Deep Adaptive Unfolded Network via Spatial Morphology Stripping and Spectral Filtration for Pan-sharpening

Hebaixu Wang Jiayi Ma†
School of Electronic Information, Wuhan University, Wuhan 430072, China
Zhongguancun Academy, Beijing 100094, China
{wanghebaixu,jyma2010}@gmail.com

Abstract

In the field of pan-sharpening, existing deep methods are hindered in deepening cross-modal complementarity in the intermediate feature, and lack effective strategies to harness the network entirety for optimal solutions, exhibiting limited feasibility and interpretability due to their black-box designs. Besides, validating pan-sharpening performance in high-level semantic tasks is intractable for the absence of datasets. To tackle these issues, we propose a deep adaptive unfolded network via spatial morphology stripping and spectral filtration for pan-sharpening, which is conceptualized as a linear inverse problem regularized by spatial and spectral priors. Specifically, we incorporate phase-oriented constraints into the spatial prior to facilitate the extraction of modal-invariant spatial morphology by intrinsic decomposition and leverage a physics-driven spectral filtration attention mechanism aligned with the spectral prior to mine the inherent spectral correlation. After transparently unfolding the model into a multi-stage network, an adaptive stage-exiting mechanism is designed to capitalize on fusion diversity by aggregating optimal image patches across candidate stages. To systematically complete the assessment, we construct the first panoptic segmentation dataset as a semantic-level benchmark for pan-sharpening performance validation. Extensive experiments are conducted to verify the merits of our method with state-of-the-arts. Code is available at https://github.com/Baixuzx7/DAPNet.

1. Introduction

Satellite systems are constrained from simultaneously capturing high-resolution (HR) multi-spectral (MS) images due to hardware limitations. In this context, the acquired data typically encompass low-resolution (LR) MS images and paired HR panchromatic (PAN) images. The former offers rich spectral attributes across multiple bands but with low spatial resolution, while the latter possesses high spatial resolution, albeit with limited spectral richness. Consequently, their inherent complementarity provides a feasible foundation for pan-sharpening [38], which super-resolves LRMS images with paired HRPAN images as texture-rich guidance. Given the excellent properties of HRMS images, the pan-sharpening technology holds significant importance across diverse domains [45, 53, 64].

So far, pan-sharpening has achieved remarkable progress owing to the powerful nonlinear capabilities of deep learning compared to the conventional methods [9, 37, 57, 60].

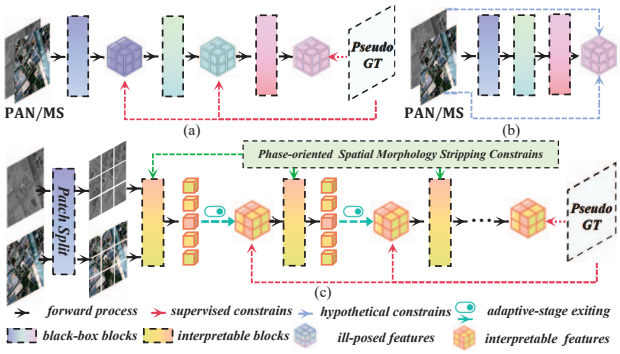


Figure 1. Overview of existing pan-sharpening paradigms: (a) Supervised learning, (b) unsupervised learning, and (c) ours.

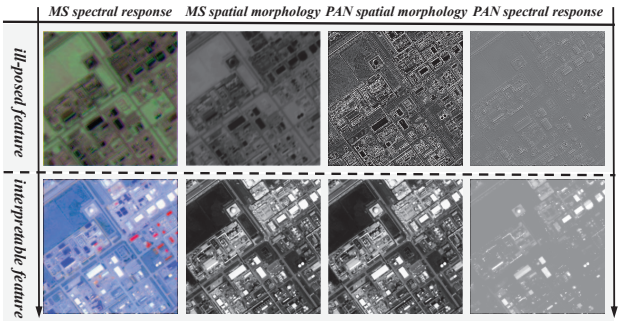
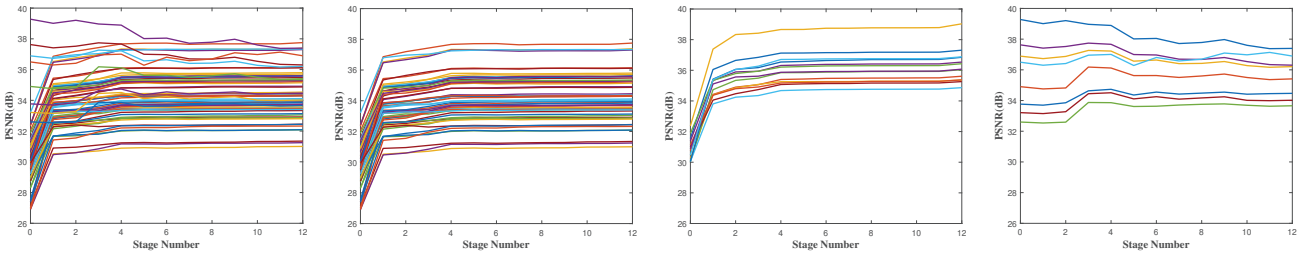


Figure 2. Examples of the internal feature of the PAN and MS images. Upper tiers are unconstrained. Lower tiers are constrained.

†Corresponding author: Jiayi Ma.



(a) All patches (b) Bottleneck patches (c) Growing patches (d) Overfitting patches

Figure 3. Stage-wise PSNR performances of different image pair patches super-resolved at 4× scaling using our 12-stage networks, where bottleneck patches, growing patches, and overfitting patches exhibit stable, improved, and deteriorating trends, respectively.

Initially, a simple structure with three convolution layers pioneered the first deep pan-sharpening paradigm [27]. Subsequently, numerous deeper and more sophisticated architectures have been investigated to further improve the fusion performance [11, 13, 15, 23, 32, 63]. Existing deep pan-sharpening methods can be generally categorized into supervised and unsupervised paradigms, as shown in Fig. 1. Supervised paradigms [32, 51, 52, 59] fabricate pseudo ground-truth (GT) to constrain the outputs of intermediate or final blocks abiding by Wald's protocol [40]. Consequently, the internal features in the intermediate layers remain ill-posed [28, 41, 43] and overfitting phenomena may occur, as illustrated in Fig. 2 and Fig. 3, respectively. Unsupervised learning discards the fixed protocols and simplifies the spectral observation priors [22, 25] to establish the hypothetical constraints, thereby improving spectral accuracy and spatial details. However, these over-simplified physical constraints struggle to maintain a balanced performance, disrupting the integration of complementary information. Besides, the existing assessments of pan-sharpening performance are inadequately comprehensive due to the absence of benchmark datasets in practical applications and excessive reliance on artificially manipulated parameters [4, 6], yielding feasibility but imprecision.

In summary, the limitations of existing pan-sharpening methods can be delineated into pivotal aspects. First, the lack of constraints on the internal features in the intermediate layers of neural networks hinders the improvement of their performance and interpretability [44, 46]. Second, the immoderate pursuit of capacity and depth in achieving multi-stage networks is not mandatory for performance improvement [17]. Specifically, not all image patches attain optimal efficacy at the terminal stage, as depicted in Fig. 3. Such architectures inherently hold underexplored fusion diversity across different stages, ultimately yielding suboptimal solutions. Last, validating pan-sharpening performance in practical applications of real-world scenarios is intractable due to the absence of datasets.

