Unfolding-Associative Encoder-Decoder Network with Progressive Alignment for Pansharpening

Supplementary Material

Overview

All source code will be made publicly available for further research.

In this supplementary material, we present more details about our study, including:

- Sec. 1 provides the mathematical model of UED-Net based on C-SALSA solver.
- Sec. 2 delves into the our configuration of UED-Net to reproduce the experimental results presented in this paper.
- Sec. 3 contains additional comparisons with state-of-the-art (SOTA) methods.
- Sec. 4 discusses more extensive ablation studies, which include the effects of the number of stages (Sec. 4.1) and different cross-stage interactions (Sec. 4.2).

1. C-SALSA for UED-Net

1.1. Model Formulation

We design the architecture of UED-Net based on the C-SALSA solver, which effectively decouples mixed constraints to optimize the high-resolution multispectral (HRMS) image processing. To reiterate the description in the main paper, the recovery of $\mathbf{H}$ from $\mathbf{L}$ and $\mathbf{P}$ is modeled using the mathematical formula of the deep unfolding network in UED-Net as follows:

$$\begin{aligned}\mathbf{H} \in \arg\min_{\mathbf{H}} &||F_{Ge}(\mathbf{H}) - \mathbf{L}|| + |F_{Se}(\mathbf{H}, \mathbf{L})| \\ &+||F_{Ga}(\mathbf{H}) - \mathbf{P}|| + |F_{Sa}(\mathbf{H}, \mathbf{P})|,\end{aligned} \tag{1}$$

where $||*||$ term represents the global degradation-aware fidelity component constrained by the ℓ_2-norm, reflecting the HRMS image's global perception of spatial and spectral modalities, while, the $|*|$ term, constrained by the ℓ_1-norm, is designed to capture sparse multi-scale prior information. Together, these terms jointly capture comprehensive degradation patterns perceived from the LRMS and PAN images, whose specific forms of these terms are defined in Eq. (13) and Eq. (14) of the main paper.

Next, we solve the constrained problem based on the C-SALSA algorithm. First, by introducing auxiliary variables $\mathbf{V}_{Ge}, \mathbf{V}_{Se}, \mathbf{V}_{Ga}$, and $\mathbf{V}_{Sa}$, we split the mixed regularization terms in Eq. (1). This reformulation leads to a new constrained optimization problem:

$$\begin{aligned}\mathbf{H} \in &\arg\min_{\mathbf{H}} ||\mathbf{V}_{Ge}|| + |\mathbf{V}_{Se}| + ||\mathbf{V}_{Ga}|| + |\mathbf{V}_{Sa}|, \\ \text{s.t.}&\mathbf{V}_{Ge} = F_{Ge}(\mathbf{H}) - \mathbf{L}, \mathbf{V}_{Se} = F_{Se}(\mathbf{H}, \mathbf{L}), \\ &\mathbf{V}_{Ga} = F_{Ga}(\mathbf{H}) - \mathbf{P}, \mathbf{V}_{Se} = F_{Sa}(\mathbf{H}, \mathbf{P}).\end{aligned} \tag{2}$$

Subsequently, we apply the Augmented Lagrangian (AL) method to incorporate penalty terms into the Eq. (2), transforming it into the following equivalent iterative optimization problem:

$$\begin{aligned}\min_{(\mathbf{H}^{(k)}, \mathbf{V}_{Ge}^{(k)}, \mathbf{V}_{Ga}^{(k)}, \mathbf{V}_{Se}^{(k)}, \mathbf{V}_{Sa}^{(k)})} &|\mathbf{V}_{Se}^{(k)}| + |\mathbf{V}_{Sa}^{(k)}| + ||\mathbf{V}_{Ge}^{(k)}|| + ||\mathbf{V}_{Ga}^{(k)}|| \\ &+||F_{Ga}(\mathbf{H}^{(k)}) - \mathbf{P} - \mathbf{V}_{Ga}^{(k)} - \mathbf{W}_{Ga}^{(k)}|| + ||F_{Sa}(\mathbf{H}^{(k)}, \mathbf{P}) - \mathbf{V}_{Sa}^{(k)} - \mathbf{W}_{Sa}^{(k)}|| \\ &+||F_{Ge}(\mathbf{H}^{(k)}) - \mathbf{L} - \mathbf{V}_{Ge}^{(k)} - \mathbf{W}_{Ge}^{(k)}|| + ||F_{Se}(\mathbf{H}^{(k)}, \mathbf{L}) - \mathbf{V}_{Se}^{(k)} - \mathbf{W}_{Se}^{(k)}||,\end{aligned} \tag{3}$$

