Deep Unfolded Network with Intrinsic Supervision for Pan-Sharpening

Hebaixu Wang, Meiqi Gong, Xiaoguang Mei, Hao Zhang, Jiayi Ma*

Electronic Information School, Wuhan University, Wuhan 430072, China
{wanghebaixu, meixiaoguang, zhpersonalbox, jyma2010}@gmail.com, meiqigong@whu.edu.cn

Abstract

Existing deep pan-sharpening methods lack the learning of complementary information between PAN and MS modalities in the intermediate layers, and exhibit low interpretability due to their black-box designs. To this end, an interpretable deep unfolded network with intrinsic supervision for pan-sharpening is proposed. Building upon the observation degradation process, it formulates the pan-sharpening task as a variational model minimization with spatial consistency prior and spectral projection prior. The former prior requires a joint component decomposition of PAN and MS images to extract intrinsic features. By being supervised in the intermediate layers, it can selectively provide high-frequency information for spatial enhancement. The latter prior constrains the intensity correlation between MS and PAN images derived from physical observations, so as to improve spectral fidelity. To further enhance the transparency of network design, we develop an iterative solution algorithm following the half-quadratic splitting to unfold the deep model. It rigorously adheres to the variational model, significantly enhancing the interpretability behind network design and efficiently alternating the optimization of the network. Extensive experiments demonstrate the advantages of our method compared to state-of-the-arts, showcasing its remarkable generalization capability to real-world scenes. Our code is publicly available at https://github.com/Baixuzx7/DISPNet.

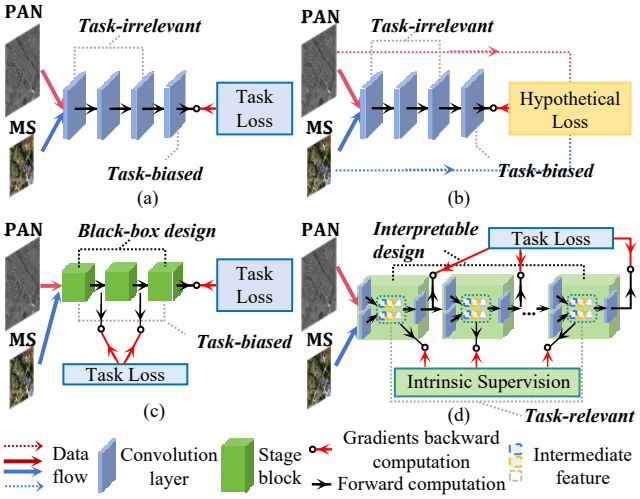

Figure 1: The overview of the four pan-sharpening training paradigm methods. (a) Supervised Learning; (b) Unsupervised Learning; (c) Progressive Supervision Learning; (d) Ours. Task loss supervises the output by references, and the hypothetical loss constrains the output with the source images based on simplified physical observations.

Introduction

Due to the physical limitations in satellite systems, it is difficult to achieve high spatial resolution and high spectral resolution simultaneously through the equipped sensors (Zhang et al. 2022). Fortunately, the captured multi-spectral (MS) images and panchromatic (PAN) images offer complementary information, making it possible to generate ideal high-resolution multi-spectral (HRMS) images. Specifically, MS images offer rich spectral information across multiple bands with lower spatial resolution. The single-band PAN images possess higher spatial resolution at the expense of spectral richness. In this context, pan-sharpening has emerged as a viable technique (Vivone et al. 2020), which aims to fuse the MS image and PAN image to produce HRMS images that preserve the spatial resolution of the PAN image, while incorporating the spectral richness of the MS image. For the excellent characteristics of HRMS images, the pan-sharpening technology holds significant importance across diverse domains (Casagli et al. 2023; Wang et al. 2023).

The traditional pan-sharpening techniques can be divided into three categories, namely Component Substitution (CS), Multi-resolution Analysis (MRA), and Model-based Optimization (MO). CS methods utilize reversible projection algorithms to extract specific components from both the MS and PAN images, which are then replaced or merged to restore the pan-sharpened image (Ghahremani and Ghassemian 2016). In the MRA approaches, a multi-resolution decomposition technique is employed to extract high-frequency spatial details from the PAN image. These details are subsequently injected into the resized MS component, aiming to enhance the spatial textures (Alparone et al. 2016). These two types of methods involve a trade-off between the preservation of spectral accuracy and the enhance-

*Corresponding author.

