Spatial-Spectral Residuals Informed Diffusion Neural Operator for Pan-sharpening

Supplementary Material

We organize the supplementary material as follows:
Structure of the Auxiliary Predictor: We introduce the structure of the auxiliary predictor in the main paper.
Plug and Play Validation: We proceed to validate the proposed spatial-spectral residual integration mechanism by applying it to a representative diffusion model.
Additional Experimental Results: We present additional experimental results on three datasets, including quantitative evaluations, qualitative comparisons, and feature maps.

S1. Structure of the Auxiliary Predictor

As illustrated in Figure A1, the auxiliary predictor (corresponding to Figure 3 in our main paper) consists of 4 convolutional residual blocks. It is designed to extract the cross-modality feature $\mathbf{V}$ and generate an initial HRMS estimate, with the latter serving as a replacement for the high-resolution reference $\mathbf{H}$ in Stage I. Specifically, the auxiliary predictor takes as input the cross-modality PAN and LRMS images together with their spatial residual. To obtain this residual, we first transform the LRMS image $\mathbf{L}$ into a single channel pseudo PAN image $\widehat{\mathbf{P}}$, and then subtract it from the real PAN image $\mathbf{P}$, supplying the spatial details absent in $\mathbf{L}$.

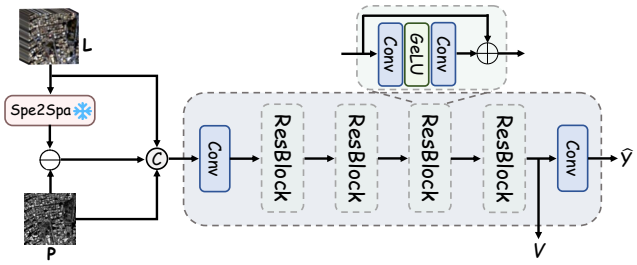

Figure A1. Structure of the auxiliary predictor in the main paper.

S2. Plug and Play Validation

We apply our proposed spatial-spectral residual integration strategy to another representative diffusion model [1] to validate its efficacy. As shown in Figure A2, integrating both spatial and spectral residuals leads to a notable improvement in PSNR over the original model. More importantly, this integration introduces only negligible overhead in terms of FLOPs, Memory, and Inference Time. These results strongly affirm the effectiveness and general applicability of the proposed residual integration approach.

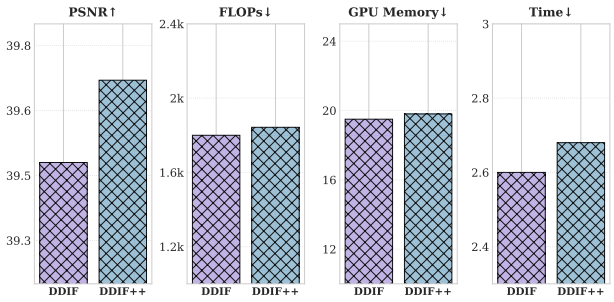

Figure A2. Ablation experiment evaluating the efficacy of our spatial-spectral residual integration strategy applied to DDIF [1]. "++" indicates the model integrating the spatial-spectral residuals.

S3. Additional Experimental Results

S3.1. Evaluation on Reduced-resolution Scene

We begin by evaluating our model on the reduced resolution GF2 and QB datasets. As summarized in the left panels of Table 1 and Table 2, our method consistently outperforms state-of-the-art approaches across all metrics for both the GF2 and QB datasets. In particular, our method achieves the optimal results in PSNR and SAM on the GF2 and QB datasets, indicating superior spatial and spectral fusion quality. Overall, these quantitative results confirm our model's superiority in preserving spectral fidelity and reconstructing spatial textures.

Visual comparisons further support these quantitative findings. As shown in Figure A4, the images generated by our model exhibit fewer visual artifacts and closer resemblance to the ground truth. This is reinforced by the mean absolute error (MAE) maps between the fused results and the ground truth, where enlarged residue regions contain notably fewer bright spots. Together, these visual observations align with and strengthen the quantitative conclusions.

We further visualize the feature maps across different denoiser layers using samples from three benchmark datasets. It is clear that the feature maps demonstrate a progressive enhancement in clarity and detail as the number of denoiser layers increases, as illustrated in Figure A3.