where $k \in \{0, 1, ..., N\}$ and N denotes the maximum number of iterations. $\mathbf{W}_{Ge}^{(k)}, \mathbf{W}_{Ga}^{(k)}, \mathbf{W}_{Se}^{(k)}$, and$\mathbf{W}_{Sa}^{(k)}$ are the Lagrange multipliers, which store the residuals between hierarchical iterative features and are updated in the k^{th} iteration through the following equation:

$$\begin{aligned}\mathbf{W}_{Ge}^{(k)} &= \mathbf{W}_{Ge}^{(k-1)} + \mathbf{f}_{Ge}^{(k)} - \mathbf{V}_{Ge}^{(k-1)}, \\ \mathbf{W}_{Ga}^{(k)} &= \mathbf{W}_{Ga}^{(k-1)} + \mathbf{f}_{Ga}^{(k)} - \mathbf{V}_{Ga}^{(k-1)}, \\ \mathbf{W}_{Se}^{(k)} &= \mathbf{W}_{Se}^{(k-1)} + \mathbf{f}_{Se}^{(k)} - \mathbf{V}_{Se}^{(k-1)}, \\ \mathbf{W}_{Sa}^{(k)} &= \mathbf{W}_{Sa}^{(k-1)} + \mathbf{f}_{Ge}^{(k)} - \mathbf{V}_{Sa}^{(k-1)},\end{aligned} \tag{4}$$

where $(\mathbf{f}_{Ge}^{(k)}, \mathbf{f}_{Se}^{(k)})/(\mathbf{f}_{Ga}^{(k)}, \mathbf{f}_{Sa}^{(k)})$ represent the feature representations of the spectral/spatial modalities of the global degradation-aware fidelity and sparse multi-scale prior at the k^{th} stage. Similarly, their specific formulations can be found in Eq. (13) and (14) of the main paper.

1.2. Model Solution

Direct solving Eq. (3) is challenging due to inseparable quadratic terms and non-smooth components. To address this, we decouple the mixed constraints, enabling us to minimize three sub-problems alternately, which include the fidelity terms $(\mathbf{V}_{Ge}^{(k)}, \mathbf{V}_{Ga}^{(k)})$, the sparse prior terms $(\mathbf{V}_{Se}^{(k)}, \mathbf{V}_{Sa}^{(k)})$, and their integration into $\mathbf{H}^{(k)}$.

Update of $(\mathbf{V}_{Se}^{(k)}, \mathbf{V}_{Sa}^{(k)})$: We separate the sparse multi-scale prior terms from Eq. (3), formulating it as a Lasso problem expressed as follows:

$$\begin{aligned}\mathbf{V}_{Se}^{(k)} &\in \arg\min_{\mathbf{V}_{Se}} |\mathbf{V}_{Se}| + ||\mathbf{V}_{Se} + \mathbf{f}_{Se}^{(k)} + \mathbf{W}_{Se}^{(k-1)}||, \\ \mathbf{V}_{Sa}^{(k)} &\in \arg\min_{\mathbf{V}_{Sa}} |\mathbf{V}_{Sa}| + ||\mathbf{V}_{Sa} + \mathbf{f}_{Sa}^{(k)} + \mathbf{W}_{Sa}^{(k-1)}||.\end{aligned} \tag{5}$$

To solve this problem effectively and promote sparsity in multi-scale features, we employ the soft-thresholding shrinkage [3] method:

$$\begin{aligned}\mathbf{V}_{Se}^{(k)} &= S_{\epsilon_e}(\mathbf{f}_{Se}^{(k)} + \mathbf{W}_{Se}^{(k-1)}), \\ \mathbf{V}_{Sa}^{(k)} &= S_{\epsilon_a}(\mathbf{f}_{Sa}^{(k)} + \mathbf{W}_{Sa}^{(k-1)}),\end{aligned} \tag{6}$$

where ϵ_e and ϵ_a are randomly initialized and stage-wise learnable parameters, which control the sparsity enforced

by shrinkage to reduce noise introduced by multi-scale sampling and enhance the representation of multi-scale details. For any ϵ, the $S_\epsilon(*)$ is defined as:

$$S_\epsilon(*) = \text{sgn}(*) \cdot \max(|*| - \epsilon, 0). \tag{7}$$

Update of $(\mathbf{V}_{Ge}^{(k)}, \mathbf{V}_{Ga}^{(k)})$: Similarly, we decouple the global degradation-aware data fidelity terms from Eq. (3), which are formulated with ℓ_2 constraints, encourage smooth degradation representation:

$$\begin{aligned} \mathbf{V}_{Ge}^{(k)} &\in \arg\min_{\mathbf{V}_{Ge}} ||\mathbf{V}_{Ge}|| + ||\mathbf{V}_{Ge} + \mathbf{f}_{Ge}^{(k)} + \mathbf{W}_{Ge}^{(k-1)}||, \\ \mathbf{V}_{Ga}^{(k)} &\in \arg\min_{\mathbf{V}_{Ga}} ||\mathbf{V}_{Ga}|| + ||\mathbf{V}_{Ga} + \mathbf{f}_{Ga}^{(k)} + \mathbf{W}_{Ga}^{(k-1)}||, \end{aligned} \tag{8}$$

The approximate solutions for $\mathbf{V}_{Ge}^{(k)}$ and $\mathbf{V}_{Ga}^{(k)}$ correspond to orthogonal projections onto an ℓ_2 ball of sufficiently small radius [1], expressed as:

$$\begin{aligned} \mathbf{V}_{Ge}^{(k)} &= \mathbf{V}_{Ge}^{(k-1)} + \frac{\mathbf{f}_{Ge}^{(k)} + \mathbf{W}_{Ge}^{(k-1)}}{||\mathbf{f}_{Ge}^{(k)} + \mathbf{W}_{Ge}^{(k-1)}||}, \\ \mathbf{V}_{Ga}^{(k)} &= \mathbf{V}_{Ga}^{(k-1)} + \frac{\mathbf{f}_{Ga}^{(k)} + \mathbf{W}_{Ga}^{(k-1)}}{||\mathbf{f}_{Ga}^{(k)} + \mathbf{W}_{Ga}^{(k-1)}||}. \end{aligned} \tag{9}$$

In the DUN context, we apply learnable normalization to enhance the generalization of this process:

$$\begin{aligned} \mathbf{V}_{Ge}^{(k)} &= \mathbf{V}_{Ge}^{(k-1)} + \text{GN}_1(F_{Ge}(\mathbf{H}^{(k-1)}, \mathbf{L}) + \mathbf{W}_{Ge}^{(k-1)}), \\ \mathbf{V}_{Ga}^{(k)} &= \mathbf{V}_{Ga}^{(k-1)} + \text{GN}_1(F_{Ga}(\mathbf{H}^{(k-1)}, \mathbf{L}) + \mathbf{W}_{Ga}^{(k-1)}), \end{aligned} \tag{10}$$

where GN_1 represents the normalization with the number of groups being 1.

Updata of $\mathbf{H}^{(k)}$: We regard updated auxiliary variables as constants and decouple the data terms about H from Eq. (3), establishing $\mathbf{H}^{(k)}$:

$$\begin{aligned} \mathbf{H}^{(k)} \in \arg\min_{\mathbf{H}} &||F_{Ga}(\mathbf{H}^{(k)}) - \mathbf{P} - \mathbf{V}_{Ga}^{(k)} - \mathbf{W}_{Ga}^{(k)}|| \\ &+ ||F_{Sa}(\mathbf{H}^{(k)}, \mathbf{P}) - \mathbf{V}_{Sa}^{(k)} - \mathbf{W}_{Sa}^{(k)}|| \\ &+ ||F_{Ge}(\mathbf{H}^{(k)}) - \mathbf{L} - \mathbf{V}_{Ge}^{(k)} - \mathbf{W}_{Ge}^{(k)}|| \\ &+ ||F_{Se}(\mathbf{H}^{(k)}, \mathbf{L}) - \mathbf{V}_{Se}^{(k)} - \mathbf{W}_{Se}^{(k)}||. \end{aligned} \tag{11}$$

The Eqs. (5) to (10) in UED-Net is outlined as encoding the degraded pattern, while Eq. (11) can be solved using a gradient descent, which is summarized as the following decoding.

We first further combine the spatially and spectrally degradation-aware auxiliary variables to obtain the feature representations of the degradation pattern at this stage, $\mathbf{f}_{spe}^{(k)}$ and $\mathbf{f}_{spa}^{(k)}$, as described in Eq. (19) and (4) of the main paper.