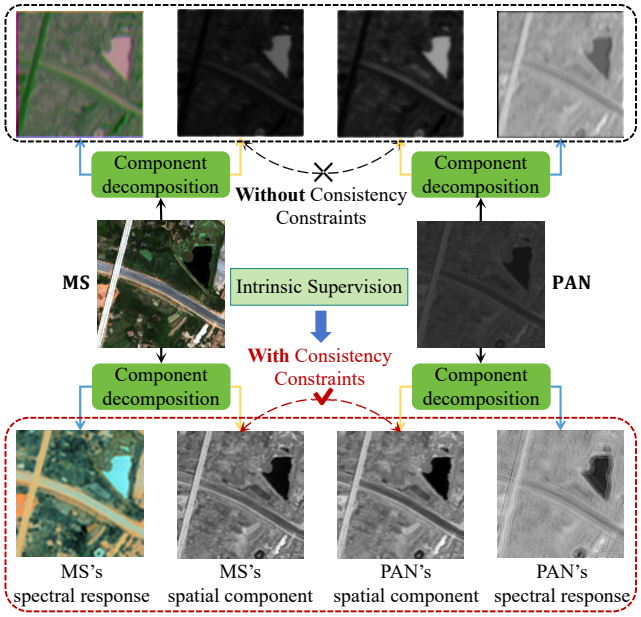

Figure 2: Examples of the component decomposition of the PAN and MS images. The bottom group and top group respectively displays extracted intermediate features under the consistency constraints or not. Intrinsic supervision on consistency constraints provides appropriate allocation of low-frequency spectral information and high-frequency spatial details for the extracted features.

ment of spatial details, typically sacrificing certain spectral fidelity to obtain richer spatial details. MO methods treat pan-sharpening as an ill-posed optimization problem and introduce specific priors to establish physically interpretable models (Vicinanza et al. 2014). These methods utilize iterative optimization algorithms to seek solutions (Ruder 2016). However, their potential is restricted by imprecise spectral and spatial assumptions in the linear space (Lu et al. 2022).

Owing to the powerful non-linear fitting ability of the neural networks, pan-sharpening has witnessed significant growth (Javan et al. 2021). The existing deep learning training paradigms for pan-sharpening can be categorized into three types, as depicted in Figure 1 (a)-(c). Firstly, traditional supervised learning applies supervision only to the final layer and propagates it backward to preceding layers, resulting in unconstrained and ill-posed internal features in the intermediate layers of the network (Meng et al. 2022). Secondly, traditional unsupervised learning relies on the employment of heuristic constraints to promote the fusion of information between PAN and MS images (Zhou et al. 2022). Nevertheless, these simplified physical constraints struggle to maintain a balanced performance between spatial and spectral information. Thirdly, progressive supervision learning typically involves introducing corresponding task losses at multiple preceding stages of a network model, aiming to enhance network performance stage by stage, and ultimately achieve improved performance in the final stage of the network (Cao et al. 2021). As a result, it makes all the stages task-biased, which weakens the network's ability to extract hierarchical feature information, leading to information homogenization and hampering performance improvement (Yang et al. 2022). Furthermore, most existing deep learning-based methods predominantly emphasize constructing deeper and more sophisticated network topologies in a black-box manner, neglecting the consideration of model interpretability and overlooking the intrinsic information between different modalities (Jin, Gu, and Xie 2022).

In summary, the existing state-of-the-art (SOTA) pan-sharpening methods suffer from two major drawbacks: the absence of supervision constraints on the internal feature information within intermediate layers of the network, and the lack of network model designs with physical interpretability. To this end, we propose an interpretable deep unfolded network with intrinsic supervision for pan-sharpening, which combines the advantages of model-based and deep-learning approaches. To begin, we consider the degradation process of the images, and formulate the pan-sharpening problem as an optimization of a variational model with two novel priors, namely spatial consistency prior and spectral projection prior. Concerning the spatial consistency prior, it can be observed that PAN and MS sensors capture the same scene in distinct modalities, inherently encompassing consistent spatial components and unique spectral response characteristics, as illustrated in Figure 2. The former corresponds to the spatial attributes of terrestrial objects, while the latter corresponds to the interactions between different spectral bands and environmental factors. Hence, we decompose the features in the intermediate layers of the network into intrinsic spatial components and spectral responses and form intrinsic supervision by imposing constraints on these intrinsic information, as shown in Figure 1 (d). This process aligns the learned features of the network more closely with physical observations, thereby facilitating subsequent network learning and enhancing the performance of pan-sharpening. Moreover, the constrained spatial components exhibit more high-frequency spatial details, while their spectral responses possess unique spectral characteristics. Regarding the spectral projection prior, the energy acquired in the wide PAN band undergoes filtration, culminating in the MS image with energy distributed across several narrow spectral bands. Consequently, the paired PAN and MS images inherently exhibit a strong intensity correlation, which is treated as a prior to promote the correlation among different spectra. Additionally, we introduce an innovative iterative algorithm to guide model design, augmenting interpretability. Extensive experiments are conducted to demonstrate the superior performance of our proposed method both qualitatively and quantitatively, showcasing its remarkable generalization capability to real-world scenes.