To tackle these issues, we propose a deep adaptive unfolded network via spatial morphology stripping and spectral filtration for pan-sharpening. We conceptualize the pan-sharpening as a special case of linear inverse problems with ill-posedness to retrieve the desired HRMS image from the lossy measurement LRMS image. Firstly, the inherent spectroscopic [20] mechanism in satellite sensors converts energy within the identical scene into distinct spectral bands with varying resolutions to acquire image data. Consequently, we extract the morphological consistency from the imaging scene, utilizing it as the spatial prior to describe the common spatial properties shared between PAN and MS modalities by incorporating the phase-oriented constraints. Secondly, the energy captured within the broad PAN band undergoes filtration, ultimately transforming into MS images with energy distributed across narrow spectral bands [5]. Based on this observation, we design a physics-driven spectral filtration attention mechanism with the inherent spectral correlations and energy distribution patterns serving as spectral priors. Thirdly, a novel effective iteration algorithm is devised to transparently and interpretably unfold the model into a multi-stage architecture. To fully leverage the fusion diversity across different stages and ensure optimal fusion outcomes, a novel adaptive-stage exiting mechanism is tailored to selectively integrate the optimal image patches from candidate stages into precise solutions. Lastly, a panoptic segmentation dataset (SegGF) is constructed, serving as a semantic-level benchmark for pan-sharpening, as illustrated in Fig. 4. It contains five representative land categories featuring over 100,000 polygons for segmentation, which are annotated at both simulated and practical scales for the typical execution of pan-sharpening. Our contributions are summarized below:

- We formulate pan-sharpening as a linear inverse problem with spatial morphology and spectral correlation priors, and devise multi-stage neural solvers with an adaptive stage-exiting mechanism, customized to generate optimal solutions and boost network efficiency.
- Phase-oriented constraints are integrated into spatial priors to facilitate the extraction of spatial morphology and spectral response, and physics-driven spectral filtration attention mechanisms are designed to mine the spectral

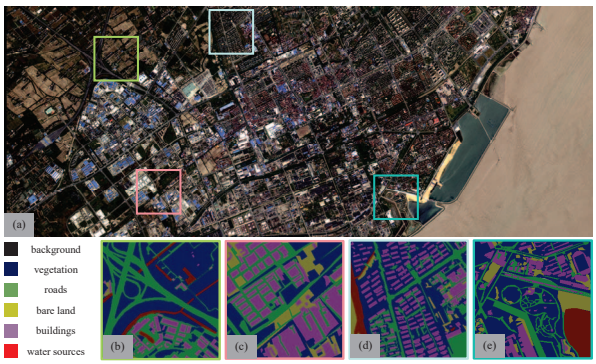


Figure 4. Examples of the MS images and annotations from the proposed SegGF datasets (displayed only on RGB bands).

correlations for spectral priors, enhancing interpretability and complementarity of the internal features.

- We construct the first semantic-level benchmark dataset for pan-sharpening assessments, which covers five typical land categories and includes over 100,000 annotated polygons at two distinct scales.

2. Method

2.1. Problem Formulation

Pan-sharpening can be viewed as a special case of linear inverse problems for multi-spectral imagery, where we have a partial LRMS measurement $y \in \mathbb{R}^{h\times w}$ that is derived from the desired HRMS image $x \in \mathbb{R}^{H\times W}$. Formally, the general form of the forward model can be stated as:

$$y = \mathcal{A}x + n, \tag{1}$$

where $\mathcal{A}$ is the forward measurement operator and n is the measurement noise. Since the mapping $x \to y$ is many-to-one, we require additional regularization to narrow down the solution space. Motivated by the observations that two modalities of paired PAN image p and MS image x essentially possess identically spatial morphology and inherent spectral correlations, we formulate them as regularization terms weighted by ω_1 and ω_2 to rewrite Eq. (1):

$$\arg\min_x \frac{1}{2}\|y-\mathcal{A}x\|_2^2+\frac{\omega_1}{2}\|\Gamma x-\Gamma p\|_2^2+\frac{\omega_2}{2}\|\phi p-\mathcal{A}x\|_2^2, \tag{2}$$

where Γ is the neural operator for common feature extraction, and ϕ distributes the energy in a wide spectrum band into multiple narrow spectral bands.

2.2. Customized Variation Solver

To circumvent the intractability of Eq. (2), we introduce two auxiliary variables z and r to simplify the formula [33, 62]:

$$\begin{aligned}\arg\min_{x,z,r}\frac{1}{2}\|y-\mathcal{A}x\|_2^2+\frac{\sigma_1}{2}\|z-x\|_2^2+\frac{\omega_1}{2}\|\Gamma z-\Gamma p\|_2^2\\ +\frac{\sigma_2}{2}\|r-x\|_2^2+\frac{\omega_2}{2}\|\phi p-\mathcal{A}r\|_2^2,\end{aligned} \tag{3}$$

where σ_1, σ_2 are penalty proportions to enforce consistency between the original and auxiliary variables. Eq. (3) can be equivalent to the following sub-problems:

$$z^{(k+1)} = \arg\min_z \frac{\sigma_1}{2}\|z-x^{(k)}\|_2^2+\frac{\omega_1}{2}\|\Gamma z-\Gamma p\|_2^2, \tag{4}$$

$$r^{(k+1)} = \arg\min_r \frac{\sigma_2}{2}\|r-x^{(k)}\|_2^2+\frac{\omega_2}{2}\|\phi p-\mathcal{A}r\|_2^2, \tag{5}$$

$$\begin{aligned}x^{(k+1)} = \arg\min_x \frac{1}{2}\|y-\mathcal{A}x\|_2^2+\frac{\sigma_1}{2}\|z^{(k+1)}-x\|_2^2\\ +\frac{\sigma_2}{2}\|r^{(k+1)}-x\|_2^2,\end{aligned} \tag{6}$$

where the subscript k denotes the iteration step. We unroll inference by optimizing three sub-problems alternatively with proximal gradient method [26].

Updating z. Given the estimated HRMS image $x^{(k)}$ at iteration k, the auxiliary variable z can be updated:

$$z^{(k+1)} = prox_{\Omega_z}(z^{(k)}-\delta_z\nabla f_z(z^{(k)})), \tag{7}$$

$$\nabla f_z(z^{(k)}) = \sigma_1(z^{(k)}-x^{(k)})+\omega_1\Gamma^*(\Gamma z^{(k)}-\Gamma p), \tag{8}$$

where $prox_{\Omega_z}(\cdot)$ is proximal operator associated with variable z, and δ_z represents the z-optimization step size. $(\cdot)^*$ indicates the derivative of corresponding operators.

Updating r. Similarly, the recursive update formula of the variable r is given by:

$$r^{(k+1)} = prox_{\Omega_r}\left(r^{(k)}-\delta_r\nabla f_r(r^{(k)})\right), \tag{9}$$

$$\nabla f_r(r^{(k)}) = \sigma_1(r^{(k)}-x^{(k)})+\omega_2\mathcal{A}^*(\mathcal{A}r^{(k)}-\phi p), \tag{10}$$

where $prox_{\Omega_r}(\cdot)$ is proximal operator corresponding to variable r, and δ_r denotes the r-optimization step size.

Updating x. After deriving the optimized variables $z^{(k+1)}$ and $r^{(k+1)}$, the update of x can be obtained accordingly, yielding the below formula:

$$x^{(k+1)} = x^{(k)}-\delta_x\nabla f_x(x^{(k)}), \tag{11}$$

$$\begin{aligned}\nabla f_x(x^{(k)}) = \mathcal{A}^*(\mathcal{A}x^{(k)}-y)+\sigma_1(x^{(k)}-z^{(k+1)})\\ +\sigma_2(x^{(k)}-r^{(k+1)}),\end{aligned} \tag{12}$$

where δ_x is the x-optimization step size.

2.3. Network Architecture

Our network architecture is presented in Fig. 5. It comprises K stages with each including three sub-networks and sharing an adaptive stage-exiting mechanism (ASE) module. We utilize four different blocks to simulate these operators, namely convolutional up-sampling blocks (CUB), convolutional layer blocks (CLB), convolutional down-sampling blocks (CDB), and spectral filtration attention (SFA).