Next, we employ PGAM to calibrate the spatial offsets of $\mathbf{f}_{spe}^{(k)}$ and $\mathbf{f}_{spa}^{(k)}$, regulating the spatial/spectral distributions at this stage to obtain the $\mathbf{f}_{mm}^{(k)}$, as detailed in Eq. (5)-(8) of the main paper.

Furthermore, we utilize the customized UAAM to capture cross-stage feature interactions, mitigating noise accumulation across stages to obtain the $\mathbf{f}_s^{(k)}$, as described in Eq. (9).

Finally, we adaptively perceive the iteration step size and the gradient descent feature representation $\nabla\mathbf{H}^{(k-1)}$ using Eq. (10)-(12) of the main paper, and reconstruct the HRMS at the k^{th} stage using the following general gradient descent formulation:

$$\mathbf{H}^{(k)} = \mathbf{H}^{(k-1)} + \nabla\mathbf{H}^{(k-1)}. \tag{12}$$

2. More Model Reproducible Details

Table 1. Training parameters and model configuration.

Configurations	Default Settings
Base Learning Rate	5×10^{-4}
Min Learning Rate	5×10^{-8}
Optimizer	ADAM
Weight Decay	0
Optimizer Momentum	0.9,0.999
Batch Size	4
Training Epochs	<1000
Learning Rate Schedule	Cosineannealing
Number of Head (T)	4
Number of Stages (N)	7
Hidden Layer Dimensions (S)	16
Convolution Initialization	Kaiming
$\mu_w^{(s)}, s \in \{1, \ldots, S\}$	$8 \cdot (\frac{k}{S-1})^{1.35} - 5$
$\mu_u^{(s)}, s \in \{1, \ldots, S\}$	$log(0.3) + \frac{((k+1)\%3-1)}{2}$
$(\mathbf{V}_{ii}^{(0)}, \mathbf{W}_{ii}^{(0)}), ii \in \{Ge, Ga, Se, Sa\}$	Zero Matrix
Other Learnable Parameters	torch.randn
Implementation	PyTorch 2.5.1
CPU	Intel i5-10600KF
GPU	NVIDIA GeForce RTX 4090

We use UED-Net with 7 reconstruction stages ($N = 7$) and 16 hidden layers ($S = 16$) as the default model, which is derived from the ablation study on the number of stages in Sec. 4. UED-Net upscales the LRMS using bicubic interpolation to initialize $\mathbf{H}^{(0)}$. Additionally, the learnable auxiliary parameters in the UAAM are initialized as described in the RWKV [10]. The auxiliary iterative variables in the SSEM are initialized as zero matrices. Other learnable parameters are initialized with random values from a normal distribution within the range [0, 1]. We summarize the key training parameters and model configurations in Tab. 1 for a better understanding of our approach.

Based on this experimental configuration, the default UED-Net requires approximately 35 milliseconds for inference on a multispectral image with 4 bands and a spatial size of 128×128.

The supplementary materials include source code and detailed experimental settings required to replicate the findings outlined in this paper. Additionally, the source code will be made publicly available to enhance accessibility and promote reproducibility.

Table 2. Comparison of UED-Net with other methods in simulated tests on reduced-resolution data, with the best result highlighted in bold **red** and the second best result highlighted in blue.

