_S, and D_λ metrics are used for full-resolution datasets.

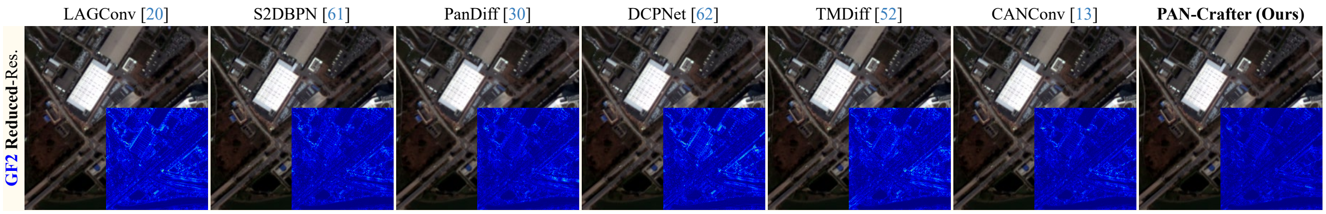

Figure 6. Visual comparison of PS results on the GF2 and QB datasets at reduced-resolution. The blue-colored insets represent error maps computed against the ground truth (GT), where brighter regions indicate higher reconstruction errors.

Methods	**GF2** Dataset						**QB** Dataset					
	Full-Resolution		Reduced-Resolution				Full-Resolution		Reduced-Resolution			
	HQNR↑	D_s↓	ERGAS↓	SCC↑	SAM↓	PSNR↑	HQNR↑	D_s↓	ERGAS↓	SCC↑	SAM↓	PSNR↑
PanNet [57]	0.929	0.052	1.038	0.975	1.050	39.197	0.851	0.092	4.856	0.966	5.273	35.563
MSDCNN [58]	0.898	0.079	0.862	0.983	0.946	40.730	0.888	0.058	4.074	0.977	4.828	37.040
FusionNet [50]	0.865	0.105	0.960	0.980	0.971	39.866	0.853	0.079	4.183	0.975	4.892	36.821
LAGConv [20]	0.895	0.078	0.816	0.985	0.886	41.147	0.892	**0.035**	3.845	0.980	4.682	37.565
S2DBPN [61]	0.935	0.046	0.686	0.990	0.772	42.686	0.908	0.036	3.956	0.980	4.849	37.314
PanDiff [30]	0.936	0.045	0.674	0.990	0.767	42.827	0.919	0.055	3.723	0.982	4.611	37.842
DCPNet [62]	0.953	0.024	0.724	0.988	0.806	42.312	0.880	0.073	3.618	0.983	**4.420**	38.079
TMDiff [52]	0.942	0.030	0.754	0.988	0.764	41.896	0.901	0.068	3.804	0.981	4.627	37.642
CANConv [13]	0.919	0.063	0.653	0.991	0.722	43.166	0.893	0.070	3.740	0.982	4.554	37.795
PAN-Crafter	**0.964**	**0.017**	**0.522**	**0.994**	**0.596**	**45.076**	**0.920**	0.039	**3.570**	**0.984**	4.426	**38.195**

Table 2. Quantitative comparison of deep learning-based PS methods on the GF2 and QB datasets. **Red** indicates the best performance.

4.3. Experimental Results

We computed all evaluation metrics using the official Pan-Collection [9] repository to ensure standardized measurement. To ensure a fair and comprehensive evaluation, we utilized the official implementations of the compared methods whenever available.

Qualitative comparison. We qualitatively compare our PAN-Crafter against very recent PS methods on both full- (Fig. 1, Fig. 5) and reduced-resolution (Fig. 6) datasets. Previous methods often produce blurry artifacts, double edges, or spectral distortions due to their inability to handle cross-modality misalignment. In contrast, PAN-Crafter effectively preserves fine details, ensuring sharper edges and clearer textures by leveraging CM3A attention for explicit cross-modality alignment. Beyond direct feature alignment, MARs further enhances robustness by leveraging PAN mode as an auxiliary self-supervision mechanism, encouraging the network to distill sharper structural details into the final HRMS output. The synergy of MARs and CM3A allows PAN-Crafter to generate high-quality HRMS images with superior spatial and spectral integrity, demonstrating its effectiveness in mitigating misalignment.