S3.2. Evaluation on Full-resolution Scene

We further evaluate our method under full resolution scenarios to verify its practical applicability and generalization ability. The results in the right panels of Table 1 and Table 2 show that our model maintains competitive performance on both the GF2 and QB datasets, mirroring the positive trends observed at reduced resolution. Notably, it achieves the highest HQNR scores on both datasets, outperforming

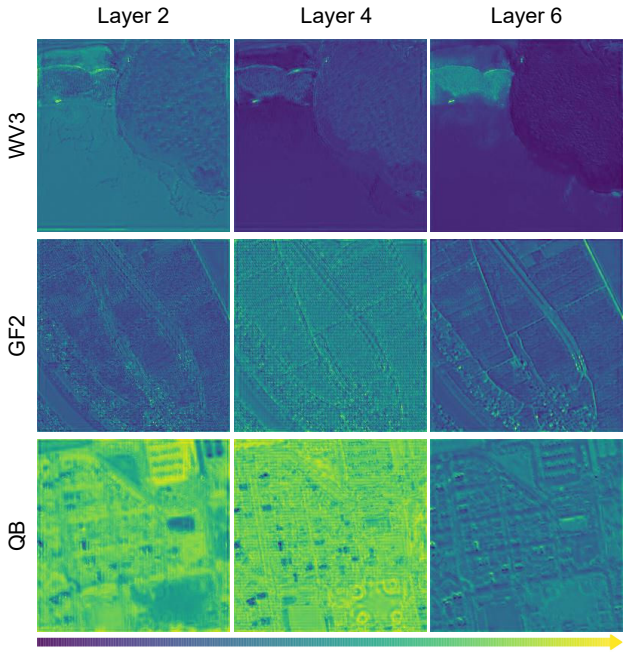

Figure A3. Feature maps across different denoiser layers.

traditional and deep-learning baselines, which indicates the superior spatial and spectral fidelity. These full resolution results highlight the robust generalization of our model in real-world applications.

Figure A5 presents visual comparisons on full resolution WV3, GF2 and QB samples. In the odd rows, our method produces images with finer spatial textures and visually consistent spectral characteristics. The even rows demonstrate the corresponding HQNR maps, where our method exhibits significantly fewer bright spots, consistent with its higher HQNR scores and indicating better spatial-spectral fidelity. These visual observations corroborate the quantitative results, jointly demonstrating the effectiveness of our approach in real-world high-resolution scenarios.

References

[1] Zihan Cao, Shiqi Cao, Liang-Jian Deng, Xiao Wu, Junming Hou, and Gemine Vivone. Diffusion model with disentangled modulations for sharpening multispectral and hyperspectral images. *Information Fusion*, 104:102158, 2024. 1

[2] Liang-Jian Deng, Gemine Vivone, Cheng Jin, and Jocelyn Chanussot. Detail injection-based deep convolutional neural networks for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 59(8):6995–7010, 2020. 3, 4

[3] 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. 3, 4

[4] Jie Huang, Haorui Chen, Jiaxuan Ren, Siran Peng, and Liangjian Deng. A general adaptive dual-level weighting mechanism for remote sensing pansharpening. *arXiv preprint arXiv:2503.13214*, 2025. 3, 4

[5] Zhi-Rui 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. AAAI Press, 2022. 3, 4

[6] Simone Lolli, Luciano Alparone, Andrea Garzelli, and Gemine Vivone. Haze correction for contrast-based multispectral pansharpening. *IEEE Geoscience and Remote Sensing Letters*, 14(12):2255–2259, 2017. 3, 4

[7] 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. 3, 4

[8] Jiangtong Tan, Jie Huang, Naishan Zheng, Man Zhou, Keyu Yan, Danfeng Hong, and Feng Zhao. Revisiting spatial-frequency information integration from a hierarchical perspective for panchromatic and multi-spectral image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25922–25931, 2024. 3, 4

[9] Gemine Vivone. Robust band-dependent spatial-detail approaches for panchromatic sharpening. *IEEE transactions on Geoscience and Remote Sensing*, 57(9):6421–6433, 2019. 3, 4

[10] Zhong-Cheng Wu, Ting-Zhu Huang, Liang-Jian Deng, Jie Huang, Jocelyn Chanussot, and Gemine Vivone. Lrtcfpan: Low-rank tensor completion based framework for pansharpening. *IEEE Transactions on Image Processing*, 32:1640–1655, 2023. 3, 4