Dataset		WorldView-II					WorldView-III					GaoFen-2					Flops (G)	Params (M)
Metrics		ERGAS↓	SSIM↑	PSNR↑	SCC↑	SAM↓	ERGAS↓	SSIM↑	PSNR↑	SCC↑	SAM↓	ERGAS↓	SSIM↑	PSNR↑	SCC↑	SAM↓		
GSA	(TGRS'07)	1.6064	0.9266	36.9802	0.4552	0.0373	9.1864	0.5335	21.8331	0.6187	0.1310	1.7981	0.8855	36.5014	0.1284	0.0345	-	-
SFIM	(IJRS'07)	1.9651	0.9147	35.8461	0.4534	0.0397	8.7638	0.5483	22.1521	0.6697	0.1243	1.5923	0.8964	37.6654	0.1697	0.0292	-	-
Wavelet	(IGARSS'01)	2.0528	0.8765	35.1764	0.3641	0.0484	9.4045	0.5002	21.6659	0.0493	0.1371	2.0188	0.8189	35.8537	0.0243	0.0280	-	-
SHIP++	(TPAMI'24)	0.9035	0.9727	42.3276	0.5893	0.0212	3.0152	0.9265	30.8038	0.8309	0.0725	0.4872	0.9900	48.3720	0.5616	0.0094	2.9012	0.1783
SFINet++	(TPAMI'24)	0.9538	0.9675	41.5874	0.5147	0.0236	3.0217	0.9261	30.7665	0.8333	0.0720	0.4376	0.9906	49.3578	0.5897	0.0087	0.7742	0.0487
HFINet	(CVPR'24)	0.9906	0.9694	41.9026	0.5697	0.0226	3.1523	0.9192	30.4808	0.8156	0.0766	0.4907	0.9891	48.3114	0.5583	0.0090	1.0104	0.0773
WINet	(TGRS'24)	0.9024	0.9714	42.0479	0.5753	0.0226	3.1753	0.9187	30.3268	0.8205	0.0802	0.4472	0.9904	49.1070	0.5630	0.0090	1.9597	0.3336
CANNet	(CVPR'24)	0.9094	0.9721	41.9689	0.5800	0.0222	3.0980	0.9225	30.5497	0.8233	0.0768	0.4893	0.9902	48.3435	0.5623	0.0094	-	-
LFormer	(MM'24)	0.8510	0.9730	42.5194	0.5954	0.0226	3.0694	0.9255	30.9709	0.8282	0.0735	0.4551	0.9898	48.9494	0.5619	0.0094	7.8588	0.4494
RFCONet	(TGRS'24)	1.0992	0.9663	40.3569	0.5285	0.0233	3.3407	0.9153	29.9527	0.8134	0.0799	0.7378	0.9839	44.4392	0.4492	0.0102	4.0509	5.2089
PDDNet	(ICCV'23)	0.9889	0.9702	41.3745	0.5819	0.0240	3.3839	0.9088	29.8436	0.8095	0.0852	0.5098	0.9892	48.9465	0.5440	0.0102	0.1284	0.0395
INNF	(AAAI'22)	0.9176	0.9706	41.8949	0.5756	0.0222	3.1000	0.9216	30.5419	0.8250	0.0745	0.4831	0.9894	48.5199	0.5664	0.0095	1.2201	0.0613
DISPNet	(AAAI'24)	0.8759	0.9720	42.2527	0.5837	0.0215	3.0096	0.9267	30.8352	0.8315	0.0720	0.4493	0.9904	49.0900	0.5679	0.0097	27.667	1.5669
NLUNet	(TGRS'23)	1.0034	0.9644	41.0635	0.5413	0.0246	3.2981	0.9140	29.9715	0.8162	0.0811	0.4976	0.9885	48.2287	0.5458	0.0099	4.6099	0.3062
LGTEUN	(IJCAI'23)	0.8968	0.9734	42.6766	0.5832	0.0211	3.0151	0.9246	30.7884	0.8291	0.0723	0.4798	0.9894	48.5291	0.5728	0.0097	3.2113	0.3004
MMNet	(ECCV'22)	0.9600	0.9710	41.5033	0.5703	0.0220	3.2175	0.9171	30.3537	0.8217	0.0778	0.6377	0.9859	45.4227	0.4801	0.0160	4.5953	0.0703
MDCUN	(CVPR'22)	1.0060	0.9635	41.1297	0.5274	0.0249	3.3531	0.9111	29.8336	0.8142	0.0828	0.4830	0.9890	48.4141	0.5461	0.0098	118.30	0.1538
GPPNN	(CVPR'21)	0.9248	0.9702	41.8381	0.5807	0.0222	3.0923	0.9226	30.5770	0.8290	0.0738	0.4812	0.9893	48.4958	0.5596	0.0096	4.1901	0.3594
UED-Net	**Defult (7stg)**	0.8349	0.9739	42.7251	0.5977	0.0203	2.8918	0.9290	31.1564	0.8370	0.0675	0.4268	0.9910	49.5648	0.5986	0.0077	2.5097	0.1682
UED-Net	**Ours (9stg)**	0.8349	0.9740	42.7211	0.5914	0.0202	2.9021	0.9292	31.1272	0.8363	0.0686	0.4282	0.9910	49.5254	0.5991	0.0082	3.2268	0.2163
UED-Net	**Ours (11stg)**	0.8350	0.9740	42.7172	0.5960	0.0203	2.8987	0.9292	31.1380	0.8365	0.0682	0.4277	0.9910	49.5032	0.5994	0.0085	3.9439	0.2643