Quantitative evaluation. We compare PAN-Crafter against deep learning-based PS methods on WV3, GF2, and QB datasets. As shown in Table 1 and Table 2, our PAN-Crafter consistently outperforms existing approaches across most evaluation metrics while maintaining low memory consumption and fast inference time. Notably, our method surpasses diffusion-based models (PanDiff [30] and TMDiff [52]) in both full- and reduced-resolution datasets, demonstrating that our CM3A module effectively handles local misalignment without the computational burden of iterative diffusion process. Specifically, PAN-Crafter achieves $328.33\times$ and $1110.78\times$ faster inference time compared to PanDiff and TMDiff, respectively. Additionally, compared to CANConv [13], which relies on k-means clustering [10] for kernel generation, PAN-Crafter is $50.11\times$ faster, highlighting the efficiency of our locally adaptive alignment strategy. Notably, on the WV3 dataset, the D_λ score is slightly lower than other methods. This is because our method aligns the generated HRMS image to the LRMS image rather than the PAN image using CM3A in MS mode, ensuring spectral fidelity at the cost of a lower PAN-HRMS correlation. For the QB dataset, which is known to be the most challenging due to its higher spectral distortion and complex scene variations, PAN-Crafter exhibits slightly lower D_s and SAM scores compared to some methods. However, it still achieves the best overall performance across both full- and reduced-resolution evaluations, confirming its robustness to diverse satellite imagery.

Generalization on unseen satellite dataset. To assess the generalization capability of PAN-Crafter, we conduct fully zero-shot evaluations on the WV2 dataset, an unseen satellite dataset not included in the training phase. As shown in Table 3 and Fig. 7, all models are trained on WV3 and directly tested on WV2 without any fine-tuning. While existing methods suffer from performance degradation due to domain shifts and sensor variations, PAN-Crafter demonstrates superior generalization ability, achieving the best performance across all key metrics. The strong general-

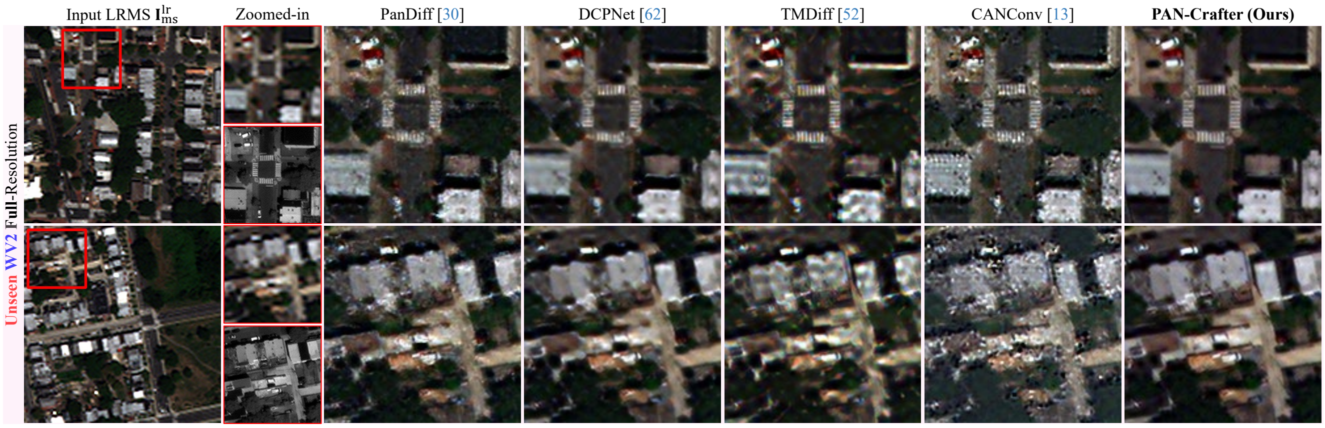

Figure 7. Visual comparison of PS results on the unseen WV2 dataset at full-resolution. The leftmost column shows the input LRMS image, with red boxes indicating zoomed-in regions for both LRMS and PAN images. Since WV2 is not included in the training phase, this evaluation represents a real-world zero-shot setting, assessing the generalization capability of PS models. Our PAN-Crafter produces sharper details with fewer distortions compared to existing methods, demonstrating superior cross-satellite robustness.