Our contributions are summarized as followed:

- We formulate pan-sharpening as an optimization of a variational model incorporating spatial consistency and spectral projection priors, with the aim of improving spatial quality and strengthening modality correlation.
- For the unfolding of deep networks under the variational framework, an iterative algorithm integrated with convo-

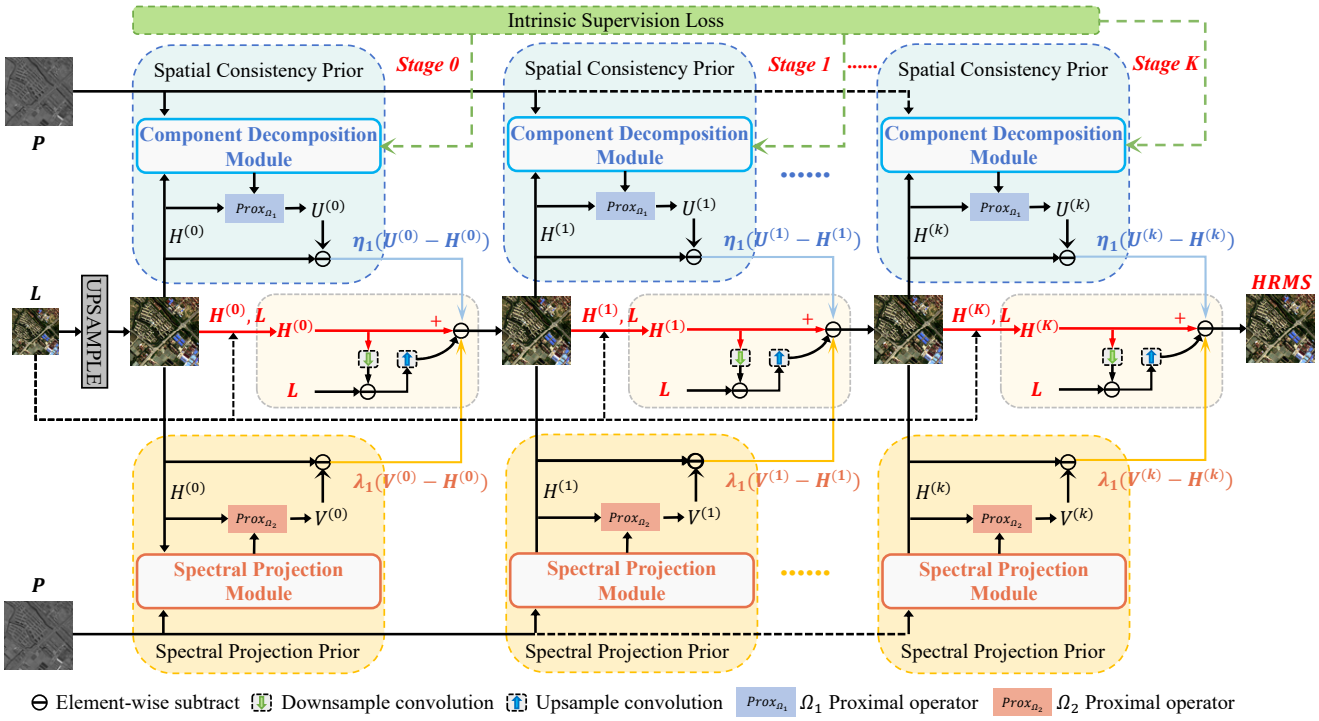

Figure 3: The workflow of our method.

lutional neural networks is introduced for transparent design to enhance the network interpretability.

- Task-relevant loss and intrinsic supervision loss are tailored to provide supervision for the final and intermediate layers of the network, which facilitates the information interaction between different modalities and improves the pan-sharpening performance.

Method

Model Formulation

Pan-sharpening can be conceptualized as a problem of single MS image super-resolution, guided by the associated PAN image. Mathematically, the low-resolution MS image L can be represented as the degradation version of the high-resolution MS H, denoted as $L = DBH + N$. In this equation, B and D stand for blurring and down-sampling operators, respectively, while N represents the noise. Based on this observation model, the high-resolution MS images can be obtained by solving the following minimization problem:

$$\arg\min_{H} \frac{1}{2}\|L - DBH\|_2^2 + \textstyle\sum_i \omega_i \Omega_i(H, P), \qquad (1)$$

where P denotes the PAN image containing complementary spatial information to assist in the restoration of the HRMS image, ω_i is the i^{th} Lagrange multiplier, and $\Omega_i(H, P)$ describes the i^{th} regularization function. Motivated by the aforementioned observations that PAN and MS images describe the same scene and share consistent spatial components under the same resolution, we introduce a spatial consistency prior $\Omega_1(P, H) = \|T_p P - T_h H\|_2^2$, where T_p and T_h represents the extraction of spatial components from the PAN image and MS image, respectively. Draw from the spectral observation model in the imaging process, the intensity correlation between the two modalities is considered as the spectral projection prior $\Omega_2(P, H) = \|IP - DBH\|_2^2$, where I functions to transfer intensity information from the PAN band to the MS bands, thereby establishing a relationship between their intensities. With the incorporation of these priors, Eq. (1) can be reformulated as:

$$\arg\min_{H} \frac{1}{2}\|L - DBH\|_2^2 + \frac{\omega_1}{2}\|T_p P - T_h H\|_2^2 + \frac{\omega_2}{2}\|IP - DBH\|_2^2. \qquad (2)$$