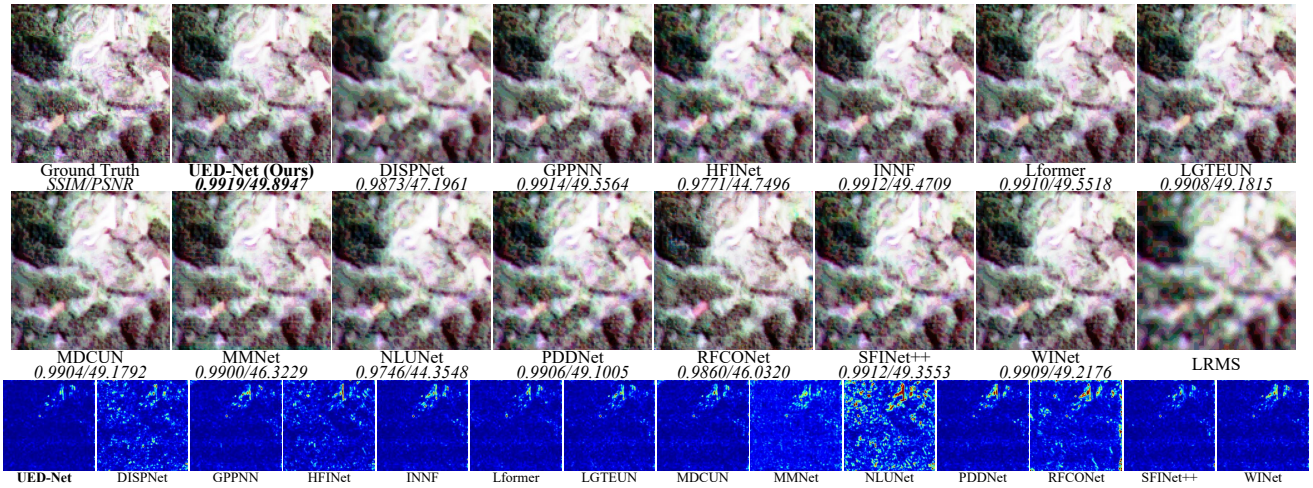

Figure 1. Visual comparison of UED-Net with other methods in simulated tests on GaoFen-2.

3. More Comparisons with SOTA Methods

In the main body of this paper, we compare the proposed UED-Net with various representative pansharpening methods, which are configured according to the best settings reported in their respective papers, including (SHIPNet++ [20], CANNet [4], SFINet++ [19], HFINet [12], WINet [17], PDDNet [5], INNF [18], LFormer [6], and RFCONet [11]), (MMNet [15], DISPNet [13], NLUNet [8], LGTEUN [8], MDCUN [16], and GPPNN [14]). In this subsection, we further highlight the significant advantages of UED-Net over these methods and provide a comparison with traditional methods (GSA [2], SFIM [9], Wavelet [7]). Additionally, we present a performance comparison of UED-Net under different representative configurations. In Tab. 2, we present the results of the further simulation tests, with the best, and second-best performances marked in red, and blue, respectively. The various configurations of UED-Net show significant performance improvements across three datasets. Specifically, on the GaoFen-2 dataset, UED-Net achieves a 0.2048 dB improvement in PSNR over the second-best algorithm, SFINet++; on the WorldView-II dataset, it outperforms the second-best algorithm, LGTEUN, by 0.0486 dB in PSNR; and on the WorldView-III dataset, UED-Net shows a 1.1849 dB improvement in PSNR compared to the second-best algorithm, DISPNet. These results indicate that our proposed method is not dataset-dependent and demonstrates reliable scalability generalization capability.

Furthermore, we present supplementary results on real test in Tab. 3, where our default UED-Net configuration

Table 3. Comparison of UED-Net with other methods in real tests on GaoFen-2 full-resolution data.