Methods	WV2 Dataset (Unseen satellite dataset)				
	HQNR↑	ERGAS↓	SCC↑	SAM↓	PSNR↑
PanNet [57]	0.875	5.481	0.876	7.040	27.120
MSDCNN [58]	0.862	4.930	0.905	5.898	27.901
FusionNet [50]	0.862	5.100	0.902	6.118	27.616
LAGConv [20]	0.902	5.133	0.885	6.094	27.525
S2DBPN [61]	0.813	5.703	0.915	7.063	26.748
PanDiff [30]	0.932	4.291	0.916	5.430	28.964
DCPNet [62]	0.797	5.507	**0.931**	10.174	27.063
TMDiff [52]	0.874	5.157	0.875	6.087	27.473
CANConv [13]	0.876	4.328	0.918	5.481	29.005
PAN-Crafter	**0.942**	**4.169**	0.924	**5.078**	**29.276**

Table 3. Quantitative comparison of deep learning-based PS methods on the unseen WV2 satellite dataset. All models are trained on WV3 and evaluated on WV2 to assess real-world generalization.

ization stems from our modality-consistent alignment strategy, which explicitly mitigates cross-modality misalignment without relying on dataset-specific priors. These results highlight the robustness of our framework in real-world PS scenarios, where test-time adaptation is often infeasible.

4.4. Ablation Studies

Impact of MARs mode. As shown in Table 4, incorporating MARs improves all evalution metric, particularly enhancing SAM and PSNR scores. This highlights the effectiveness of PAN as an auxiliary self-supervision signal, allowing the network to learn sharper spatial details while maintaining spectral fidelity. Despite a slight increase in memory usage, MARs provides a substantial performance gain, justifying its inclusion in our framework.

Analysis of CM3A. In Table 4, we further examine the impact of our Cross-Modality Alignment-Aware Attention (CM3A). Without CM3A, the model struggles to mitigate cross-modality misalignment, leading to degraded spectral and spatial consistency. The results show that CM3A significantly improves HQNR and ERGAS scores while reducing SAM, demonstrating its effectiveness in aligning MS textures with the corresponding PAN structures.

CM3A	MARs	WV3 Dataset					
		HQNR↑	ERGAS↓	SAM↓	PSNR↑	Time↓	Memory↓
		0.948	2.232	2.980	37.245	0.006	1.537
✓		0.949	2.212	2.970	37.285	0.007	1.556
	✓	0.956	2.122	2.873	37.602	0.009	1.701
✓	✓	**0.958**	**2.040**	**2.787**	**37.956**	0.009	1.711

Table 4. Ablation studies on CM3A and MARs on the WV3 dataset. Notably, the combination of both components achieves the best performance, highlighting their synergistic effect in jointly refining spatial and spectral consistency.

5. Conclusion

We introduce PAN-Crafter, a modality-consistent alignment framework for PAN-sharpening that explicitly addresses cross-modality misalignment. MARs enables joint HRMS and PAN reconstruction, leveraging PAN's high-frequency details as auxiliary self-supervision, while CM3A ensures bidirectional alignment between MS and the corresponding PAN images. PAN-Crafter achieves SOTA performance across multiple benchmarks, preserving fine details and spectral integrity. It also generalizes well to unseen satellite data, demonstrating strong zero-shot robustness. With superior efficiency in inference speed and memory usage, PAN-Crafter offers a practical and scalable solution for real-world remote sensing applications.

Acknowledgement. This work was supported by National Research Foundation of Korea (NRF) grant funded by the Korean Government [Ministry of Science and ICT (Information and Communications Technology)] (Project Number: RS2024-00338513, Project Title: AI-based Computer Vision Study for Satellite Image Processing and Analysis, 100%).

