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

Supplementary Material

A. Additional Discussions on Results

A.1. Additional Qualitative Comparisons

Fig. 8 and Fig. 9 provide additional qualitative comparisons of PS results on the WorldView-3 (WV3), QuickBird (QB), and GaoFen-2 (GF2) datasets [9] at full-resolution. Fig. 10 and Fig. 11 provide additional qualitative comparisons of PS results on the WV3, QB, GF2 datasets at reduced-resolution. Our PAN-Crafter consistently generates pan-sharpened images with minimal artifacts, preserving fine details around buildings and vehicles, whereas existing methods often produce blurring or structural distortions.

A.2. Additional Quantitative Evaluation

To provide a more comprehensive analysis, we present extended quantitative evaluations in the *Supplementary Material*. Table 6, Table 7, and Table 8 provide detailed results on the WV3, GF2, and QB datasets, respectively. The (i) metrics, (ii) full-/no-reference, and (iii) Wald's evaluation protocol [42, 44] used for evaluations are quite commonly known for PS restoration in remote sensing. We confirm that all comparison methods were trained using their official codebases and were evaluated, following the same protocol [42, 44] used in the prior work [13, 30, 52, 62]. To ensure fairness, we applied the same data splits [9], random seeds, data augmentations. PAN-Crafter consistently achieves strong performance across various evaluation metrics, further demonstrating its effectiveness in preserving both spatial and spectral fidelity. These extended results reinforce the robustness of our approach across different datasets and imaging conditions.

A.3. Generalization on Unseen Satellite Dataset

To further evaluate the zero-shot generalization capability of PAN-Crafter, we provide additional quantitative and qualitative results on the unseen WorldView-2 (WV2) dataset [9]. Table 9 and Fig. 12 present quantitative and qualitative results, respectively. Despite not being trained on WV2, PAN-Crafter outperforms existing methods in both spatial and spectral fidelity, demonstrating its robustness to cross-sensor variations. The results highlight the effectiveness of our cross-modality alignment strategy, enabling strong generalization without requiring additional fine-tuning.

A.4. Computational Complexity

Efficiency is a critical factor in PS applications, particularly for real-time and large-scale remote sensing tasks. We evaluate the computational complexity of PAN-Crafter against state-of-the-art methods in terms of inference time, memory consumption, FLOPs, and the number of parameters, as summarized in Table 10. Our PAN-Crafter achieves a significant speedup over diffusion-based models, with over 1110.78× faster inference time compared to TMDiff [52] and over 328.33× faster than PanDiff [30], demonstrating the efficiency of our attention-based alignment mechanism. Compared to CANConv [13], which utilizes k-means clustering [10] for spatial adaptation, PAN-Crafter achieves 50.11× faster inference while maintaining competitive reconstruction quality.

B. Limitations

B.1. Misalignment between multi-spectral bands

Our method addresses cross-modality misalignment but does not explicitly handle misalignment between multi-spectral bands. A potential solution is to apply depth-wise separable convolutional layers in CM3A for MS feature projection, preventing information mixing across spectral bands.

C. Further Ablation Studies

C.1. Ablation studies on MARs and CM3A

Table 11, Table 12, and Table 13 present extended ablation studies on MARs and CM3A across WV3, GF2, and QB datasets. The results demonstrate the significant impact of MARs, which consistently improves both spatial and spectral fidelity by leveraging auxiliary PAN self-supervision. While CM3A alone provides only marginal benefits, its effectiveness is significantly amplified when combined with MARs. The bidirectional interaction between PAN and MS reconstruction in MARs enables CM3A to refine cross-modality alignment more effectively, leading to a synergistic enhancement in both spatial consistency and spectral preservation. These findings further validate the importance of jointly leveraging MARs and CM3A for robust PAN-sharpening.

Ablation settings. We clarify the ablation setups for the main components: (i) without MARs – we remove the PAN mode entirely, including all learnable parameters related to modality switching (i.e., α, β, and γ). This turns the architecture into a single-task (PS) network; (ii) without CM3A – we remove the concatenated original inputs ($\mathbf{I}_{\text{ms}}^{\text{lr},\downarrow}, \mathbf{I}_{\text{pan}}^{\text{rep},\downarrow}$) from the attention block, disabling cross-modality conditioning in the alignment mechanism.

C.2. Additional ablation studies

Additional component-wise ablations on the WV3, GF2, and QB datasets were done: (i) without modulation parameters ($\boldsymbol{\beta}$, $\boldsymbol{\gamma}$) – the modulation is not applied to the feature maps (Table 14); (ii) without combination parameter ($\boldsymbol{\alpha}$) – we eliminate the learnable fusion weight between features (Table 14); (iii) varying local attention window size k in CM3A (Table 15); (iv) two-stage learning with pretrain on the PAN back-reconstruction and finetune for PS (Table 16).