Optimization

Following the framework of half-quadratic splitting (HQS) (Sun et al. 2020), two auxiliary variables U and V are introduced to reformulate Eq. (2):

$$\arg\min_{H,U,V} \frac{1}{2}\|L - DBH\|_2^2 + \frac{\eta_1}{2}\|U - H\|_2^2 + \frac{\eta_2}{2}\Omega_1(P, U) + \frac{\lambda_1}{2}\|V - H\|_2^2 + \frac{\lambda_2}{2}\Omega_2(P, V), \qquad (3)$$

where η_1, η_2, λ_1 and λ_2 are penalty parameters. To achieve the unrolling inference, Eq. (3) can be divided into the following three sub-problems and solved alternatively:

$$U^{(k)} = \arg\min_{U} \frac{\eta_1}{2}\left\|U - H^{(k)}\right\|_2^2 + \frac{\eta_2}{2}\left\|T_p P - T_h U\right\|_2^2, \qquad (4)$$

$$V^{(k)} = \arg\min_{V} \frac{\lambda_1}{2}\left\|V - H^{(k)}\right\|_2^2 + \frac{\lambda_2}{2}\left\|DBV - IP\right\|_2^2, \qquad (5)$$

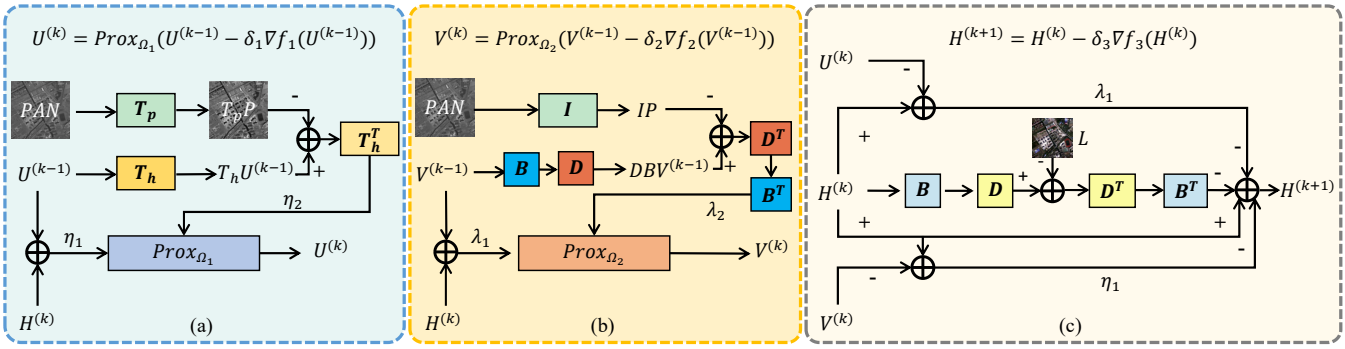

Figure 4: Architecture of the overall network. (a) UNet, (b) VNet, (c) HNet. Each network corresponds to one operation of the iterative step for updating U, V and H, respectively.

$$H^{(k+1)} = \arg\min_H \frac{1}{2}\|L - DBH\|_2^2 + \frac{\eta_1}{2}\|U^{(k)} - H\|_2^2 + \frac{\lambda_1}{2}\|V^{(k)} - H\|_2^2, \quad (6)$$

where k denotes the iteration index. We employ the proximal gradient projection method to solve these three subproblems:

$$U^{(k)} = prox_{\Omega_1}\left(U^{(k-1)} - \delta_1 \nabla f_1(U^{(k-1)})\right), \quad (7)$$

$$V^{(k)} = prox_{\Omega_2}\left(V^{(k-1)} - \delta_2 \nabla f_2(V^{(k-1)})\right), \quad (8)$$

$$H^{(k+1)} = H^{(k)} - \delta_3 \nabla f_3(H^{(k)}), \quad (9)$$

where $prox_{\Omega_1}(\cdot)$ and $prox_{\Omega_2}(\cdot)$ are proximal operators corresponding to penalty $\Omega_1(\cdot)$ and $\Omega_2(\cdot)$, $\delta_{(\cdot)}$ is the step size. Furthermore, the gradient-related notations are detailed as:

$$\nabla f_1(U^{(k-1)}) = \eta_1(U^{(k-1)} - H^{(k)}) + \eta_2 T_h^T(T_h U^{(k-1)} - T_p P), \quad (10)$$

$$\nabla f_2(V^{(k-1)}) = \lambda_1(V^{(k-1)} - H^{(k)}) + \lambda_2(DB)^T(DBV^{(k-1)} - IP), \quad (11)$$

$$\nabla f_3(H^{(k)}) = (DB)^T(DBH^{(k)} - L) + \eta_1(H^{(k)} - U^{(k)}) + \lambda_1(H^{(k)} - V^{(k)}). \quad (12)$$

Deep Unfolded Network

The constructed deep unfolded network comprises K stages, deliberately designed to align with K iterations in the optimization algorithm, as illustrated in Figure 3. In each stage, three variables, namely U, V, and H, are alternately updated, corresponding to the UNet, VNet, and HNet architectures as illustrated in Figure 4. Specifically, the operators T_p, T_h, and T_h^T are realized using three convolutional network layers employing 3×3 kernels along with the activation function $PRelu(\cdot)$. The degraded and blurred operator DB is emulated through a single convolution layer featuring 3×3 kernels and $Relu(\cdot)$ activation function, augmented with a max-pooling layer for downsampling at the tail end. Moreover, the inverse operator $(DB)^T$ diverges from DB by incorporating an upsampling layer rather than a downsample layer at its terminus. The operator I encompasses a max-pooling operation for downsampling and a three-layer convolution network utilizing 3×3 kernels with $PRelu(\cdot)$ activation. It is worth noting that the VNet and HNet share identical network structures for the operators DB and its inverse operators $(DB)^T$ while having distinct parameters.