Metrics	Traditional Methods			Pure DL-Based Methods									Deep Unfolding Methods						
	GSA	SFIM	Wavelet	SHIP++	SFINet++	HFINet	WINet	CANNet	LFormer	RFCONet	PDDNet	INNF	DISPNet	NLUNet	LGTEUN	MMNet	MDCUN	GPPNN	UED-Net
$D_\lambda \downarrow$	0.17344	0.16949	0.33271	0.07208	0.07836	0.09837	0.11871	0.07476	0.08352	0.25559	0.07984	0.07080	0.08068	0.07628	0.11336	0.07182	0.07649	0.09124	**0.06801**
$D_s \downarrow$	0.44336	0.30266	0.40957	0.09838	0.08855	0.08491	0.08273	0.09752	0.08593	0.13758	0.10634	0.10758	0.08224	0.10741	0.17165	0.17053	0.08860	0.09485	**0.07011**
$QNR \uparrow$	0.46010	0.57915	0.39399	0.84930	0.84003	0.82506	0.80838	0.83502	0.83773	0.64309	0.82231	0.82923	0.84371	0.82450	0.73445	0.76990	0.84168	0.82257	**0.86665**

Table 4. Ablation of the stage number.

Num of Stage	3	5	**7 (Defult)**	9	11	13
SSIM↑	0.9906	0.9907	0.9910	0.9910	0.9910	0.9910
PSNR↑	49.3232	49.3721	49.5648	49.5254	49.5032	49.5121
$QNR \uparrow$	0.7756	0.8083	0.8666	0.8645	0.8639	0.8652
FLOPs	1.0756	1.7927	2.5097	3.2268	3.9439	4.6609
Params	0.0721	0.1202	0.1682	0.2163	0.2643	0.3124

Table 5. Ablation of cross-stage feature interactions.

Method	Simulated Tests (GaoFen-2)			Simulated Tests (WorldView-II)			Real Tests (GaoFen-2)			Calculate Costs	
	ERGAS↓	SSIM↑	PSNR↑	ERGAS↓	SSIM↑	PSNR↑	$D_\lambda \downarrow$	$D_s \downarrow$	$QNR \uparrow$	FOLPs	Params
Net1	0.4616	0.9896	48.8597	0.8825	0.9718	42.2939	0.0680	0.0768	0.8604	1.5051	0.1065
Net2	0.4462	0.9901	49.1483	0.8623	0.9721	42.4275	0.9791	0.0100	0.8359	6.7954	0.4294
Net3	0.4443	0.9902	49.2000	0.8593	0.9723	42.4554	0.0797	0.1522	0.7802	3.6190	0.2364
Defult	**0.4268**	**0.9910**	**49.5648**	**0.8349**	**0.9739**	**42.7251**	**0.0680**	**0.0701**	**0.8666**	2.5097	0.1682

continues to exhibit superior performance, particularly in terms of QNR metrics and maintaining a balanced hardware load. Additionally, compared to DUN-based methods, our method incurs the lowest computational cost.

Finally, we include a visual comparison on the GaoFen-2 dataset for simulated tests,as shown in Fig. 1, highlighting both the reconstruction details and visualizations of the mean squared error (MSE). Notably, the comparisons further confirm that our method outperforms other algorithms across multiple scenarios, while offering a competitive performance-to-computation-cost ratio comparable to pure deep learning methods. This validates our hypothesis on the importance of effective cross-modal and cross-stage interactions at different abstraction levels in successful pansharpening.

4. More Ablation Studies

4.1. Number of Stages

We conduct an ablation study on the GaoFen-2 dataset to investigate how the performance of UED-Net varies with computational cost. As shown in Fig. 1 of main papet and Tab. 4, performance improves as we increase the number of stages. We observe that both performance and cost increase significantly with the number of stages. At 7 iterations, we achieve an impressive PSNR of 49.5648 dB in simulated tests and the highest QNR in real tests. After this point, the performance continues to improve slightly in simulated tests but shows some fluctuation. Additionally, we present the results for 9 and 11 stages in WorldView-II and WorldView-III tests, as shown in Tab. 2. Based on the performance-cost trade-off, we use 7 iterations stages as the default configuration for UED-Net.

4.2. Cross-stage feature interactions

Building on the ablation study of the UAAM presented in Sec. 4.2 of main paper, we further investigate the benefits of associative attention as a cross-stage interaction mechanism within our proposed UAAM by substituting it with various cross-stage interaction methods. Specifically, Net1 serves as a baseline network without any cross-stage interaction. For Net2, we implement the method proposed in MDCUN [16], which utilizes stacked intermediate variables. For Net3, we employ the LSTM-like stage interaction method outlined in MMNet [15]. As demonstrated in Tab. 5, our method outperforms Net1, Net2, and Net3 across all evaluation metrics, not only in the simulated tests of GaoFen-2 and WorldView-II but also in real tests, while maintaining superior computational efficiency.