C.3. Justification of ablation results

The reason that U-Net without MAR and CM3A is superior to existing methods is that our multi-scale window-based local attention [35] in U-Net is still effective to constitute a strong baseline. We additionally ablated this component and its result can be seen in the (Table 17). Without it (replacement of the local attention layer with convolution layer in the baseline model), the performance drop is significant.

D. Discussion on Various Cross-Attention Approaches

While the two prior works employ cross-attention in multi-frame restoration [24] and reference-based SR [54], their setups are substantially different from ours. Siamtrans [24] applies cross-attention after warping adjacent video frames to a query frame, assuming strong temporal correlation and accurate alignment. Similarly, TTSR [54] uses global cross-attention between an LR image and a semantically unrelated HR reference. [24, 54] rely on global attention [12, 43], which is computationally expensive and less suitable for local misalignment patterns. In contrast, our CM3A module is specially tailored for PS, where PAN and MS images are often not significantly misaligned and share similar spatial structures. To effectively handle this, we introduce a novel MARs-mode-dependent local cross-/self-attention. Also, we replace fixed positional embeddings (PE) with downsampled original images concatenated to the attention inputs to implicitly learn the relative misalignment between modalities. The distinction between [24, 54] and our CM3A is summarized as Table 5.

E. Local Attention Mechanisms

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 fuction (LocalAttn) [35] computes attention scores within the $k \times k$ local receptive field as follows:

$$\begin{aligned}
&\mathsf{Attn}_{i,j,m,n} = \mathbf{Q}_{i,j}\mathbf{K}_{i+m,j+n}, \\
&\mathsf{Attn} \leftarrow \mathsf{SoftMax}\left(\mathsf{Attn}/\sqrt{C}\right), \\
&\mathsf{LocalAttn}(\mathbf{Q},\mathbf{K},\mathbf{V})_{i,j} \\
&= \sum_{m=-k'}^{k'} \sum_{n=-k'}^{k'} \mathsf{Attn}_{i,j,m,n}\mathbf{V}_{i+m,j+n},
\end{aligned} \tag{13}$$

where $\mathsf{Attn} \in \mathbb{R}^{H\times W\times k\times k}$ is the attention score, and SoftMax is applied along the last two dimensions.

Computational complexity analysis. The computational complexity of a global self-attention layer for a feature map $\mathbf{x}$ of size (H, W, C) is:

$$\mathcal{O}(2(HW)^2C), \tag{14}$$

due to pairwise interactions across all spatial locations. In contrast, CM3A leverages local attention with a fixed receptive field size of $k \times k$, reducing the complexity to:

$$\mathcal{O}(2(HW)k^2C). \tag{15}$$

Since $k^2 \ll HW$, our approach significantly reduces computational overhead while maintaining effective cross-modality feature alignment. By restricting attention to local neighborhoods, CM3A balances efficiency with the ability to capture localized structural discrepancies between PAN and MS images.