Loss Functions

To stabilize the network's performance, we employ ℓ_2 supervised loss at each stage of the network, which can be expressed as:

$$\mathcal{L}_{ref} = \sum_{k=1}^{K} \left\|H^{(k)} - GT\right\|_F^2, \quad (13)$$

where $H^{(k)}$ is the output of the k^{th} stage, and GT is the reference. As previously mentioned, images of different modalities are decomposed into spatial components and spectral responses through the specific network:

$$(L_{MS}, R_{MS}) = T_h(H), \quad (L_{PAN}, R_{PAN}) = T_p(P), \quad (14)$$

where L_{MS} and R_{MS} refer to the spatial component and spectral response of MS modalities, L_{PAN} and R_{PAN} refer to the spatial component and spectral response of PAN modalities, respectively. Within identical scenes, spatial components of distinct modalities are anticipated to exhibit consistency. The reconstruction of the spectral image can be feasible under the given spatial component and spectral responses. Moreover, spatial components should integrate enhanced texture details, while the spectral responses possess smoothness. Hence, the following loss functions are designed for the supervision of intermediate features generated by T_h and T_p:

$$\mathcal{L}_{equal} = \left\|L_{MS} - L_{PAN}\right\|_1, \quad (15)$$

$$\mathcal{L}_{rec} = \sum_{i\,=\,MS,PAN} \sum_{j\,=\,MS,PAN} \left\|R_j \odot L_i - j\right\|_1, \quad (16)$$

$$\mathcal{L}_{sharpen} = \sum_{i\,=\,MS,PAN} \left\|\nabla R_i \cdot \exp(-\lambda \nabla L_i)\right\|, \quad (17)$$

where $\odot$ denotes the Hamilton product, $\mathcal{L}_{equal}$ and $\mathcal{L}_{rec}$ contribute to extracting consistent spatial components, and $\mathcal{L}_{sharpen}$ is utilized to adjust the sharpness of the spatial components. The loss for intrinsic supervision is then defined as:

$$\mathcal{L}_{intrinsic} = \sum_{k=1}^{K} \alpha \mathcal{L}_{sharpen}^k + \beta \mathcal{L}_{rec}^k + \gamma \mathcal{L}_{equal}^k, \quad (18)$$

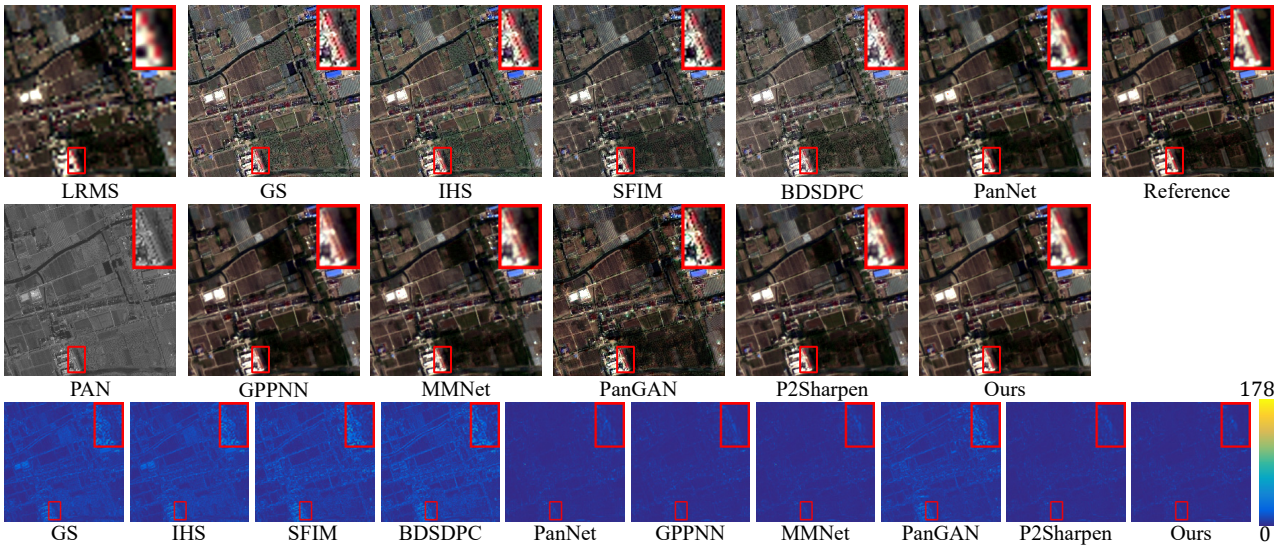

Figure 5: Qualitative results of reduced-scale on the GaoFen-2 datasets. Top group: the fused results. Bottom group: the error between fused results and reference.