Adaptive Stage Exiting. To alleviate the overfitting phenomenon and fully exploit fusion diversity across different stages, we develop the ASE to dynamically determine

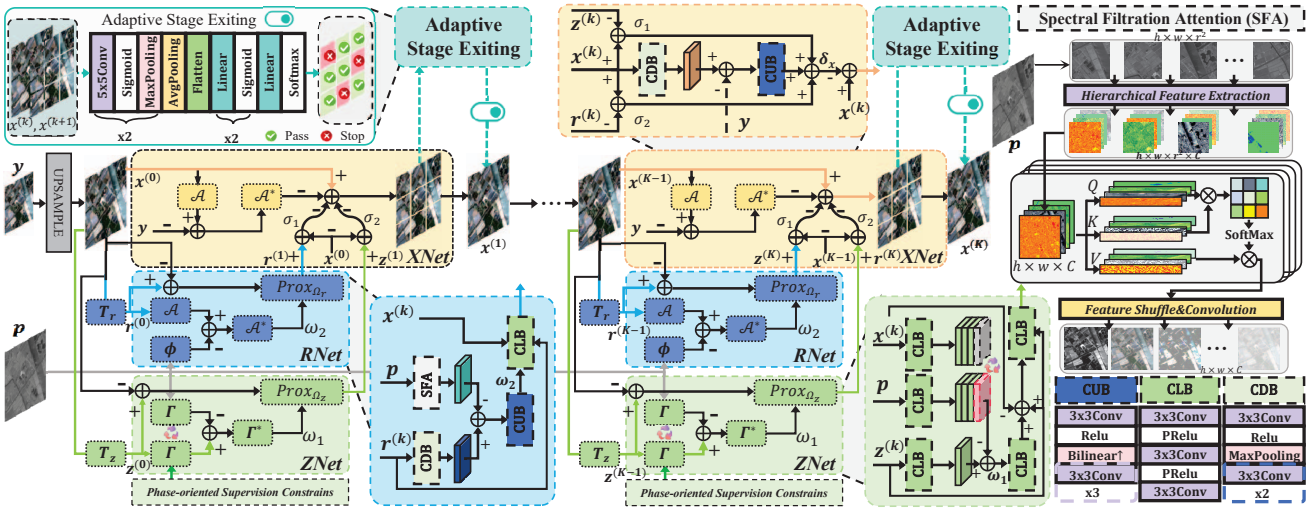


Figure 5. The architecture of the proposed methods includes an adaptive stage-exiting mechanism module for patch-wise discrimination and a pan-sharpening network that comprises three unfolded sub-networks, namely XNet, RNet, and ZNet.

whether image patches proceed to the next stage or are retained at the current stage, ensuring optimal outcomes and maintaining convergence. It adopts images from adjacent stages as input, generating corresponding confidence scores and binary labels. Afterwards, patches with positive labels are retained to filter out the optimal solution stage by stage.

Spectral Filtration Attention. Observation of the energy distribution across the spectrum reveals the spectral coherence between PAN and MS modalities. We utilize the SFA to simulate ϕ operator. Initially, PAN that carries the full-band energy is divided into patches for hierarchical feature extraction with convolution layers. Subsequently, these features are shuffled and rearranged into groups to form multiple heads. Then query (Q), key (K), and value (V) are generated by linear transformation to describe the energy distribution patterns. Spectral correlations are simulated by the self-attention [12] between these components. The outputs are obtained through feature shuffle and convolution.

2.4. Loss Functions

To stabilize the network performance, we employ ℓ_1 loss at each network stage to minimize the distance between estimated images and reference images in the spatial domain:

$$\mathcal{L}_{ref} = \sum_{k=1}^{K} |x^{(k)} - y|_1, \tag{13}$$

where $x^{(k)}$ is the output of the k-th stage. Considering data slice $x(\lambda)$ obtained in narrow spectrum can be represented as the multiplication of spectral response $R(\lambda)$ and spatial morphology $S(\lambda)$, we attain the dichromatic constraints to internal features to enforce the separability of spatial and spectral attributes with identical operators Γ:

$$\Gamma(x^{(k)}), \Gamma(\widetilde{p}) = (S_{x^{(k)}}, R_{x^{(k)}}), (S_{\widetilde{p}}, R_{\widetilde{p}}), \tag{14}$$

where $\widetilde{p}$ is p stacked in the spectral dimension. Spatial morphology depicts consistent texture and geometric shapes within an identical scene, and spectral response varies with wavelength λ. We utilize the exponential weight of $\nabla S(\lambda)$ to manipulate the smoothness of corresponding $\nabla R(\lambda)$:

$$\mathcal{L}_{smooth} = \|\nabla R_i \cdot e^{(-\lambda \nabla S_i)}\|, \ i = x^{(k)}, \widetilde{p}, \tag{15}$$

where ∇ is the gradient operator. As spatial disturbances are more prominent in the phase perspective for $\mathcal{F}(S(i+\Delta i, j+\Delta j) - S(i,j)) = S(u,v)[e^{j2\pi(\frac{\Delta i}{M}u + \frac{\Delta j}{N}v)} - 1]$, we use $\Psi(\cdot)$ to extract the phase and employ the below constraints to promote spatial morphology across bands:

$$\mathcal{L}_{equal} = \|S_{x^{(k)}} - S_{\widetilde{p}}\|_1, \ \mathcal{L}_{phase} = \|\Psi(S_{x^{(k)}}) - \Psi(S_{\widetilde{p}})\|_1, \tag{16}$$

where $\mathcal{L}_{equal}$ and $\mathcal{L}_{phase}$ define the spatial and phase consistency constraints, respectively. Furthermore, we adopt the reconstruction loss $\mathcal{L}_{rec}$ to enhance the correlation among response, morphology, and captured data, formulated as:

$$\begin{aligned}\mathcal{L}_{rec} = &\|R_{x^{(k)}} \circ S_{\widetilde{p}} - x^{(k)}\|_1 + \|R_{\widetilde{p}} \circ S_{x^{(k)}} - \widetilde{p}\|_1 \\ &+ \|R_{x^{(k)}} \circ S_{x^{(k)}} - x^{(k)}\|_1 + \|R_{\widetilde{p}} \circ S_{\widetilde{p}} - \widetilde{p}\|_1,\end{aligned} \tag{17}$$

where $\circ$ denotes the element-wise product. In conclusion, the phase-oriented supervision $\mathcal{L}_{morph}$ for intermediate features can be formulated as:

$$\mathcal{L}_{morph} = \sum_{k=1}^{K} \alpha \mathcal{L}_{smooth}^{(k)} + \beta \mathcal{L}_{equal}^{(k)} + \gamma \mathcal{L}_{phase}^{(k)} + \kappa \mathcal{L}_{rec}^{(k)}, \tag{18}$$

where α, β, γ and κ are weight parameters. The total loss function for pan-sharpening network is denoted as:

$$\mathcal{L}_{ps} = \mathcal{L}_{ref} + \rho \mathcal{L}_{morph}, \tag{19}$$

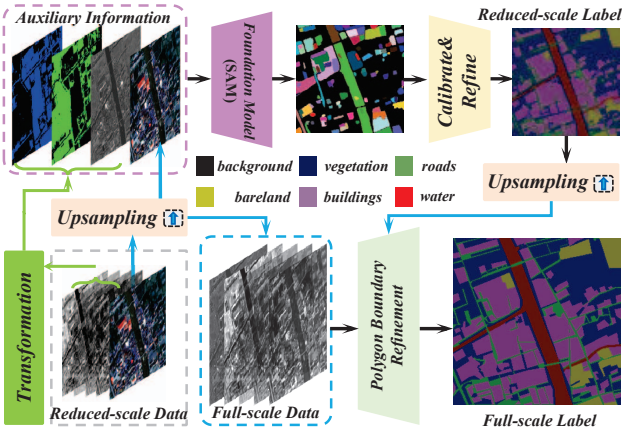


Figure 6. The construction pipeline of SegGF datasets.

where ρ represents the balancing parameter. After completing the training of pan-sharpening network, ASE is separately optimized under binary cross entropy loss [7]:

$$\mathcal{L}_{bce} = -\frac{1}{N}\sum_{N} y_i \log(p_i) + (1 - y_i)\log(1 - p_i), \quad (20)$$

where y_i represents the label of patch samples, and p_i stands for the predicted probability. Notably, the labels are derived by comparing PSNR [14] of the results produced in adjacent stages. If the PSNR of the subsequent stage surpasses that of the preceding stage, the label is set to true; vice versa.

2.5. Datasets Annotations

To systematically complete pan-sharpening assessments, we annotate the paired MS and PAN images acquired from the GaoFen-2 satellite to create the SegGF dataset, serving as a semantic-level benchmark for pan-sharpening. Two critical factors are considered: the segmentation categories and the scale of annotation. We adhere to the typical framework for land cover mapping, categorizing the data into five representative types excluding the background: vegetation, buildings, roads, bareland, and water. Due to the absence of real HRMS images, we meticulously develop a multi-scale annotation pipeline that facilitates a transition from coarse to fine-grained calibration at both reduced and full scales, as illustrated in Fig. 6. Specifically, the visible spectrum bands from the original data are sent to SAM [18] to obtain a coarse segmentation template, while the remaining spectral bands are utilized to generate multi-spectral indices, like NDVI [36] and NDWI [10], serving as auxiliary information for determining the categories and boundaries of the targets along with the PAN image. Then experts conduct calibration and refine segmentation labels. Subsequently, the reduced-scale labels are resampled to produce plausible full-scale labels, followed by the refinement of internal polygons to obtain high-quality segmentation labels.