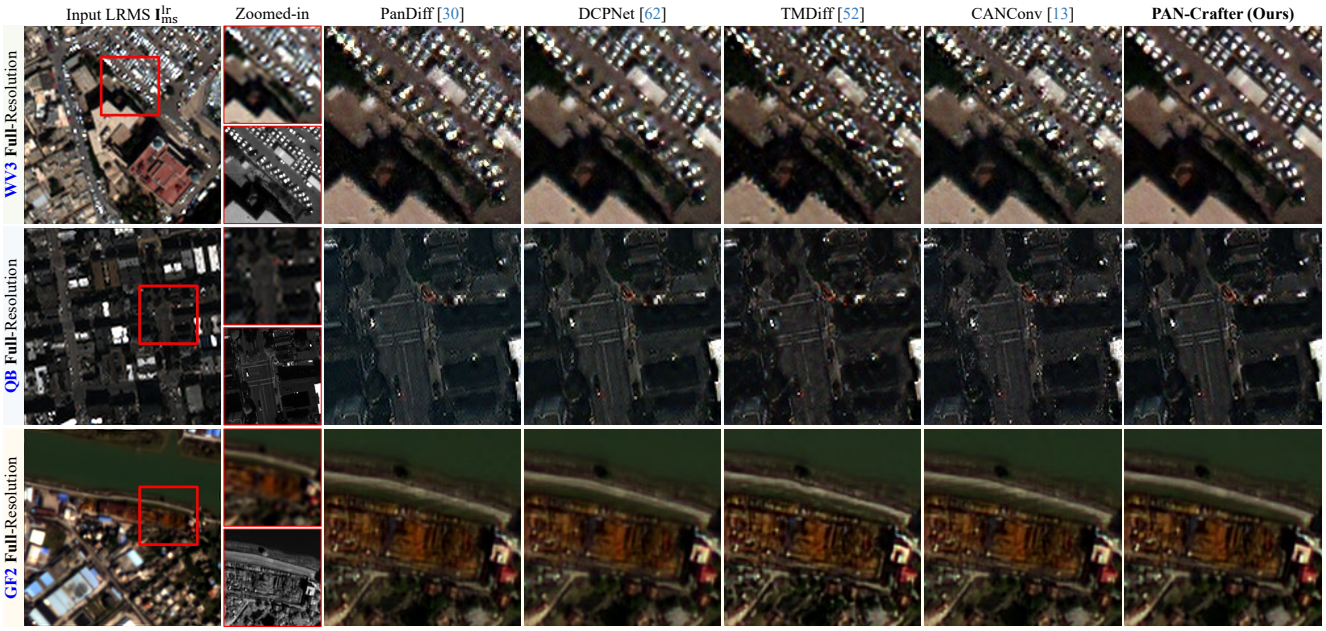

Figure 8. Visual comparison of PAN-Sharpening (PS) results on the WV3, QB, and GF2 datasets at full-resolution. The leftmost column shows the input LRMS images, with red boxes indicating zoomed-in regions for both LRMS and PAN images. Our PAN-Crafter method generates pan-sharpened images with minimal artifacts, particularly around buildings and vehicles, whereas other methods frequently produce blurry or distorted outputs.

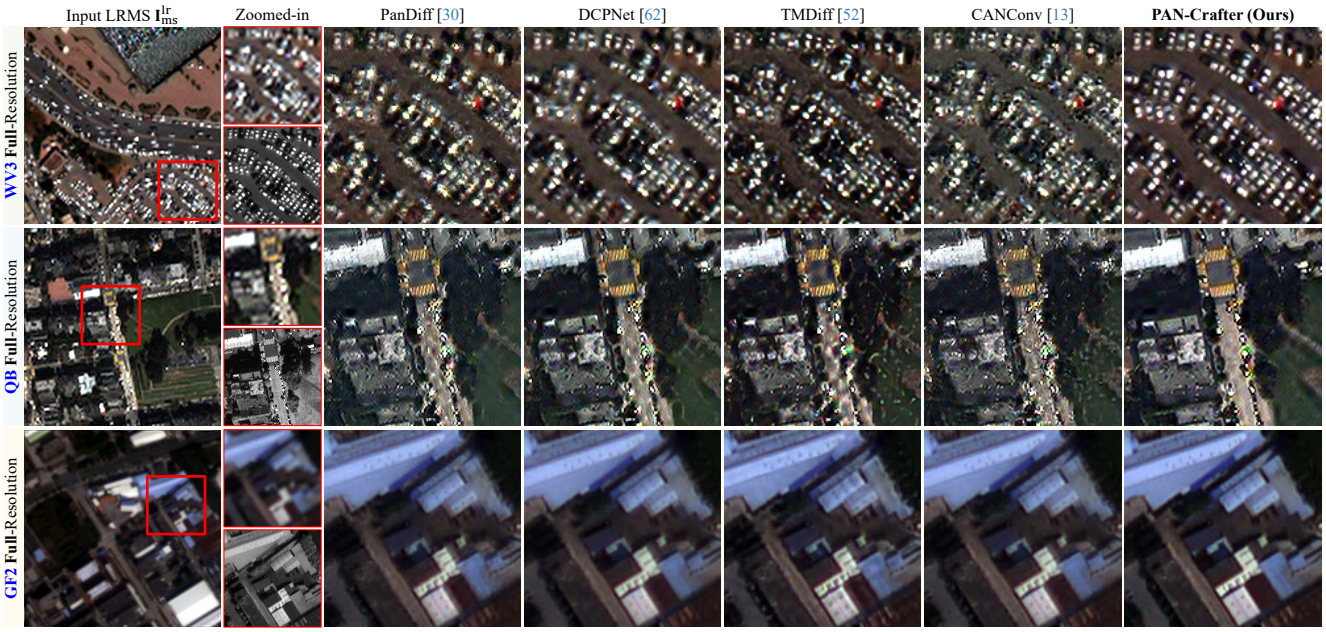

Figure 9. Visual comparison of PAN-Sharpening (PS) results on the WV3, QB, and GF2 datasets at full-resolution. The leftmost column shows the input LRMS images, with red boxes indicating zoomed-in regions for both LRMS and PAN images. Our PAN-Crafter method generates high-quality of pan-sharpened images with minimal artifacts, particularly around vehicles and crosswalks, whereas other methods frequently produce blurry or distorted outputs.