Method	GaoFen-2 (Four bands)				QuickBird (Four bands)				WorldView-II (Eight bands)			
	ERGAS↓	RMSE↓	SAM↓	SSIM↑	ERGAS↓	RMSE↓	SAM↓	SSIM↑	ERGAS↓	RMSE↓	SAM↓	SSIM↑
GS	3.6570	14.8546	2.7030	0.6358	2.2270	4.8449	2.6610	0.8964	6.6222	10.0180	8.6770	0.6597
IHS	3.6833	14.8988	2.6302	0.6206	2.5771	5.2654	3.1451	0.8663	7.3297	10.9009	10.3707	0.6137
SFIM	3.9183	15.6493	2.7172	0.6008	3.0894	6.4807	3.0498	0.8503	9.5545	14.2953	9.8337	0.5978
BDSDPC	4.2799	17.0594	3.4726	0.5725	2.5860	6.0181	3.2624	0.8759	7.1908	10.9640	9.1282	0.6435
P2Sharpen	<u>1.3011</u>	<u>5.2200</u>	<u>1.5055</u>	<u>0.9166</u>	<u>1.3302</u>	<u>2.8508</u>	<u>1.6926</u>	<u>0.9546</u>	<u>5.3133</u>	<u>7.8671</u>	6.7204	0.7407
MMNet	1.6224	6.7085	1.6365	0.9050	1.4557	2.9987	1.8101	0.9510	5.3889	7.9310	<u>6.6438</u>	<u>0.7493</u>
PanGAN	3.1412	12.6124	2.5724	0.6728	2.1395	4.6383	2.5596	0.8981	9.2109	14.1030	11.3481	0.4705
PanNet	1.6936	6.9974	1.6174	0.9056	1.8602	4.1136	2.0534	0.9340	7.5865	10.5018	7.8314	0.6855
GPPNN	1.6442	6.8207	1.7363	0.9022	1.4699	3.0603	1.8934	0.9517	5.5859	8.2194	6.6775	0.7354
Ours	**1.2490**	**5.0351**	**1.3907**	**0.9271**	**1.0445**	**2.2080**	**1.3870**	**0.9668**	**5.1130**	**7.8559**	**6.4468**	**0.7512**

Table 1: The quantitative results of reduced scale over three datasets. The best and the second best values are highlighted by bold and underline, respectively. The up or down arrows indicate higher or lower values correspond to better results.

where α, β, and γ are weight parameters, and k indicates the stage number. The final loss function is denoted as:

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

where ρ represents the balancing parameter.

Experiment

Dataset and Benchmark

Extensive experiments are conducted over three satellite datasets, namely GaoFen-2, Quickbird and WorldView-II. The GaoFen-2 and Quickbird satellites capture four spectral bands, while the WorldView-II satellite contains eight bands. For the unavailability of reference image, the Wald's protocol (Wald, Ranchin, and Mangolini 1997) is followed. Specifically, the MS image $H \in \mathbf{R}^{M \times N \times C}$ and PAN image $P \in \mathbf{R}^{rM \times rN \times 1}$ are downsampled with ratio $r = 4$ to generate the reduced MS image $L \in \mathcal{R}^{\frac{M}{r} \times \frac{N}{r} \times C}$ and reduced PAN image $p \in \mathbf{R}^{M \times N \times 1}$. In the training stage, the reduced image pairs are treated as inputs, while H is regarded as a reference. In the testing stage, the reduced image pairs and original image pairs are employed as inputs to evaluate the method for reduced-scale and full-scale, respectively.

Three satellite images are adopted to construct the image datasets over 10000 patches. For each database, PAN images are cropped into patches with the size of 128×128 pixels, while the corresponding MS images are cropped with the size of 32×32 pixels. For numerical stability, each patch is normalized by dividing the maximum value to make the pixels range from 0 to 1. To evaluate the proposed method, several state-of-the-art pan-sharpening methods are selected, including four promising traditional methods (SFIM (Liu 2000), BDSDPC (Vivone 2019), GS (Laben and Brower 2000), IHS (Haydn 1982) and five deep-learning methods (GPPNN (Xu et al. 2021), PanNet (Yang et al. 2017), PanGAN (Ma et al. 2020), MMNet (Yang et al. 2022), P2Sharpen (Zhang et al. 2023)).