[11] Man Zhou, Jie Huang, Chun-Le Guo, and Chongyi Li. Fourmer: An efficient global modeling paradigm for image restoration. In *International conference on machine learning*, pages 42589–42601. PMLR, 2023. 3, 4

[12] Man Zhou, Naishan Zheng, Xuanhua He, Danfeng Hong, and Jocelyn Chanussot. Probing synergistic high-order interaction for multi-modal image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3, 4

Method	Reduced Resolution				Full Resolution		
	PSNR(± std)	SAM(± std)	ERGAS(± std)	$Q2^n$(± std)	D_λ(± std)	D_s(± std)	HQNR(± std)
BDSD-PC [9]	35.180±2.317	1.681±0.360	1.667±0.445	0.892±0.035	0.076±0.030	0.155±0.028	0.781±0.041
BT-H [6]	36.054±2.236	1.649±0.360	1.528±0.409	0.918±0.025	0.060±0.025	0.131±0.019	0.817±0.031
LRTCFPan [10]	37.599±2.331	1.298±0.312	1.272±0.343	0.935±0.030	0.033±0.027	0.090±0.014	0.881±0.023
FusionNet [2]	39.639±2.270	0.974±0.212	0.988±0.222	0.964±0.009	0.035±0.012	0.101±0.013	0.867±0.018
LAGConv [5]	42.735±1.447	0.786±0.148	0.687±0.113	0.980±0.009	0.028±0.013	0.079±0.014	0.895±0.020
Fourmer [11]	40.670±1.903	0.976±0.209	0.885±0.185	0.970±0.011	0.047±0.039	**0.038±0.010**	0.917±0.035
HFIN [8]	42.189±1.752	0.843±0.148	0.735±0.126	0.977±0.011	0.027±0.020	0.062±0.009	0.912±0.018
HOIF [12]	40.982±1.802	0.943±0.205	0.841±0.162	0.974±0.009	0.029±0.015	0.051±0.011	0.922±0.019
PanDiff [7]	42.326±1.635	0.875±0.133	0.727±0.115	0.981±0.008	0.028±0.020	0.073±0.010	0.902±0.021
LFormer [3]	44.196±1.800	0.648±0.130	0.578±0.112	0.985±0.007	0.021±0.010	0.050±0.001	0.930±0.013
ADWM [4]	43.884±1.714	0.672±0.130	0.597±0.107	0.985±0.006	0.022±0.012	0.052±0.011	0.928±0.014
SRINO (ours)	**44.228±1.709**	**0.646±0.129**	**0.573±0.101**	**0.985±0.006**	**0.019±0.010**	0.044±0.008	**0.938±0.011**
Ideal value	∞	**0**	**0**	**1**	**0**	**0**	**1**

Table 1. Quantitative results for reduced and full resolution GF2 samples, comparing several representative state-of-the-art methods. Bold: Best; Underline: Second best.