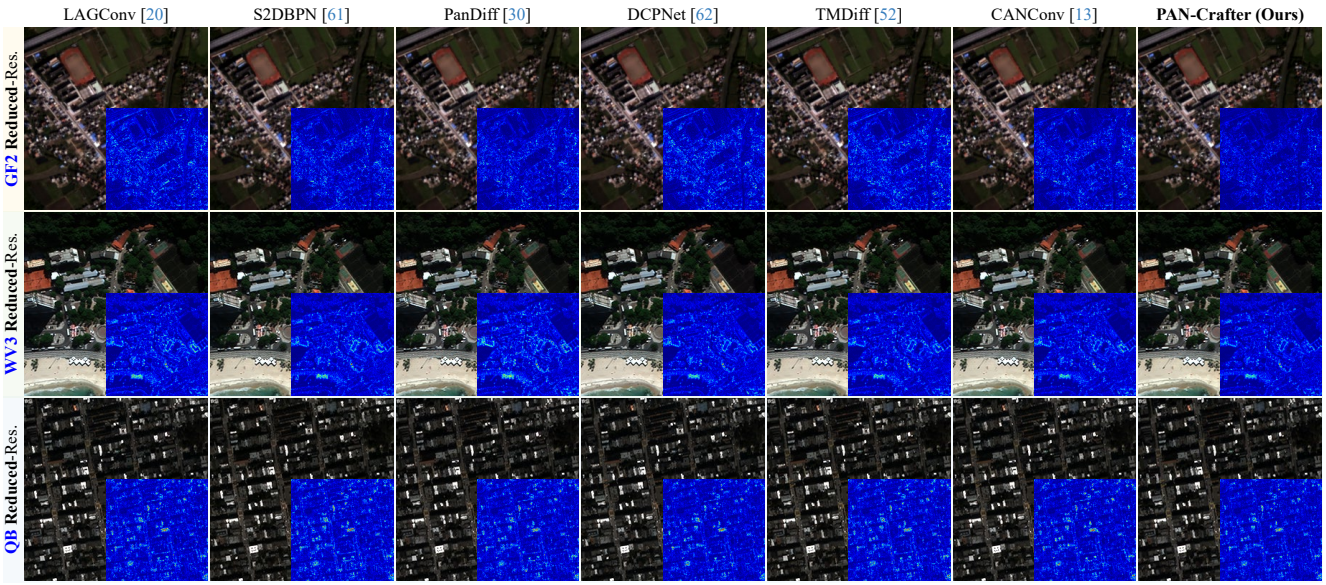

Figure 10. Visual comparison of PS results on the GF2 and QB datasets at reduced-resolution. The blue-colored insets represent error maps computed against the ground truth (GT), where brighter regions indicate higher reconstruction errors.

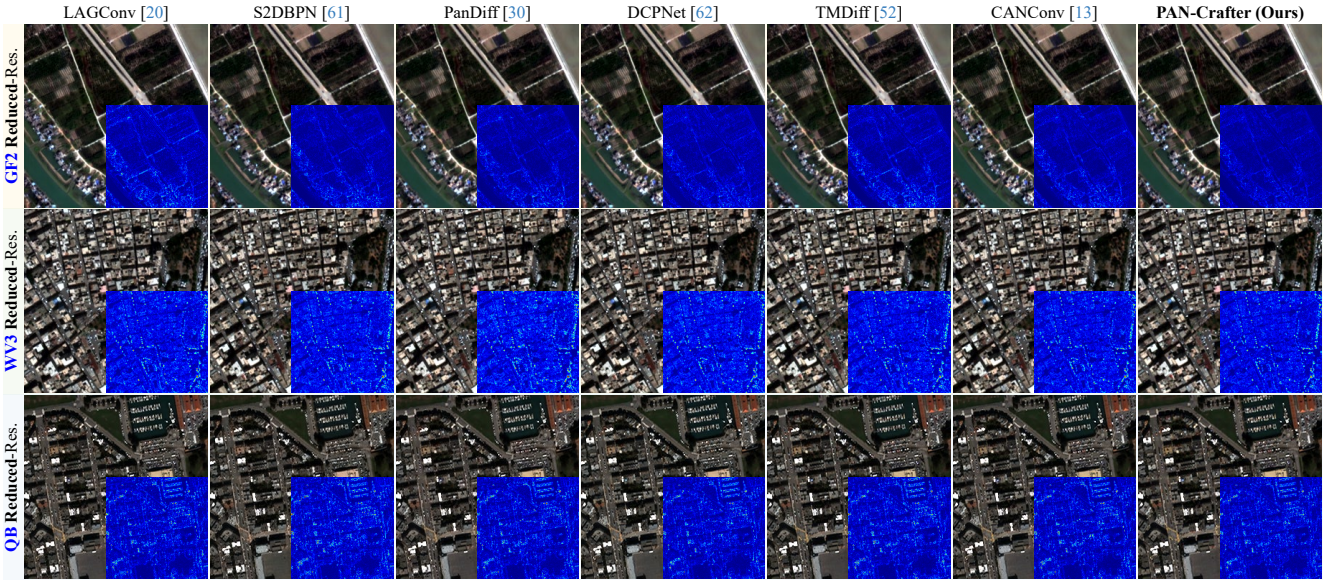

Figure 11. Visual comparison of PS results on the GF2 and QB datasets at reduced-resolution. The blue-colored insets represent error maps computed against the ground truth (GT), where brighter regions indicate higher reconstruction errors.