Implementation Details

The implementation is based on the Pytorch framework. For optimization, the learning rate is set to 1×10^{-4}. The Adam

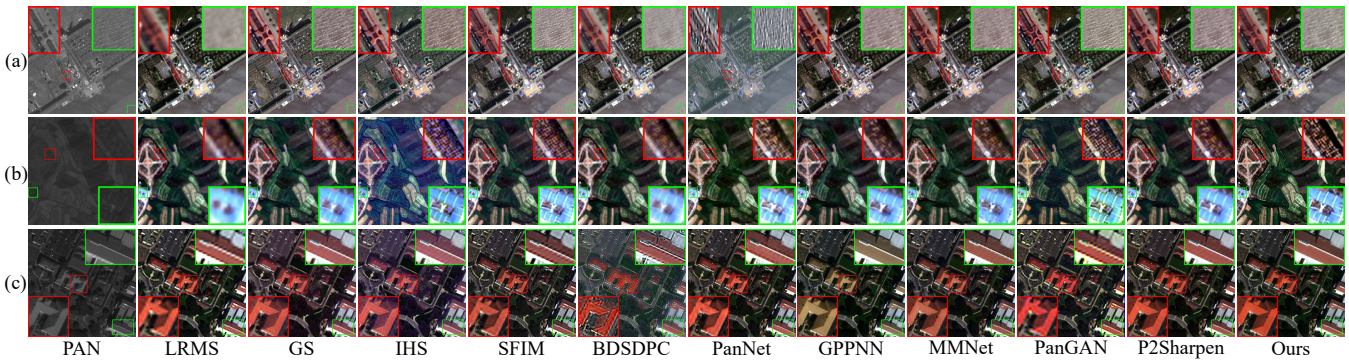

Figure 6: Qualitative results of full-scale over three datasets. (a) GaoFen-2 (b) QuickBird (c) WorldView-II.

Method	GaoFen-2 (4 bands)			QuickBird (4 bands)			WorldView-II (8 bands)		
	QNR↑	D_λ↓	D_s↓	QNR↑	D_λ↓	D_s↓	QNR↑	D_λ↓	D_s↓
GS	0.5849	0.1234	0.3342	0.8363	0.0417	0.1292	0.8654	0.0292	0.1096
IHS	0.5850	0.1294	0.3296	0.7580	0.0737	0.1835	0.8613	0.0443	0.1001
SFIM	0.8018	0.0757	0.1331	0.8538	0.0552	0.0971	0.8960	0.0460	**0.0617**
BDSDPC	0.8553	**0.0270**	0.1210	0.9164	0.0284	0.0573	0.8867	0.0452	0.0726
P2Sharpen	0.8320	0.0618	0.1133	0.9252	0.0276	0.0487	0.8931	0.0283	0.0809
MMNet	0.8574	0.0493	0.0982	0.9149	0.0335	0.0537	0.8843	0.0400	0.0793
PanGAN	0.7813	0.0858	0.1459	0.8555	0.0710	0.0802	0.8513	0.0754	0.0793
PanNet	0.7068	0.1840	0.1321	0.9144	0.0359	0.0517	0.8971	0.0284	0.0768
GPPNN	0.8218	0.0662	0.1202	0.9080	0.0400	0.0545	0.8888	0.0349	0.0792
Ours	**0.9223**	0.0344	**0.0449**	**0.9592**	**0.0166**	**0.0247**	**0.9050**	**0.0274**	0.0695

Table 2: The quantitative results of full-scale over three datasets. The best and the second best values are highlighted by bold and underline, respectively. The up or down arrows indicate higher or lower values correspond to better results.

optimizer is employed with to update the network parameters for 600 epochs with the batch size of 16. The number of unfolding stages is $K = 4$, other coefficients are $\alpha{=}0.1, \beta{=}1, \gamma{=}0.01, \rho{=}0.1$ and $\lambda{=}-10$. All the experiments are conducted on a desktop with 2.6GHz AMD EPYC 7H12, NVIDIA GeForce RTX 3090. For the reduced-scale testing, four widely used assessment metrics are selected including ERGAS (Wald 2002), RMSE, SAM (Alparone et al. 2008) and SSIM (Wang et al. 2004). For the full-scale testing, D_λ, D_s, and QNR (Alparone et al. 2008) are used.

Comparison Experiments

Qualitative Comparison on Reduced-scale The qualitative comparison is provided to verify the performance of our methods, as shown in Figures 5. The top row is the fused results of each method and the bottom row is error maps between reference and fused images. Obviously, traditional methods suffer from the imbalance in spectral and spatial information, causing excessive texture details or apparent spectral distortion. Oppositely, DL-based methods maintain the balance more effectively. In contrast, our method has minor spectral, spatial distortion for preserving reasonable spectral distribution and accurate texture details. Besides, our residual maps are the darkest, which implies our fused images are the most similar with references.

Quantitative Evaluation on Reduced-scale The quantitative results over three datasets are shown in Table 1. As can be seen clearly, our method achieves the best overall performance in all metrics over the satellite datasets, verifying the flexibility and effectiveness of our method are far ahead among other comparative methods both in spectral and spatial assessment. Hence, our method is the most promising.

Qualitative Comparison on Full-scale To further verify the performance of our method, the full-scale qualitative experiments over three datasets are illustrated in Figure 6. It is observed that the four traditional methods exhibit trade-off limitations, sacrificing either the accuracy of spectral information or the richness of spatial information to enhance the other aspect, leading to spatially over-textured results and abnormal alterations in spectral distribution. The other DL-based methods generate severe spectral artifacts and lack sufficient texture details. By comparison, our proposed method achieves better preservation of spectral distribution, and our fused results share the most consistent spatial component with the PAN image.