References

[1] Manya V. Afonso, José M. Bioucas-Dias, and Mário A. T. Figueiredo. A fast algorithm for the constrained formulation of compressive image reconstruction and other linear inverse problems. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4034–4037, 2010. 2

[2] Bruno Aiazzi, Stefano Baronti, and Massimo Selva. Improving component substitution pansharpening through multivariate regression of ms +pan data. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10):3230–3239, 2007. 3

[3] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. 1

[4] 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)*, pages 27738–27747, 2024. 3

[5] Xuanhua He, Keyu Yan, Rui Li, Chengjun Xie, Jie Zhang, and Man Zhou. Pyramid dual domain injection network for pan-sharpening. In *2023 IEEE/CVF International Confer-*

ence on Computer Vision (ICCV), pages 12862–12871, 2023. 3

[6] 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*, pages 1486–1494, New York, NY, USA, 2024. Association for Computing Machinery. 3

[7] R.L. King and Jianwen Wang. A wavelet based algorithm for pan sharpening landsat 7 imagery. In *IGARSS 2001. Scanning the Present and Resolving the Future. Proceedings. IEEE 2001 International Geoscience and Remote Sensing Symposium (Cat. No.01CH37217)*, pages 849–851 vol.2, 2001. 3

[8] Xingxing Li, Yujia Li, Guangyao Shi, Liping Zhang, Weisheng Li, and Dajiang Lei. Pansharpening method based on deep nonlocal unfolding. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–11, 2023. 3

[9] J. G. Liu. Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details. *International Journal of Remote Sensing*, 21(18): 3461–3472, 2000. 3

[10] Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, Eric Alcaide, Stella Biderman, Eugene Cheah, Xingjian Du, Teddy Ferdinan, Haowen Hou, Przemysław Kazienko, Kranthi Kiran GV, Jan Kocoń, Bartłomiej Koptyra, Satyapriya Krishna, Ronald McClelland Jr. au2, Jiaju Lin, Niklas Muennighoff, Fares Obeid, Atsushi Saito, Guangyu Song, Haoqin Tu, Cahya Wirawan, Stanisław Woźniak, Ruichong Zhang, Bingchen Zhao, Qihang Zhao, Peng Zhou, Jian Zhu, and Rui-Jie Zhu. Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence, 2024. 2

[11] Jiahui Qu, Xuyao Liu, Wenqian Dong, Yang Liu, Tongzhen Zhang, Yang Xu, and Yunsong Li. Progressive multi-iteration registration-fusion co-optimization network for unregistered hyperspectral image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. 3

[12] 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 (CVPR)*, pages 25922–25931, 2024. 3

[13] Hebaixu Wang, Meiqi Gong, Xiaoguang Mei, Hao Zhang, and Jiayi Ma. Deep unfolded network with intrinsic supervision for pan-sharpening. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5419–5426, 2024. 3

[14] Shuang Xu, Jiangshe Zhang, Zixiang Zhao, Kai Sun, Junmin Liu, and Chunxia Zhang. Deep gradient projection networks for pan-sharpening. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1366–1375, 2021. 3

[15] Keyu Yan, Man Zhou, Li Zhang, and Chengjun Xie. Memory-augmented model-driven network for pansharpening. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIX*, pages 306–322. Springer, 2022. 3, 4

[16] Gang Yang, Man Zhou, Keyu Yan, Aiping Liu, Xueyang Fu, and Fan Wang. Memory-augmented deep conditional unfolding network for pansharpening. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1778–1787, 2022. 3, 4

[17] Jie Zhang, Xuanhua He, Ke Ren Yan, Ke Cao, Rui Li, Chengjun Xie, Man Zhou, and Danfeng Hong. Pansharpening with wavelet-enhanced high-frequency information. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. 3

[18] Man Zhou, Jie Huang, Yanchi Fang, Xueyang Fu, and Aiping Liu. Pan-sharpening with customized transformer and invertible neural network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):3553–3561, 2022. 3

[19] Man Zhou, Jie Huang, Keyu Yan, Danfeng Hong, Xiuping Jia, Jocelyn Chanussot, and Chongyi Li. A general spatial-frequency learning framework for multimodal image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2024. 3

[20] 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*, pages 1–18, 2024. 3