Methods	Tasks	Attn. types	Characteristics
Siamtrans [24]	Multi-frame restoration	Global	Only cross-attention with PE
TTSR [54]	Reference-based SR	Global	Only cross-attention with PE
Ours	PAN-sharpening	Local	MARs-mode-dependent cross-/self-attention

Table 5. Comparison with ours CM3A with existing cross-attention methods.

Figure 12. 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 proposed PAN-Crafter significantly outperforms the existing methods by effectively preserving both fine structural and spectral details of the input MS and PAN images

WV3 Dataset	Full-Resolution			Reduced-Resolution					
Methods	HQNR↑	D_s↓	D_λ↓	ERGAS↓	SCC↑	SAM↓	Q8↑	PSNR↑	SSIM↑
PanNet [57]	0.918 ± 0.031	0.049 ± 0.019	0.035 ± 0.014	2.538 ± 0.597	0.979 ± 0.006	3.402 ± 0.672	0.913 ± 0.087	36.148 ± 1.958	0.966 ± 0.011
MSDCNN [58]	0.924 ± 0.030	0.050 ± 0.020	0.028 ± 0.013	2.489 ± 0.620	0.979 ± 0.007	3.300 ± 0.654	0.914 ± 0.087	36.329 ± 1.748	0.967 ± 0.010
FusionNet [50]	0.920 ± 0.030	0.053 ± 0.021	0.029 ± 0.011	2.428 ± 0.621	0.981 ± 0.007	3.188 ± 0.628	0.916 ± 0.087	36.569 ± 1.666	0.968 ± 0.009
LAGNet [20]	0.915 ± 0.033	0.055 ± 0.023	0.033 ± 0.012	2.380 ± 0.617	0.981 ± 0.007	3.153 ± 0.608	0.916 ± 0.087	36.732 ± 1.723	0.970 ± 0.009
S2DBPN [61]	0.946 ± 0.018	0.030 ± 0.010	0.025 ± 0.010	2.245 ± 0.541	0.985 ± 0.005	3.019 ± 0.588	0.917 ± 0.091	37.216 ± 1.888	0.972 ± 0.009
PanDiff [30]	0.952 ± 0.009	0.034 ± 0.005	**0.014** ± 0.005	2.276 ± 0.545	0.984 ± 0.004	3.058 ± 0.567	0.913 ± 0.084	37.029 ± 1.796	0.971 ± 0.008
DCPNet [62]	0.923 ± 0.027	0.036 ± 0.012	0.043 ± 0.018	2.301 ± 0.569	0.984 ± 0.005	3.083 ± 0.537	0.915 ± 0.092	37.009 ± 1.735	0.972 ± 0.008
TMDiff [52]	0.924 ± 0.015	0.059 ± 0.009	0.018 ± 0.007	2.151 ± 0.458	0.986 ± 0.004	2.885 ± 0.549	0.915 ± 0.086	37.477 ± 1.923	0.973 ± 0.008
CANConv [13]	0.951 ± 0.013	0.030 ± 0.008	0.020 ± 0.008	2.163 ± 0.481	0.985 ± 0.005	2.927 ± 0.536	0.918 ± 0.082	37.441 ± 1.788	0.973 ± 0.008
PAN-Crafter	**0.958** ± 0.009	**0.027** ± 0.004	0.016 ± 0.006	**2.040** ± 0.459	**0.988** ± 0.003	**2.787** ± 0.523	**0.922** ± 0.082	**37.956** ± 1.771	**0.976** ± 0.006

Table 6. Quantitative comparison of deep learning-based PS methods on the WV3 dataset. **Red** indicates the best performance.