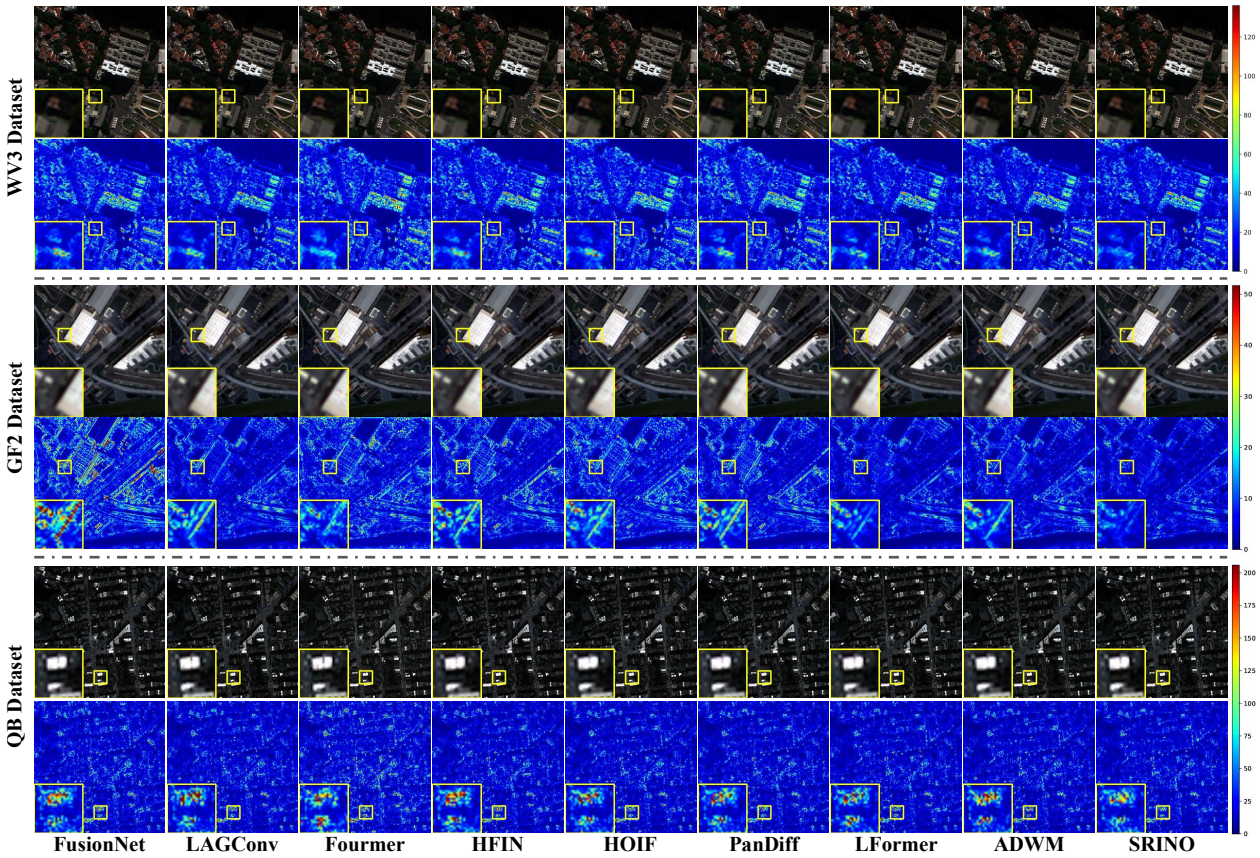

Figure A4. The visual results (odd rows) and the corresponding mean absolute error (MAE) maps (even rows) of all compared DL-based methods on reduced resolution samples from the WV3, GF2 and QB datasets, respectively.

Method	Reduced Resolution				Full Resolution		
	PSNR(± std)	SAM(± std)	ERGAS(± std)	$Q2^n$(± std)	D_λ(± std)	D_s(± std)	HQNR(± std)
BDSD-PC [9]	32.547±3.202	8.089±1.980	7.515±0.800	0.831±0.090	0.198±0.033	0.164±0.048	0.672±0.058
BT-H [6]	32.648±3.273	7.194±1.552	7.401±0.838	0.833±0.088	0.230±0.072	0.165±0.017	0.643±0.065
LRTCFPan [10]	33.260±3.272	7.187±1.711	6.928±0.812	0.855±0.087	**0.023±0.012**	0.071±0.035	0.909±0.044
FusionNet [2]	37.532±2.518	4.923±0.908	4.159±0.321	0.925±0.090	0.057±0.018	0.052±0.009	0.894±0.021
LAGConv [5]	38.181±2.456	4.547±0.830	3.826±0.420	0.934±0.088	0.086±0.024	0.068±0.014	0.852±0.018
Fourmer [11]	36.797±2.597	5.079±1.009	4.494±0.410	0.924±0.080	0.033±0.013	0.050±0.010	0.920±0.018
HFIN [8]	38.247±2.403	4.542±0.805	3.813±0.322	0.934±0.085	0.067±0.025	0.078±0.019	0.860±0.018
HOIF [12]	38.242±2.119	4.521±0.811	3.825±0.510	0.933±0.094	0.077±0.026	0.059±0.025	0.869±0.045
PanDiff [7]	38.538±2.401	4.524±0.792	3.688±0.343	0.937±0.084	0.058±0.022	0.064±0.025	0.882±0.041
LFormer [3]	38.674±2.370	4.382±0.789	3.616±0.314	0.939±0.082	0.037±0.016	0.089±0.026	0.877±0.021
ADWM [4]	38.466±2.420	4.450±0.809	3.705±0.346	0.937±0.085	0.064±0.025	**0.024±0.016**	0.914±0.036
SRINO (ours)	**38.864±2.346**	**4.351±0.773**	**3.542±0.313**	**0.940±0.085**	0.041±0.013	0.032±0.013	**0.927±0.024**
Ideal value	∞	**0**	**0**	**1**	**0**	**0**	**1**

Table 2. Quantitative results for reduced and full resolution QB samples, comparing several representative state-of-the-art methods. Bold: Best; Underline: Second best.