The SegGF dataset contains 522 annotated image pairs with varying resolutions, which consists of over 100,000 polygons for each scale. Notably, this dataset not only serves as the first evaluation benchmark by leveraging semantic-level labels for pan-sharpening performance validation, but also provides indispensable data support for the joint learning of high- and low-level vision applications in pan-sharpening. It can be effectively utilized for land cover mapping and panoptic segmentation tasks.

3. Experiment

3.1. Datasets and Benchmark

Extensive experiments are conducted over three datasets with a scaling factor of 4, namely QuickBird (four bands), GaoFen-2 (four bands), and Worldview-II (eight bands). All image pairs are adaptively cropped and augmented to construct over 10,000 samples. Twelve representative methods are selected for comprehensive comparisons at both reduced and full scales. Traditional methods include AWLP [29], BT-H [24], MTF_GLP [1] and MF [30]. Deep learning methods are PanNet [55], HFIN [35], P2Sharpen [58], PanGAN [25], PanFlowNet [54], ZeroSharpen [42] and DISPNet [41]. Five reference metrics are selected for reduced-scale performance evaluation, including ERGAS [39], RMSE, QAVE [47], SAM [2], SSIM [48], while three mainstream metrics are used to assess the full-scale performance, including D_λ, D_s and QNR [3]. Notably, the non-reference metrics are evaluated on artificially manipulated parameters, not necessarily coinciding with the visual effects.

3.2. Implemented Details

During the training process of the pan-sharpening network, the learning rate is set to 3×10^{-4} and decreases by half per 200 epochs with Adam optimizer for 600 epochs with a batch size of 16. The number of unfolding stages is set to $K = 4$, while the coefficients for the loss functions are specified as $\alpha = 0.1, \beta = 0.01, \gamma = 0.01, \kappa = 1, \rho = 0.1$, and $\lambda = -10$. For the training process of the ASE, we employ Adam optimizer with the learning rate as 1×10^{-3} to train the network for 200 epochs. All the experiments are conducted on a desktop with 2.6GHz AMD EPYC 7H12 CPU, NVIDIA GeForce RTX 3090, and 128 GB RAM based on the Pytorch framework.

3.3. Comparative Experiment

Qualitative Comparison. Visual effects on reduced scale are illustrated in Fig. 7 and Fig. 8, respectively. For reduced-scale evaluation, our method completely integrates spatial details and effectively balances the preservation of spectral distribution and spatial information. The darkest residual error maps verify our method is optimal and out-

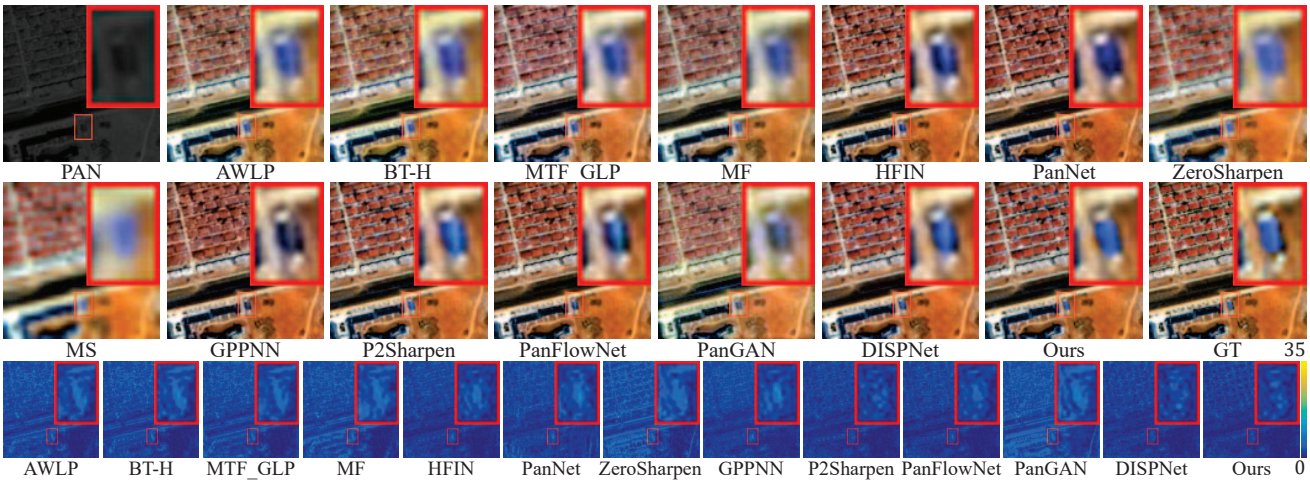


Figure 7. Visual comparisons of reduced scale on the Quickbird dataset.

Method	QuickBird (Four bands)					GaoFen-2 (Four bands)					WorldView-II (Eight bands)				
	ERGAS↓	RMSE↓	QAVE↑	SAM↓	SSIM↑	ERGAS↓	RMSE↓	QAVE↑	SAM↓	SSIM↑	ERGAS ↓	RMSE↓	QAVE↑	SAM ↓	SSIM↑
AWLP	1.8257	4.3035	0.9699	2.5602	0.9137	4.1512	17.1964	0.9418	2.2174	0.5526	9.3649	14.4759	0.7616	9.5110	0.5689
BT-H	1.7754	4.0556	0.9720	2.5618	0.9156	4.2208	16.5500	0.9210	3.5306	0.6097	10.1136	15.1895	0.7816	9.8384	0.5778
MTF_GLP	2.0052	4.5653	0.9700	2.7397	0.9005	4.7303	18.8908	0.9397	2.9984	0.5168	9.7847	14.7671	0.7754	10.2048	0.5607
MF	2.0459	4.5743	0.9688	2.7025	0.9092	3.6580	14.5560	0.9504	2.4179	0.6344	9.4076	13.5805	0.8018	8.0697	0.5778
PanNet	1.3864	3.0994	0.9864	1.8030	0.9486	1.5100	6.1709	0.9752	1.5714	0.9052	9.4621	12.1032	0.7709	9.0218	0.5301
P2Sharpen	1.1443	2.5785	0.9886	1.6353	0.9602	1.3108	5.1808	0.9756	1.5258	0.9168	5.2455	7.9507	0.8558	6.7570	0.6711
PanGAN	2.2133	5.1647	0.9609	3.0033	0.8869	2.9645	11.8193	0.9495	2.5425	0.6975	9.1934	13.8314	0.7041	10.2032	0.4273
ZeroSharpen	2.0543	4.8037	0.9715	2.4797	0.8923	2.9073	11.4732	0.9390	2.8784	0.7345	11.8535	25.8705	0.6517	13.6379	0.4348
PanFlowNet	1.1210	2.5557	0.9892	1.5636	0.9599	1.4016	5.7743	0.9782	1.4189	0.9254	4.9416	7.2028	0.8890	5.5391	0.6951
DISPNet	1.1164	2.5044	0.9901	1.5299	0.9625	1.2515	4.9532	0.9771	1.4484	0.9223	4.9051	6.9986	0.8954	5.3024	0.7064
GPPNN	1.2934	2.8157	0.9869	1.8032	0.9560	1.4322	5.6885	0.9723	1.6507	0.9126	5.5625	8.0170	0.8574	6.4577	0.6777
HFIN	1.1600	2.6269	0.9885	1.6251	0.9591	1.3090	5.3292	0.9772	1.4643	0.9225	5.5925	7.4880	0.8896	5.5475	0.7004
Ours	**1.0177**	**2.3009**	**0.9916**	**1.4113**	**0.9666**	**1.1118**	**4.4235**	**0.9811**	**1.2879**	**0.9330**	**4.6692**	**6.7729**	**0.8991**	**5.2666**	**0.7117**

Table 1. Quantitative results of reduced scale over three datasets. **bold** is the best and underline is the second best.

performs the others. For full-scale assessment, our method shows the finest visual effect among them, showcasing the remarkable generalization capability to real-world scenes.