GF2 Dataset	Full-Resolution			Reduced-Resolution					
Methods	HQNR↑	D_s↓	D_λ↓	ERGAS↓	SCC↑	SAM↓	Q4↑	PSNR↑	SSIM↑
PanNet [57]	0.929 ± 0.013	0.052 ± 0.009	0.020 ± 0.012	1.038 ± 0.214	0.975 ± 0.006	1.050 ± 0.209	0.963 ± 0.009	39.197 ± 2.009	0.959 ± 0.011
MSDCNN [58]	0.898 ± 0.016	0.079 ± 0.011	0.026 ± 0.014	0.862 ± 0.141	0.983 ± 0.003	0.946 ± 0.166	0.972 ± 0.009	40.730 ± 1.564	0.971 ± 0.006
FusionNet[50]	0.865 ± 0.018	0.105 ± 0.013	0.034 ± 0.013	0.960 ± 0.193	0.980 ± 0.005	0.971 ± 0.195	0.967 ± 0.008	39.866 ± 1.955	0.966 ± 0.009
LAGNet [20]	0.895 ± 0.021	0.078 ± 0.013	0.030 ± 0.014	0.816 ± 0.121	0.985 ± 0.003	0.886 ± 0.140	0.974 ± 0.009	41.147 ± 1.384	0.974 ± 0.005
S2DBPN [61]	0.935 ± 0.011	0.046 ± 0.007	0.020 ± 0.012	0.686 ± 0.125	0.990 ± 0.002	0.772 ± 0.149	0.981 ± 0.007	42.686 ± 1.676	0.980 ± 0.005
PanDiff [30]	0.936 ± 0.011	0.045 ± 0.009	0.020 ± 0.014	0.674 ± 0.110	0.990 ± 0.002	0.767 ± 0.134	0.981 ± 0.007	42.827 ± 1.462	0.980 ± 0.005
DCPNet [62]	0.953 ± 0.019	0.024 ± 0.008	0.024 ± 0.022	0.724 ± 0.138	0.988 ± 0.003	0.806 ± 0.153	0.980 ± 0.007	42.312 ± 1.682	0.979 ± 0.005
TMDiff [52]	0.942 ± 0.016	0.030 ± 0.010	0.029 ± 0.011	0.754 ± 0.143	0.988 ± 0.003	0.764 ± 0.155	0.979 ± 0.007	41.896 ± 1.765	0.977 ± 0.005
CANConv [13]	0.919 ± 0.011	0.063 ± 0.009	**0.019** ± 0.010	0.653 ± 0.124	0.991 ± 0.002	0.722 ± 0.138	0.983 ± 0.006	43.166 ± 1.705	0.982 ± 0.004
PAN-Crafter	**0.964** ± 0.015	**0.017** ± 0.007	0.020 ± 0.013	**0.552** ± 0.093	**0.994** ± 0.001	**0.596** ± 0.110	**0.988** ± 0.006	**45.076** ± 1.610	**0.988** ± 0.003

Table 7. Quantitative comparison of deep learning-based PS methods on the GF2 dataset. **Red** indicates the best performance.

QB Dataset	Full-Resolution			Reduced-Resolution					
Methods	HQNR↑	D_s↓	D_λ↓	ERGAS↓	SCC↑	SAM↓	Q4↑	PSNR↑	SSIM↑
PanNet [57]	0.851 ± 0.035	0.092 ± 0.021	0.063 ± 0.019	4.856 ± 0.590	0.966 ± 0.015	5.273 ± 0.946	0.911 ± 0.094	35.563 ± 1.930	0.939 ± 0.012
MSDCNN [58]	0.888 ± 0.037	0.058 ± 0.027	0.058 ± 0.014	4.074 ± 0.244	0.977 ± 0.010	4.828 ± 0.824	0.925 ± 0.098	37.040 ± 1.778	0.954 ± 0.007
FusionNet [50]	0.853 ± 0.041	0.079 ± 0.025	0.074 ± 0.022	4.183 ± 0.266	0.975 ± 0.011	4.892 ± 0.822	0.923 ± 0.100	36.821 ± 1.765	0.952 ± 0.007
LAGNet [20]	0.892 ± 0.024	**0.035** ± 0.009	0.075 ± 0.019	3.845 ± 0.323	0.980 ± 0.009	4.682 ± 0.785	0.930 ± 0.095	37.565 ± 1.721	0.958 ± 0.006
S2DBPN [61]	0.908 ± 0.044	0.036 ± 0.023	0.059 ± 0.026	3.956 ± 0.291	0.980 ± 0.008	4.849 ± 0.822	0.928 ± 0.093	37.314 ± 1.782	0.956 ± 0.006
PanDiff [30]	0.919 ± 0.010	0.055 ± 0.012	**0.028** ± 0.011	3.723 ± 0.280	0.982 ± 0.007	4.611 ± 0.768	0.935 ± 0.084	37.842 ± 1.721	0.959 ± 0.006
DCPNet [62]	0.880 ± 0.013	0.073 ± 0.013	0.051 ± 0.017	3.618 ± 0.313	0.983 ± 0.010	**4.420** ± 0.710	0.935 ± 0.095	38.079 ± 1.454	**0.963** ± 0.004
TMDiff [52]	0.901 ± 0.011	0.068 ± 0.012	0.034 ± 0.016	3.804 ± 0.279	0.981 ± 0.008	4.627 ± 0.814	0.930 ± 0.096	37.642 ± 1.831	0.958 ± 0.006
CANConv [13]	0.893 ± 0.010	0.070 ± 0.017	0.039 ± 0.012	3.740 ± 0.304	0.982 ± 0.007	4.554 ± 0.788	0.935 ± 0.087	37.795 ± 1.801	0.960 ± 0.006
PAN-Crafter	**0.920** ± 0.027	0.039 ± 0.020	0.043 ± 0.011	**3.570** ± 0.286	**0.984** ± 0.008	4.426 ± 0.740	**0.938** ± 0.087	**38.195** ± 1.597	**0.963** ± 0.005

Table 8. Quantitative comparison of deep learning-based PS methods on the QB dataset. **Red** indicates the best performance.