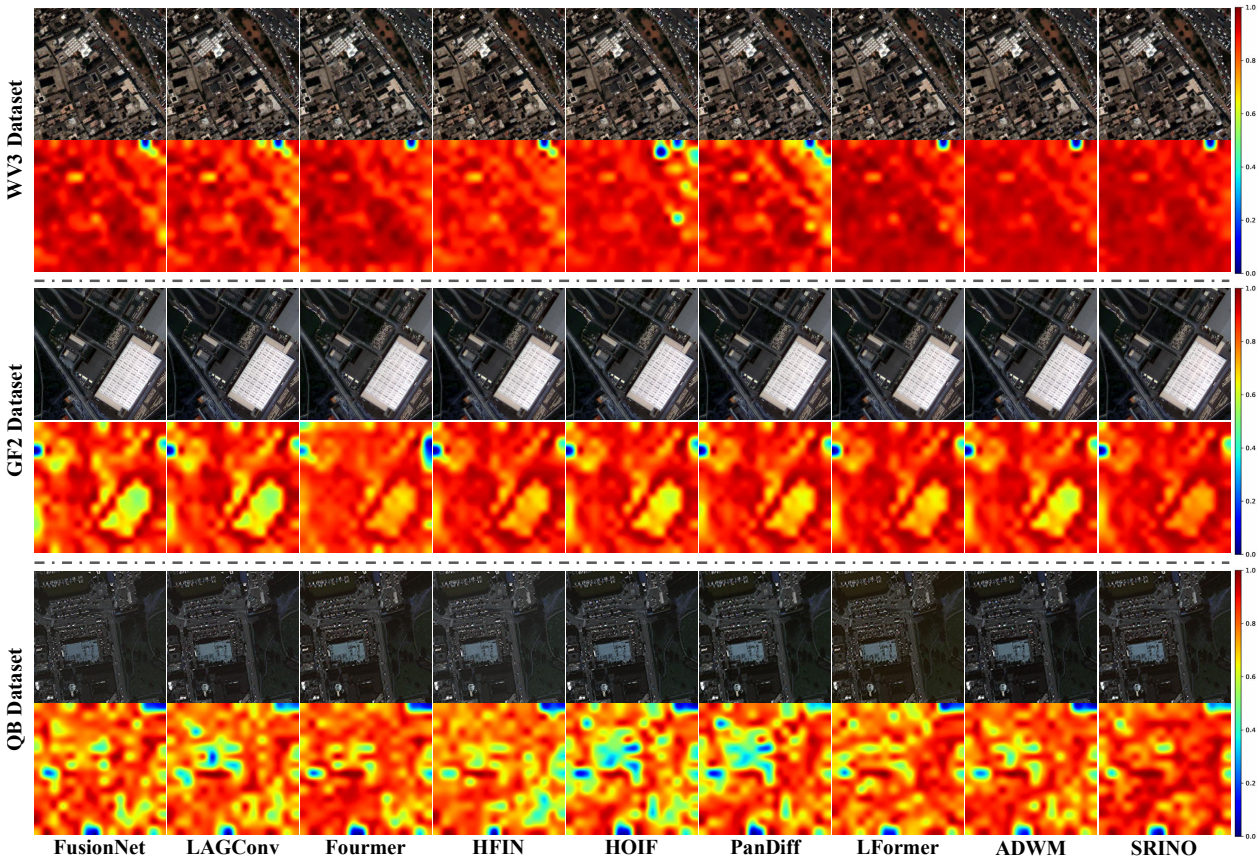

Figure A5. The visual results (odd rows) and the corresponding HQNR maps (even rows) of all compared DL-based methods on full-resolution samples from the WV3, GF2 and QB datasets, respectively.