References

[1] 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. 2

[2] Bruno Aiazzi, Luciano Alparone, Stefano Baronti, Andrea Garzelli, and Massimo Selva. Mtf-tailored multiscale fusion of high-resolution ms and pan imagery. *Photogrammetric Engineering & Remote Sensing*, 72(5):591–596, 2006. 2

[3] Coloma Ballester, Vicent Caselles, Laura Igual, Joan Verdera, and Bernard Rougé. A variational model for p+ xs image fusion. *International Journal of Computer Vision*, 69:43–58, 2006. 2

[4] Wele Gedara Chaminda Bandara and Vishal M Patel. Hypertransformer: A textural and spectral feature fusion transformer for pansharpening. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1767–1777, 2022. 2

[5] Xiangyong Cao, Yang Chen, and Wenfei Cao. Proximal pannet: A model-based deep network for pansharpening. In *Proceedings of the AAAI conference on artificial intelligence*, pages 176–184, 2022. 2

[6] Wjoseph Carper, Thomasm Lillesand, Ralphw Kiefer, et al. The use of intensity-hue-saturation transformations for merging spot panchromatic and multispectral image data. *Photogrammetric Engineering and remote sensing*, 56(4): 459–467, 1990. 2

[7] 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. 2

[8] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 3

[9] 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. 2, 6, 7

[10] Yufei Ding, Yue Zhao, Xipeng Shen, Madanlal Musuvathi, and Todd Mytkowicz. Yinyang k-means: A drop-in replacement of the classic k-means with consistent speedup. In *International conference on machine learning*, pages 579–587. PMLR, 2015. 7

[11] Jeonghyeok Do, Jaehyup Lee, and Munchurl Kim. C-diffset: Leveraging latent diffusion for sar-to-eo image translation with confidence-guided reliable object generation. *arXiv preprint arXiv:2411.10788*, 2024. 1

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[13] Yule Duan, Xiao Wu, Haoyu Deng, and Liang-Jian Deng. Content-adaptive non-local convolution for remote sensing pansharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27738–27747, 2024. 2, 3, 6, 7, 8

[14] 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*, pages 10265–10274, 2019. 2

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

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[17] Junming Hou, Qi Cao, Ran Ran, Che Liu, Junling Li, and Liang-jian Deng. Bidomain modeling paradigm for pansharpening. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 347–357, 2023. 2

[18] Junming Hou, Zihan Cao, Naishan Zheng, Xuan Li, Xiaoyu Chen, Xinyang Liu, Xiaofeng Cong, Danfeng Hong, and Man Zhou. Linearly-evolved transformer for pansharpening. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1486–1494, 2024. 2

[19] Yuming Jiang, Kelvin CK Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Robust reference-based super-resolution via c2-matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2103–2112, 2021. 3

[20] 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*, pages 1113–1121, 2022. 2, 3, 6, 7, 8

[21] 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. *arXiv preprint arXiv:2412.06243*, 2024. 3

[22] Jaehyup Lee, Soomin Seo, and Munchurl Kim. Sipsa-net: Shift-invariant pan sharpening with moving object alignment for satellite imagery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10166–10174, 2021. 2, 3

[23] 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. 1

[24] Lin Liu, Shanxin Yuan, Jianzhuang Liu, Xin Guo, Youliang Yan, and Qi Tian. Siamtrans: zero-shot multi-frame image restoration with pre-trained siamese transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1747–1755, 2022. 3

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3

[26] Laetitia Loncan, Luis B De Almeida, José M Bioucas-Dias, Xavier Briottet, Jocelyn Chanussot, Nicolas Dobigeon, Sophie Fabre, Wenzhi Liao, Giorgio A Licciardi, Miguel Simoes, et al. Hyperspectral pansharpening: A review. *IEEE Geoscience and remote sensing magazine*, 3(3):27–46, 2015. 2

[27] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[29] Giuseppe Masi, Davide Cozzolino, Luisa Verdoliva, and Giuseppe Scarpa. Pansharpening by convolutional neural networks. *Remote Sensing*, 8(7):594, 2016. 2

[30] 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. 2, 3, 6, 7, 8

[31] Xiangchao Meng, Yiming Xiong, Feng Shao, Huanfeng Shen, Weiwei Sun, Gang Yang, Qiangqiang Yuan, Randi Fu, and Hongyan Zhang. A large-scale benchmark data set for evaluating pansharpening performance: Overview and implementation. *IEEE Geoscience and Remote Sensing Magazine*, 9(1):18–52, 2020. 2

[32] Robbe Neyns and Frank Canters. Mapping of urban vegetation with high-resolution remote sensing: A review. *Remote sensing*, 14(4):1031, 2022. 1

[33] Keiron O'shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015. 2

[34] 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. 2

[35] Xuran Pan, Tianzhu Ye, Zhuofan Xia, Shiji Song, and Gao Huang. Slide-transformer: Hierarchical vision transformer with local self-attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2082–2091, 2023. 3, 5