WV2 Dataset	Full-Resolution (**Unseen** satellite dataset)			Reduced-Resolution (**Unseen** satellite dataset)					
Methods	HQNR↑	D_s↓	D_λ↓	ERGAS↓	SCC↑	SAM↓	Q8↑	PSNR↑	SSIM↑
PanNet [57]	0.875 ± 0.064	0.032 ± 0.005	0.096 ± 0.066	5.481 ± 0.326	0.876 ± 0.018	7.040 ± 0.417	0.786 ± 0.084	27.120 ± 1.827	0.770 ± 0.053
MSDCNN [58]	0.862 ± 0.050	0.029 ± 0.013	0.113 ± 0.041	4.930 ± 0.378	0.905 ± 0.009	5.898 ± 0.490	0.812 ± 0.090	27.901 ± 1.812	0.804 ± 0.040
FusionNet [50]	0.862 ± 0.034	0.038 ± 0.005	0.104 ± 0.032	5.100 ± 0.367	0.902 ± 0.011	6.118 ± 0.533	0.786 ± 0.083	27.616 ± 1.765	0.788 ± 0.042
LAGNet [20]	0.902 ± 0.045	**0.024** ± 0.018	0.076 ± 0.032	5.133 ± 0.432	0.885 ± 0.015	6.094 ± 0.559	0.792 ± 0.081	27.525 ± 2.008	0.777 ± 0.054
S2DBPN [61]	0.813 ± 0.066	0.065 ± 0.019	0.129 ± 0.080	5.703 ± 0.257	0.915 ± 0.011	7.063 ± 0.421	0.805 ± 0.092	26.748 ± 1.892	0.804 ± 0.041
DCPNet [62]	0.797 ± 0.134	0.034 ± 0.022	0.176 ± 0.129	5.507 ± 0.264	**0.931** ± 0.009	10.174 ± 1.115	0.843 ± 0.094	27.063 ± 1.541	0.855 ± 0.021
PanDiff [30]	0.932 ± 0.019	0.043 ± 0.010	0.026 ± 0.019	4.291 ± 0.418	0.916 ± 0.010	5.430 ± 0.601	0.840 ± 0.087	28.964 ± 1.709	0.832 ± 0.033
TMDiff [52]	0.874 ± 0.013	0.088 ± 0.021	0.042 ± 0.020	5.157 ± 0.604	0.875 ± 0.008	6.087 ± 0.786	0.777 ± 0.079	27.473 ± 1.634	0.762 ± 0.045
CANConv [13]	0.876 ± 0.044	0.060 ± 0.022	0.068 ± 0.049	4.328 ± 0.413	0.918 ± 0.008	5.481 ± 0.595	0.841 ± 0.087	29.005 ± 1.719	0.837 ± 0.031
PAN-Crafter	**0.942** ± 0.019	0.036 ± 0.010	**0.022** ± 0.008	**4.169** ± 0.397	0.924 ± 0.009	**5.078** ± 0.561	**0.846** ± 0.085	**29.276** ± 1.621	**0.839** ± 0.029

Table 9. Quantitative comparison of deep learning-based PS methods on the unseen WV2 dataset. All models are trained on WV3 and evaluated on WV2 to assess real-world generalization. **Red** indicate the best performance in each metric.

Methods	LAGConv [20]	S2DBPN [61]	PanDiff [30]	DCPNet [62]	TMDiff [52]	CANConv [13]	**PAN-Crafter**
Time (s)	0.004	0.005	2.955	0.109	9.997	0.451	0.009
Memory (MB)	3360.1	2444.0	2383.6	7386.8	10147.4	2777.6	1751.9
FLOPs (G)	8.43	158.94	62.07	105.40	1284.42	52.21	79.03
Params. (M)	0.15	16.19	9.52	1.414	154.10	0.79	7.17

Table 10. Computational efficiency comparison of deep learning-based PS methods. We report inference time (s), memory usage (MB), FLOPs (G), and parameter count (M).

WV3 Dataset		Full-Resolution			Reduced-Resolution						Inference Time↓ (s)	Memory↓ (GB)
CM3A	MARs	HQNR↑	D_s↓	D_λ↓	ERGAS↓	SCC↑	SAM↓	Q8↑	PSNR↑	SSIM↑		
		0.948	0.035	0.018	2.232	0.985	2.980	0.913	37.245	0.972	0.006	1.537
✓		0.949	0.035	0.016	2.212	0.985	2.970	0.915	37.285	0.973	0.007	1.556
	✓	0.956	0.028	0.017	2.122	0.987	2.873	0.919	37.602	0.974	0.009	1.701
✓	✓	**0.958**	**0.027**	**0.016**	**2.040**	**0.988**	**2.787**	**0.922**	**37.956**	**0.976**	0.009	1.711

Table 11. Ablation studies on CM3A and MARs on the WV3 dataset.