Quantitative Evaluation. Quantitative results of each method on reduced and full scales are shown in Tab. 1 and Tab. 2, respectively. Apparently, our method surpasses the other methods on five reference metrics and demonstrates competitive performance on three non-reference metrics.

3.4. Ablation Study

To fully exploit the effectiveness of each design, we conduct ablation studies that include four categories: (i) Effects of the number of stages K, (ii) influence of different core modules, (iii) impact of loss functions $\mathcal{L}_{phase}$, and (iv) robustness of ASE module under different criteria.

Effects of the Number of Stages. We present the quantitative analysis for different K, as illustrated in Tab. 3. Both the activation and inactivation of ASE mechanism are considered. The results indicate reference assessments improve as stage K increases, but yield the unpredictable trend in the non-reference evaluation. The activation of ASE brings performance improvement in all stage settings. Besides, as the network depth increases, the proportion of overfitting samples also rises, which demonstrates the excessive network depth deteriorates the quality of samples across the stages. Notably, we select $K = 4$ in our implementation to achieve an optimal balance between the network performance and computational efficiency.

Method	QuickBird (Four bands)			GaoFen-2 (Four bands)			WorldView-II (Eight bands)		
	QNR↑	D_λ↓	D_s↓	QNR↑	D_λ↓	D_s↓	QNR↑	D_λ↓	D_s↓
AWLP	0.8406	0.0736	0.0934	0.7402	0.1052	0.1736	0.8437	0.0680	0.0970
BT-H	0.8456	0.0655	0.0959	0.6863	0.1141	0.2258	0.8155	0.0786	0.1190
MTF_GLP	0.8091	0.0916	0.1104	0.7083	0.1282	0.1885	0.8148	0.0899	0.1077
MF	0.8038	0.0909	0.1167	0.7080	0.1316	0.1856	0.8278	0.0819	0.1011
PanNet	0.9402	0.0264	0.0349	0.7662	0.1427	0.1059	0.8866	0.0366	0.0799
P2Sharpen	0.9327	0.0237	0.0450	0.8494	0.0547	0.1015	0.8802	0.0310	0.0924
PanGAN	0.8462	0.0815	0.0795	0.7809	0.0884	0.1439	0.8306	0.0935	0.0838
ZeroSharpen	0.9471	**0.0141**	0.0399	0.9117	**0.0067**	0.0823	**0.9366**	**0.0221**	**0.0423**
PanFlowNet	0.9370	0.0252	0.0390	0.8555	0.0717	0.0787	0.8747	0.0489	0.0815
DISPNet	0.9424	0.0232	0.0355	0.9007	0.0380	0.0638	0.8933	0.0326	0.0766
GPPNN	0.9119	0.0377	0.0542	0.8172	0.0673	0.1240	0.8194	0.0770	0.1161
HFIN	0.9256	0.0255	0.0503	0.8418	0.0570	0.1074	0.8904	0.0371	0.0753
Ours	**0.9500**	0.0193	**0.0315**	**0.9443**	0.0294	**0.0272**	0.8996	0.0304	0.0725

Table 2. Quantitative results of full-scale over three datasets. **Bold** is the best and underline is the second best.

Influence of Different Core Modules. Our core modules

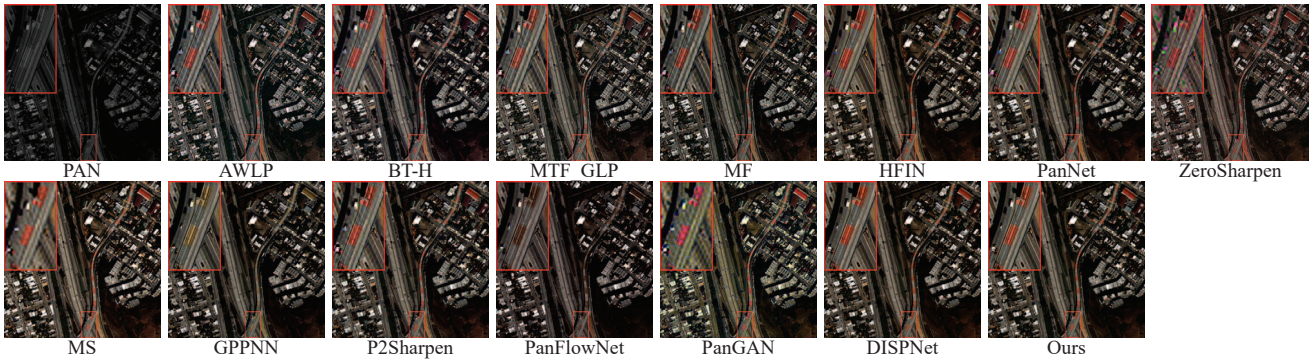


Figure 8. Visual comparisons on full scale on the WorldView-II dataset.

K	Time(s)	ERGAS↓	RMSE↓	QAVE↑	SAM ↓	SSIM↑	QNR↑	D_λ↓	D_s↓	Proportion
1	0.0177	1.2461	4.9091	0.9790	1.3803	0.9248	0.9161	0.0369	0.0489	-
2	0.0294	1.1847	4.6857	0.9793	1.3673	0.9280	0.9099	0.0429	0.0494	26.80%
	0.0314	1.1810	4.6705	0.9794	1.3664	0.9282	0.9104	0.0432	0.0486	
3	0.0400	1.1394	4.5278	0.9803	1.3235	0.9308	0.9091	0.0392	0.0539	27.80%
	0.0429	1.1374	4.5190	0.9804	1.3248	0.9308	0.9119	0.0382	0.0520	
4	0.0504	1.1140	4.4320	0.9810	1.2865	0.9343	0.9440	0.0292	0.0277	27.60%
	0.0584	1.1118	4.4235	0.9811	1.2879	0.9330	0.9443	0.0294	0.0272	
5	0.0608	1.0936	4.3311	0.9822	1.2283	0.9374	0.9271	0.0338	0.0406	33.14%
	0.0703	1.0910	4.3202	0.9823	1.2323	0.9374	0.9264	0.0344	0.0407	
6	0.0726	1.0731	4.2632	0.9823	1.2271	0.9392	0.9187	0.0390	0.0441	32.76%
	0.0896	1.0674	4.2405	0.9824	1.2297	0.9390	0.9200	0.0391	0.0427	
10	0.1134	1.0331	4.1071	0.9830	1.1796	0.9421	0.9326	0.0319	0.0368	33.46%
	0.1305	1.0288	4.0894	0.9831	1.1846	0.9422	0.9332	0.0323	0.0358	

Table 3. Quantitative results, running efficiency, overfitting patches proportion of our method with different stage numbers on the GaoFen-2 dataset. Colored column means activating ASE.

comprise $\mathcal{L}_{morph}$, ASE and SFA. The influence of three variants is investigated, as presented in Tab. 4. Obviously, reference metrics are improved upon the adoption of each module, and the best results are achieved by employing all of them. In conclusion, these proposed modules are crucial for performance improvement.

Impact of Loss Functions. Two types of supervision on a four-stage network are considered: one incorporates $\mathcal{L}_{phase}$ and the other discards it. Visual effects are presented in Fig. 9. Evidently, $\mathcal{L}_{phase}$ promotes the consistency of sharp spatial morphology and reconstruction accuracy with the darker residual error map, indicating better decomposition performance. Furthermore, we utilize PSNR to measure the similarity between synthesized image $(R_{x^k} \circ S_{x^k}, R_{\widetilde{p}} \circ S_{\widetilde{p}})$ and the original image $x^k, \widetilde{p}$, as presented in Tab. 5. The introduction of $\mathcal{L}_{phase}$ improves the performance in all stages.