[36] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[37] Siran Peng, Xiangyu Zhu, Haoyu Deng, Liang-Jian Deng, and Zhen Lei. Fusionmamba: Efficient remote sensing image fusion with state space model. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 2

[38] Sheida Rahmani, Melissa Strait, Daria Merkurjev, Michael Moeller, and Todd Wittman. An adaptive ihs pan-sharpening method. *IEEE Geoscience and Remote Sensing Letters*, 7(4): 746–750, 2010. 2

[39] Vijay P Shah, Nicolas H Younan, and Roger L King. An efficient pan-sharpening method via a combined adaptive pca approach and contourlets. *IEEE transactions on geoscience and remote sensing*, 46(5):1323–1335, 2008. 2

[40] Mark J Shensa. The discrete wavelet transform: wedding the a trous and mallat algorithms. *IEEE Transactions on signal processing*, 40(10):2464–2482, 1992. 2

[41] Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik Learned-Miller, and Jan Kautz. Pixel-adaptive convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11166–11175, 2019. 3

[42] 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. 2

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[44] 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. 6

[45] 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. 2, 6

[46] 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"*, pages 99–103. SEE/URISCA, 2000. 6

[47] Yudong Wang, Liang-Jian Deng, Tian-Jing Zhang, and Xiao Wu. Ssconv: Explicit spectral-to-spatial convolution for pansharpening. In *Proceedings of the 29th ACM international conference on multimedia*, pages 4472–4480, 2021. 2

[48] 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. 6

[49] 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. 1

[50] Xiao Wu, Ting-Zhu Huang, Liang-Jian Deng, and Tian-Jing Zhang. Dynamic cross feature fusion for remote sensing pansharpening. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14687–14696, 2021. 2, 6, 7, 8

[51] Zhong-Cheng Wu, Ting-Zhu Huang, Liang-Jian Deng, Jin-Fan Hu, and Gemine Vivone. Vo+ net: An adaptive approach using variational optimization and deep learning for panchromatic sharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021. 2

[52] 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*, 2024. 3, 6, 7, 8

[53] Yonghao Xu, Tao Bai, Weikang Yu, Shizhen Chang, Peter M Atkinson, and Pedram Ghamisi. Ai security for geoscience and remote sensing: Challenges and future trends. *IEEE Geoscience and Remote Sensing Magazine*, 11(2):60–85, 2023. 1

[54] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5791–5800, 2020. 3

[55] 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*, pages 1788–1797, 2022. 2

[56] Jun Yang, Peng Gong, Rong Fu, Minghua Zhang, Jingming Chen, Shunlin Liang, Bing Xu, Jiancheng Shi, and Robert Dickinson. The role of satellite remote sensing in climate change studies. *Nature climate change*, 3(10):875–883, 2013. 1

[57] Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, and John Paisley. Pannet: A deep network architecture for pan-sharpening. In *Proceedings of the IEEE international conference on computer vision*, pages 5449–5457, 2017. 2, 6, 7, 8

[58] 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. 2, 6, 7, 8

[59] Qiangqiang Yuan, Huanfeng Shen, Tongwen Li, Zhiwei Li, Shuwen Li, Yun Jiang, Hongzhang Xu, Weiwei Tan, Qianqian Yang, Jiwen Wang, et al. Deep learning in environmental remote sensing: Achievements and challenges. *Remote sensing of Environment*, 241:111716, 2020. 1

[60] 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. 6

[61] 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. 2, 6, 7, 8

[62] 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. 2, 6, 7, 8

[63] Zhilu Zhang, Haolin Wang, Ming Liu, Ruohao Wang, Jiawei Zhang, and Wangmeng Zuo. Learning raw-to-srgb mappings with inaccurately aligned supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4348–4358, 2021. 3

[64] Yu Zhong, Xiao Wu, Liang-Jian Deng, Zihan Cao, and Hong-Xia Dou. Ssdiff: Spatial-spectral integrated diffusion model for remote sensing pansharpening. *Advances in Neural Information Processing Systems*, 37:77962–77986, 2024. 3

[65] Zeyu Zhu, Xiangyong Cao, Man Zhou, Junhao Huang, and Deyu Meng. Probability-based global cross-modal upsampling for pansharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14039–14048, 2023. 2