GF2 Dataset		Full-Resolution			Reduced-Resolution					
CM3A	MARs	HQNR↑	D_s↓	D_λ↓	ERGAS↓	SCC↑	SAM↓	Q4↑	PSNR↑	SSIM↑
		0.959	0.021	0.021	0.632	0.992	0.723	0.984	43.476	0.984
✓		0.953	0.025	0.023	0.624	0.992	0.718	0.984	43.618	0.984
	✓	0.945	0.032	0.023	0.574	0.993	0.651	0.986	44.298	0.986
✓	✓	**0.964**	**0.017**	**0.020**	**0.552**	**0.994**	**0.596**	**0.988**	**45.076**	**0.988**

Table 12. Ablation studies on CM3A and MARs on the GF2 dataset.

QB Dataset		Full-Resolution			Reduced-Resolution					
CM3A	MARs	HQNR↑	D_s↓	D_λ↓	ERGAS↓	SCC↑	SAM↓	Q4↑	PSNR↑	SSIM↑
		0.856	0.086	0.064	4.907	0.977	5.200	0.923	35.476	0.947
✓		0.879	0.062	0.063	4.869	0.975	5.168	0.922	35.538	0.947
	✓	0.896	0.047	0.060	3.857	0.980	4.661	0.930	37.557	0.959
✓	✓	**0.920**	**0.039**	**0.043**	**3.570**	**0.984**	**4.426**	**0.938**	**38.195**	**0.963**

Table 13. Ablation studies on CM3A and MARs on the QB dataset.

α	β, γ	**WV3 / GF2 / QB** Datasets			
		HQNR↑	ERGAS↓	SAM↓	PSNR↑
		0.945 / 0.951 / 0.908	2.214 / 0.623 / 3.758	2.901 / 0.642 / 4.523	37.210 / 44.321 / 37.842
✓		0.949 / 0.957 / 0.915	2.150 / 0.589 / 3.669	2.829 / 0.618 / 4.472	37.562 / 44.758 / 38.021
	✓	0.947 / 0.956 / 0.913	2.185 / 0.601 / 3.690	2.841 / 0.624 / 4.488	37.433 / 44.612 / 37.935
✓	✓	**0.958 / 0.964 / 0.920**	**2.040 / 0.552 / 3.570**	**2.787 / 0.596 / 4.426**	**37.956 / 45.076 / 38.195**

Table 14. Ablation studies on α, β, and γ on the WV3, GF2, and QB datasets.

k	**WV3 / GF2 / QB** Datasets					
	HQNR↑	ERGAS↓	SAM↓	PSNR↑	Time↓	Memory↓
3	**0.958** / 0.964 / 0.920	2.040 / **0.552** / **3.570**	2.787 / **0.596** / **4.426**	37.956 / **45.076** / 38.195	**0.009**	**1.711**
5	0.953 / **0.965** / 0.919	**2.021** / 0.555 / 3.577	**2.785** / 0.600 / 4.433	**37.966** / 45.001 / **38.201**	0.019	3.429
7	0.955 / 0.961 / **0.921**	2.033 / 0.553 / 3.575	2.790 / 0.599 / 4.429	37.949 / 45.010 / 38.190	0.042	5.243

Table 15. Ablation studies on k on the WV3, GF2, and QB datasets.

Training Strategy	**WV3 / GF2 / QB** Datasets			
	HQNR↑	ERGAS↓	SAM↓	PSNR↑
w/o MARs	0.949 / 0.953 / 0.879	2.212 / 0.624 / 4.869	2.970 / 0.718 / 5.168	37.285 / 43.618 / 35.538
Two-stage	0.945 / 0.953 / 0.890	2.199 / 0.602 / 4.551	2.899 / 0.688 / 4.907	37.345 / 43.921 / 36.081
w/ MARs	**0.958** / **0.964** / **0.920**	**2.040** / **0.552** / **3.570**	**2.787** / **0.596** / **4.426**	**37.956** / **45.076** / **38.195**

Table 16. Ablation studies on training strategy on the WV3, GF2, and QB datasets.

Layer Type	**WV3 / GF2 / QB** Datasets					
	HQNR↑	ERGAS↓	SAM↓	PSNR↑	Time↓	Memory↓
Attn.	**0.948** / **0.959** / **0.856**	**2.232** / **0.632** / **4.907**	**2.980** / **0.723** / **5.200**	**37.245** / **43.476** / **35.476**	0.006	1.537
Conv.	0.937 / 0.943 / 0.850	2.322 / 0.741 / 5.142	3.120 / 0.831 / 5.463	36.988 / 42.590 / 35.218	**0.004**	**1.209**

Table 17. Ablation studies on layer type on the WV3, GF2, and QB datasets.