Robustness of ASE Module under Different Criteria. We freeze the pansharpening backbone and retrain ASE under different criteria and patch number N to verify its robustness, as shown in Tab. 6. Clearly, different criteria lead to different overfitting portion, which also increases as granularity gets finer. ASE performs well under PSNR and SSIM criteria but underperforms under the RGB-based

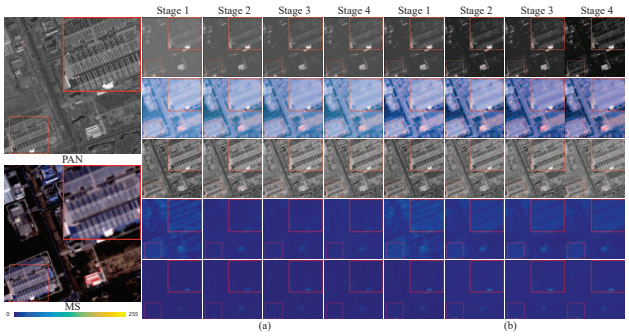


Figure 9. Qualitative verification of $\mathcal{L}_{phase}$ on GaoFen-2 dataset. From top to bottom are $R_{\widetilde{p}}$, R_{x^k}, S_{x^k}, $|R_{x^k} \circ S_{x^k} - x^k|$, $|R_{\widetilde{p}} \circ S_{x^k} - \widetilde{p}|$, respectively. (a) / (b): Feature with / without $\mathcal{L}_{phase}$.

	$\mathcal{L}_{morph}$	SFA	ASE	ERGAS↓	RMSE↓	QAVE↑	SAM ↓	SSIM↑	QNR↑	D_λ↓	D_s↓
(1)	×	✓	×	1.2078	4.8138	0.9780	1.4149	0.9241	0.8988	0.0474	0.0566
(2)	✓	✓	×	1.1140	4.4320	0.9810	**1.2865**	**0.9343**	0.9440	**0.0292**	0.0277
(3)	✓	✓	✓	**1.1118**	**4.4235**	**0.9811**	1.2879	0.9330	**0.9443**	0.0294	**0.0272**

Table 4. Quantitative comparison on the GaoFen-2 dataset. **Bold** indicates the best, and underline indicates the second best.

Config		Stage 1	Stage 2	Stage 3	Stage 4
× $\mathcal{L}_{phase}$	$R_{x^k} \circ S_{x^k} \sim x$	32.7524 dB	37.4948 dB	39.5504 dB	37.5197 dB
	$R_{\widetilde{p}} \circ S_{\widetilde{p}} \sim \widetilde{p}$	48.4023 dB	43.8845 dB	46.8753 dB	42.3337 dB
✓ $\mathcal{L}_{phase}$	$R_{x^k} \circ S_{x^k} \sim x$	33.2380 dB	40.1566 dB	41.2205 dB	39.3948 dB
	$R_{\widetilde{p}} \circ S_{\widetilde{p}} \sim \widetilde{p}$	49.4039 dB	50.8125 dB	48.2691 dB	50.3425 dB

Table 5. Quantitative verification of $\mathcal{L}_{phase}$ on GaoFen-2 dataset.

LPIPS criterion due to modality incompatibility. For the balance of performance and efficiency, we choose $N = 16$ and utilize PSNR as the criterion.

3.5. Application Verification on Segmentation

SegGF dataset is specifically designed to alleviate the unreliability of non-reference metrics by providing meticulously crafted and highly reliable semantic labels. We employ SegFormer [50] to assess the results of each method, including PAN images and resized low-quality MS images. Qualitative results of two scales are illustrated in Fig. 10. Clearly,

Criterion	N	ERGAS↓	PSNR↑	SAM ↓	SSIM↑	QNR↑	D_λ↓	D_s↓	Portion
×	×	1.1140	35.4968	1.2865	0.9343	0.9440	0.0292	0.0277	-
PSNR	16	1.1118	35.5174	1.2879	0.9330	0.9443	0.0294	0.0272	27.60%
	64	**1.0851**	**35.7213**	**1.2625**	0.9341	**0.9454**	**0.0282**	0.0271	44.31%
SSIM	16	1.1130	35.5026	1.2898	0.9344	0.9445	0.0294	0.0269	23.88%
	64	1.0919	35.6688	1.2721	**0.9350**	0.9441	0.0295	0.0270	36.43%
LPIPS	16	1.1663	35.0977	1.3475	0.9273	0.9436	0.0307	0.0265	79.07%
	64	1.1283	35.3778	1.3159	0.9311	0.9436	0.0308	**0.0264**	67.53%

Table 6. Robustness evaluation of ASE on the GaoFen-2 dataset.

	Reduced-scale						Full-scale					
	mIoU	vegetation	roads	bareland	buildings	water	mIoU	vegetation	roads	bare land	buildings	water
PAN	30.72	39.38	25.60	32.92	41.18	45.23	34.80	48.58	33.93	33.59	47.73	44.97
LRMS	37.77	60.72	31.54	47.43	49.71	37.21	40.04	62.81	37.80	50.65	53.56	35.43
AWLP	38.30	59.84	32.43	46.55	48.14	42.82	40.56	61.98	37.97	47.52	50.68	45.25
BT-H	38.72	59.58	33.54	46.88	49.73	42.59	40.96	63.14	39.51	49.95	52.30	40.84
MTF_GLP	37.43	57.84	32.93	45.22	48.57	40.04	39.30	62.12	39.13	46.76	51.10	36.70
MF	38.69	60.76	30.84	47.46	48.64	44.45	41.84	62.25	39.58	49.67	49.76	49.75
PanNet	39.43	62.18	34.62	46.58	50.73	42.46	43.32	64.25	42.59	50.63	54.40	48.07
P2Sharpen	40.83	61.81	36.18	48.82	51.56	46.62	43.88	64.41	43.78	50.75	53.43	50.88
PanGAN	39.69	60.29	35.20	47.62	49.38	45.63	42.31	61.80	39.89	49.06	52.13	50.95
ZeroSharpen	39.81	60.23	34.58	48.27	50.84	44.96	42.99	63.54	42.90	51.64	54.54	45.33
PanFlowNet	40.78	62.79	36.12	46.62	50.60	48.54	43.73	63.60	41.94	51.24	53.00	52.62
DISPNet	41.02	61.70	36.87	48.82	52.15	46.58	44.37	65.00	43.98	51.12	54.75	51.38
GPPNN	40.76	61.07	35.62	48.42	50.29	49.18	43.84	63.81	42.35	50.57	54.86	51.48
HFIN	40.92	61.56	36.22	48.88	51.49	47.39	43.88	64.54	43.84	51.39	54.84	48.65
Ours	**41.23**	61.91	36.68	48.44	51.11	49.25	**45.18**	64.42	42.60	51.46	54.96	57.66

Table 7. Segmentation performance on the SegGF dataset. **Bold** means the best average segmentation performance.

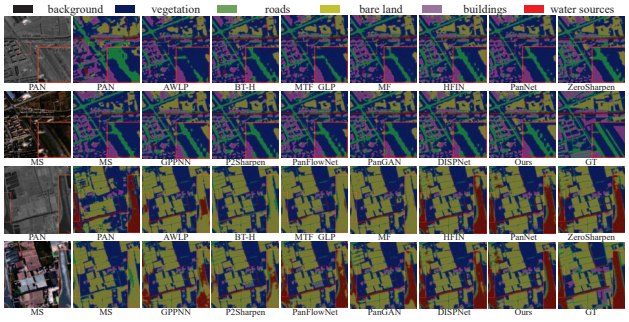


Figure 10. Segmentation results on SegGF dataset. Top and bottom groups are the results of reduced- and full-scale, respectively.

the fused images can improve accuracy in high-level tasks. Notably, our method achieves the best segmentation performance, particularly in road and water regions. Besides, we provide quantitative evaluation using the intersection over Union (IoU) metric in Tab. 7. The segmentation performance of each method at a reduced scale aligns closely with its fusion performance evaluations, indicating the rationality of using segmentation performance to evaluate the full-scale performance. In the full-scale evaluation, our method has the highest mean IoU, indicating the best performance.

3.6. Hyperspectral Image Reconstruction

Hyperspectral (HS) sensors also encounter adverse physical conditions, where the acquisition of a large amount of spectral information leads to an extreme reduction in resolution. Hence, we perform the HS image reconstruction on the CAVE dataset [56] to examine the generalization capacity of our method under more intricate conditions with more spectral bands and a higher spatial degradation ratio. We follow the experimental settings of [16] with a scaling factor of 8. Besides, eight representative methods are investigated for performance comparisons, including two conventional methods (CSU [19], HySure [31]) and six deep-learning methods (MCNet [21], MoG-DCN [8], SSRNet [61], DUNet [49], DSPNet [34], DHIF-Net [16]). Qualitative results are simply displayed by presenting the average absolute error of each band, as provided in Fig. 11. Our error map has the lowest brightness, showcasing the highest competitiveness. Moreover, four reference metrics are selected to complete the quantitative evaluation, as reported in Tab. 8. In summary, our method achieves the best performance with a reasonable quantity of parameters, which demonstrates its high scalability.

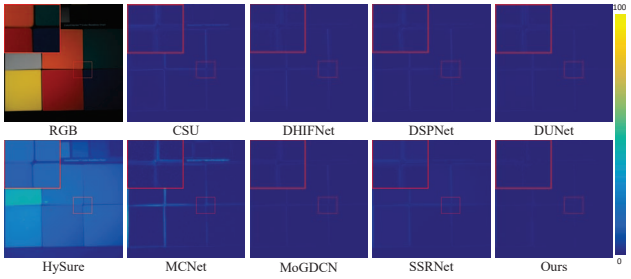


Figure 11. Residual error maps between reconstructed images and groundtruth of sponges on CAVE dataset with a scale factor of 8.

Method	PSNR↑	SAM↓	ERGAS↓	SSIM↑	Param. (M)	FLOPs (G)
CSU	41.4308	5.4780	1.1945	0.9830	-	-
HySure	21.1736	13.0832	13.3782	0.8371	-	-
MCNet	32.9982	4.4830	6.1478	0.9073	2.961	10218.500
MoG-DCN	<u>50.0094</u>	**2.0068**	<u>0.8959</u>	<u>0.9963</u>	7.071	4390.184
SSRNet	44.8474	3.5848	1.5021	0.9926	0.026	6.802
DUNet	49.2487	2.1462	0.9563	0.9958	3.514	289.675
DSPNet	48.8350	2.2589	0.9972	0.9956	6.055	422.189
DHIF-Net	48.5191	2.1999	1.0523	0.9956	22.669	3616.306
Ours	**50.6182**	<u>2.0193</u>	**0.7840**	**0.9969**	7.002	1628.253

Table 8. Quantitative results on the CAVE dataset. **Bold** is the best and <u>underline</u> is the second best.

4. Conclusion

In this paper, we propose a deep adaptive unfolded network via spatial morphology stripping and spectral filtration for pan-sharpening in a linear inverse optimization perspective. We fully exploit spatial and spectral priors by incorporating phase-oriented constraints and spectral filtration attention, enhancing the interpretability and transparency of the model. The adaptive stage-exiting mechanism is adopted to leverage the fusion diversity for optimal solutions and prevent overfitting. The first semantic-level benchmark dataset is constructed for pan-sharpening assessments. Extensive experiments on pan-sharpening, segmentation, and hyperspectral reconstruction are conducted to showcase the superior performance of our method.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (62276192, 62225113, 624B2109, 623B2079), in part by the Zhongguancun Academy Project (20240308), and in part by the Fund of National Key Laboratory of Multispectral Information Intelligent Processing Technology (61421132302).

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 Trans. Geosci. Remote. Sens.*, 40(10):2300–2312, 2002. 5

[2] Luciano Alparone, Lucien Wald, Jocelyn Chanussot, Claire Thomas, Paolo Gamba, and Lori Mann Bruce. Comparison of pansharpening algorithms: Outcome of the 2006 grs-s data-fusion contest. *IEEE Trans. Geosci. Remote. Sens.*, 45 (10):3012–3021, 2007. 5

[3] Luciano Alparone, Bruno Aiazzi, Stefano Baronti, Andrea Garzelli, Filippo Nencini, and Massimo Selva. Multispectral and panchromatic data fusion assessment without reference. *Photogramm. Eng. Remote Sens.*, 74(2):193–200, 2008. 5

[4] Alberto Arienzo, Gemine Vivone, Andrea Garzelli, Luciano Alparone, and Jocelyn Chanussot. Full-resolution quality assessment of pansharpening: Theoretical and hands-on approaches. *IEEE Geosci. Remote Sens. Mag.*, 10(3):168–201, 2022. 2

[5] Xun Cao, Hao Du, Xin Tong, Qionghai Dai, and Stephen Lin. A prism-mask system for multispectral video acquisition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(12):2423–2435, 2011. 2

[6] Matteo Ciotola, Giuseppe Guarino, Gemine Vivone, Giovanni Poggi, Jocelyn Chanussot, Antonio Plaza, and Giuseppe Scarpa. Hyperspectral pansharpening: Critical review, tools, and future perspectives. *IEEE Geosci. Remote Sens. Mag.*, 13(1):311–338, 2025. 2

[7] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Ann. Oper. Res.*, 134:19–67, 2005. 5

[8] Weisheng Dong, Chen Zhou, Fangfang Wu, Jinjian Wu, Guangming Shi, and Xin Li. Model-guided deep hyperspectral image super-resolution. *IEEE Trans. Image Process*, 30: 5754–5768, 2021. 8

[9] Xueyang Fu, Zihuang Lin, Yue Huang, and Xinghao Ding. A variational pan-sharpening with local gradient constraints. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 10265–10274, 2019. 1

[10] Bo-Cai Gao. Ndwi—a normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sens. Environ.*, 58(3):257–266, 1996. 5

[11] Meiqi Gong, Hao Zhang, Hebaixu Wang, Jun Chen, Jun Huang, Xin Tian, and Jiayi Ma. Cross-scale domain adaptation with comprehensive information for pansharpening. In *Proc. Int. Joint Conf. Artif. Intell.*, pages 785–793, 2024. 2

[12] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1):87–110, 2022. 4

[13] Xuanhua He, Ke Cao, Jie Zhang, Keyu Yan, Yingying Wang, Rui Li, Chengjun Xie, Danfeng Hong, and Man Zhou. Pan-mamba: Effective pan-sharpening with state space model. *Inf. Fusion*, 115F:102779, 2024. 2

[14] Alain Horé and Djemel Ziou. Is there a relationship between peak-signal-to-noise ratio and structural similarity index measure? *IET Image Process.*, 7(1):12–24, 2013. 5

[15] 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 *Proc. ACM Int. Conf. Multimed.*, pages 1486–1494, 2024. 2

[16] Tao Huang, Weisheng Dong, Jinjian Wu, Leida Li, Xin Li, and Guangming Shi. Deep hyperspectral image fusion network with iterative spatio-spectral regularization. *IEEE Trans. Comput. Imaging*, 8:201–214, 2022. 8

[17] Yanan Jia, Qiming Hu, Renwei Dian, Jiayi Ma, and Xiaojie Guo. Paps: Progressive attention-based pan-sharpening. *IEEE CAA J. Autom. Sin.*, 11(2):391–404, 2024. 2

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proc. IEEE Int. Conf. Comput. Vision*, pages 4015–4026, 2023. 5

[19] Charis Lanaras, Emmanuel Baltsavias, and Konrad Schindler. Hyperspectral super-resolution by coupled spectral unmixing. In *Proc. IEEE Int. Conf. Comput. Vision*, pages 3586–3594, 2015. 8

[20] Jiaxin Li, Danfeng Hong, Lianru Gao, Jing Yao, Ke Zheng, Bing Zhang, and Jocelyn Chanussot. Deep learning in multimodal remote sensing data fusion: A comprehensive review. *Int. J. Appl. Earth Obs. Geoinf.*, 112:102926, 2022. 2

[21] Qiang Li, Qi Wang, and Xuelong Li. Mixed 2d/3d convolutional network for hyperspectral image super-resolution. *Remote Sens.*, 12(10):1660, 2020. 8

[22] Huangxing Lin, Yuhang Dong, Xinghao Ding, Tianpeng Liu, and Yongxiang Liu. Unsupervised pan-sharpening via mutually guided detail restoration. In *Proc. AAAI Conf. Artif. Intell.*, pages 3386–3394, 2024. 2

[23] Jinyuan Liu, Runjia Lin, Guanyao Wu, Risheng Liu, Zhongxuan Luo, and Xin Fan. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *International Journal of Computer Vision*, 132(5):1748–1775, 2024. 2

[24] Simone Lolli, Luciano Alparone, Andrea Garzelli, and Gemine Vivone. Haze correction for contrast-based multispectral pansharpening. *IEEE Geosci. Remote Sens. Lett.*, 14(12): 2255–2259, 2017. 5

[25] Jiayi Ma, Wei Yu, Chen Chen, Pengwei Liang, Xiaojie Guo, and Junjun Jiang. Pan-gan: An unsupervised pan-sharpening method for remote sensing image fusion. *Inf. Fusion*, 62: 110–120, 2020. 2, 5

[26] Yura Malitsky and Konstantin Mishchenko. Adaptive proximal gradient method for convex optimization. *Adv. Neural Inform. Process. Syst.*, 37:100670–100697, 2025. 3

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

[28] Ge Meng, Jingjia Huang, Yingying Wang, Zhenqi Fu, Xinghao Ding, and Yue Huang. Progressive high-frequency reconstruction for pan-sharpening with implicit neural representation. In *Proc. AAAI Conf. Artif. Intell.*, pages 4189–4197, 2024. 2

[29] 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 Trans. Geosci. Remote. Sens.*, 43(10):2376–2385, 2005. 5

[30] Rocco Restaino, Gemine Vivone, Mauro Dalla Mura, and Jocelyn Chanussot. Fusion of multispectral and panchromatic images based on morphological operators. *IEEE Trans. Image Process*, 25(6):2882–2895, 2016. 5

[31] Miguel Simoes, José Bioucas-Dias, Luis B Almeida, and Jocelyn Chanussot. A convex formulation for hyperspectral image superresolution via subspace-based regularization. *IEEE Trans. Geosci. Remote. Sens.*, 53(6):3373–3388, 2014. 8

[32] Jiechong Song, Bin Chen, and Jian Zhang. Dynamic path-controllable deep unfolding network for compressive sensing. *IEEE Trans. Image Process*, 32:2202–2214, 2023. 2

[33] Yubao Sun, Ying Yang, Qingshan Liu, Jiwei Chen, Xiao-Tong Yuan, and Guodong Guo. Learning non-locally regularized compressed sensing network with half-quadratic splitting. *IEEE Trans. Multimedia*, 22(12):3236–3248, 2020. 3

[34] Yucheng Sun, Han Xu, Yong Ma, Minghui Wu, Xiaoguang Mei, Jun Huang, and Jiayi Ma. Dual spatial-spectral pyramid network with transformer for hyperspectral image fusion. *IEEE Trans. Geosci. Remote. Sens.*, 61:5526016, 2023. 8

[35] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 25922–25931, 2024. 5

[36] Compton J Tucker. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sens. Environ.*, 8(2):127–150, 1979. 5

[37] Gemine Vivone. Robust band-dependent spatial-detail approaches for panchromatic sharpening. *IEEE Trans. Geosci. Remote. Sens.*, 57(9):6421–6433, 2019. 1

[38] 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 Geosci. Remote Sens. Mag.*, 9(1):53–81, 2020. 1

[39] Lucien Wald. *Data fusion: definitions and architectures: fusion of images of different spatial resolutions*. Presses des MINES, 2002. 5

[40] Lucien Wald, Thierry Ranchin, and Marc Mangolini. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogramm. Eng. Remote Sens.*, 63(6):691–699, 1997. 2

[41] Hebaixu Wang, Meiqi Gong, Xiaoguang Mei, Hao Zhang, and Jiayi Ma. Deep unfolded network with intrinsic supervision for pan-sharpening. In *Proc. AAAI Conf. Artif. Intell.*, pages 5419–5426, 2024. 2, 5

[42] Hebaixu Wang, Hao Zhang, Xin Tian, and Jiayi Ma. Zero-sharpen: A universal pansharpening method across satellites for reducing scale-variance gap via zero-shot variation. *Inf. Fusion*, 101:102003, 2024. 5

[43] Jiaming Wang, Tao Lu, Xiao Huang, Ruiqian Zhang, and Xiaoxiao Feng. Pan-sharpening via conditional invertible neural network. *Inf. Fusion*, 101:101980, 2024. 2

[44] Jiaming Wang, Xitong Chen, Xiao Huang, Ruiqian Zhang, Yu Wang, and Tao Lu. Rethinking the role of panchromatic images in pan-sharpening. *IEEE Trans. Multimedia*, 2025. 2

[45] Xinya Wang, Qian Hu, Yingsong Cheng, and Jiayi Ma. Hyperspectral image super-resolution meets deep learning: A survey and perspective. *IEEE CAA J. Autom. Sin.*, 10(8): 1664–1687, 2023. 1

[46] Yingying Wang, Yunlong Lin, Ge Meng, Zhenqi Fu, Yuhang Dong, Linyu Fan, Hedeng Yu, Xinghao Ding, and Yue Huang. Learning high-frequency feature enhancement and alignment for pan-sharpening. In *Proc. ACM Int. Conf. Multimed.*, pages 358–367, 2023. 2

[47] Zhou Wang and Alan C Bovik. A universal image quality index. *IEEE Signal Process. Lett.*, 9(3):81–84, 2002. 5

[48] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process*, 13(4): 600–612, 2004. 5

[49] Jiajun Xiao, Jie Li, Qiangqiang Yuan, and Liangpei Zhang. A dual-unet with multistage details injection for hyperspectral image fusion. *IEEE Trans. Geosci. Remote. Sens.*, 60:1–13, 2021. 8

[50] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Proces. Syst.*, 34:12077–12090, 2021. 7

[51] Shuang Xu, Jiangshe Zhang, Zixiang Zhao, Kai Sun, Junmin Liu, and Chunxia Zhang. Deep gradient projection networks for pan-sharpening. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 1366–1375, 2021. 2

[52] Keyu Yan, Man Zhou, Li Zhang, and Chengjun Xie. Memory-augmented model-driven network for pansharpening. In *Proc. Eur. Conf. Comput. Vis.*, pages 306–322. Springer, 2022. 2

[53] Gang Yang, Man Zhou, Keyu Yan, Aiping Liu, Xueyang Fu, and Fan Wang. Memory-augmented deep conditional unfolding network for pan-sharpening. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 1788–1797, 2022. 1

[54] Gang Yang, Xiangyong Cao, Wenzhe Xiao, Man Zhou, Aiping Liu, Xun Chen, and Deyu Meng. Panflownet: A flow-based deep network for pan-sharpening. In *Proc. IEEE Int. Conf. Comput. Vision*, pages 16857–16867, 2023. 5

[55] Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, and John Paisley. Pannet: A deep network architecture for pan-sharpening. In *Proc. IEEE Int. Conf. Comput. Vision*, pages 5449–5457, 2017. 5

[56] Fumihito Yasuma, Tomoo Mitsunaga, Daisuke Iso, and Shree K Nayar. Generalized assorted pixel camera: post-capture control of resolution, dynamic range, and spectrum. *IEEE Trans. Image Process*, 19(9):2241–2253, 2010. 8

[57] Bing Zhang, Yuanfeng Wu, Boya Zhao, Jocelyn Chanussot, Danfeng Hong, Jing Yao, and Lianru Gao. Progress and challenges in intelligent remote sensing satellite systems. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 15:1814–1822, 2022. 1

[58] Hao Zhang, Hebaixu Wang, Xin Tian, and Jiayi Ma. P2sharpen: A progressive pansharpening network with deep spectral transformation. *Inf. Fusion*, 91:103–122, 2023. 5

[59] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 3217–3226, 2020. 2

[60] Kai Zhang, Feng Zhang, Wenbo Wan, Hui Yu, Jiande Sun, Javier Del Ser, Eyad Elyan, and Amir Hussain. Panchromatic and multispectral image fusion for remote sensing and earth observation: Concepts, taxonomy, literature review, evaluation methodologies and challenges ahead. *Inf. Fusion*, 93: 227–242, 2023. 1

[61] Xueting Zhang, Wei Huang, Qi Wang, and Xuelong Li. Ssr-net: Spatial–spectral reconstruction network for hyperspectral and multispectral image fusion. *IEEE Trans. Geosci. Remote. Sens.*, 59(7):5953–5965, 2020. 8

[62] Yanan Zhao, Yuelong Li, Haichuan Zhang, Vishal Monga, and Yonina C Eldar. Deep, convergent, unrolled half-quadratic splitting for image deconvolution. *IEEE Trans. Comput. Imaging*, 10:574–588, 2024. 3

[63] Huanyu Zhou, Qingjie Liu, and Yunhong Wang. Panformer: A transformer based model for pan-sharpening. In *Proc. IEEE Int. Conf. Multimedia Expo*, pages 1–6, 2022. 2

[64] 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 Trans. Pattern Anal. Mach. Intell.*, 47(7):5281–5298, 2025. 1