Quantitative Evaluation on Full-scale To evaluate the quality of the full-scale performance, three non-reference metrics over three datasets are presented in Table 2. Clearly, our approach achieved the best results in the comprehensive evaluation index QNR, while maintaining a good balance between spectral distortion index D_λ and spatial distortion index D_s. It demonstrates that our method exhibits superior generalization capabilities at the full scale. Therefore, our method outperforms others and is the most competitive.

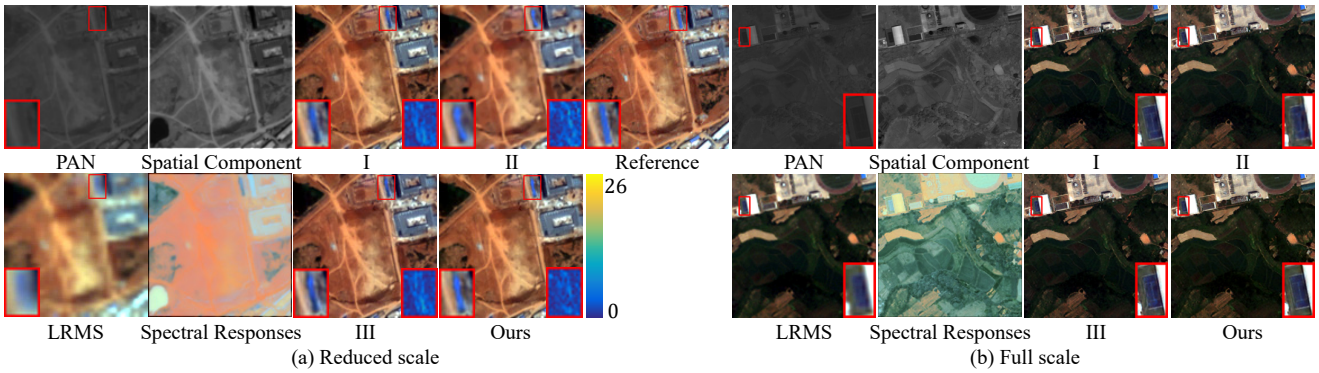

Figure 7: Visual comparisons in the ablation study on the QuickBird datasets.

Config	Ω_1	Ω_2	$\mathcal{L}_{intrinsic}$	ERGAS↓	RMSE↓	SAM↓	SSIM↑	QNR↑	D_λ↓	D_s↓
I	✓	✗	✗	1.1195	2.3286	1.4586	0.9650	0.9605	0.0176	0.0223
II	✗	✓	✗	1.3766	3.0752	1.6722	0.9411	**0.9697**	**0.0156**	**0.0149**
III	✓	✓	✗	1.0865	2.2556	1.4167	0.9664	0.9538	0.0202	0.0267
Ours-IV	✓	✓	✓	**1.0445**	**2.2080**	**1.3870**	**0.9668**	0.9592	0.0166	0.0247

Table 3: The quantitative results of ablation experiments on the QuickBird datasets. The best values are highlighted by bold. "✓" indicates enablement, and "✗" indicates disablement. ↑ and ↓ means positive and negative correlation with image quality.

K	ERGAS↓	RMSE↓	SAM↓	SSIM↑	QNR↑	D_λ↓	D_s↓
1	1.0836	2.2820	1.4473	0.9648	0.9553	0.0187	0.0266
2	1.0705	2.2501	1.4183	0.9658	**0.9605**	0.0161	**0.0238**
3	1.0485	2.2250	1.4006	0.9664	0.9547	0.0175	0.0284
4	1.0445	2.2080	**1.3870**	**0.9668**	0.9592	0.0166	0.0247
5	1.0433	2.2027	1.3991	0.9667	0.9532	0.0175	0.0299
6	**1.0355**	**2.2005**	1.3949	0.9667	0.9595	**0.0153**	0.0257

Table 4: The quantitative results of our method with different number of stages on QuickBird. The best and the second best values are highlighted by bold and underline, respectively.

Ablation Study

The ablation experiments encompass the following three aspects, and the visual comparisons are illustrated in Figure 7.

Effects of the number of stages To investigate the effects of stages number on performance, Table 4 presents the quantitative analysis for different unfolded stages values of K. As K increases from 1 to 4, there is an obvious improvement in the reduced-scale assessments metrics. However, upon further increasing K from 4 to 6, the metrics exhibit varying degrees of fluctuation, albeit maintaining an overarching incremental trend. In terms of non-reference evaluation of full-scale, no substantial disparities are observed in the overall variability of the metrics. As K increases, both training and testing durations correspondingly increase. So we choose $K = 4$ in our implementation to strike a balance between performance and computational complexity.

Influence of Different Priors Two different priors, namely spatial consistency prior (Ω_1) and spectral projection prior (Ω_2) are utilized in the proposed model. We therefore conduct ablation studies to investigate the influence of different priors. As demonstrated in Table 3 (I-III), the best performance of reduced scale is achieved when utilizing both two priors as well as the visual effect. However, the full-scale performance measured by non-reference metrics does not always align with the visual perception. This is because D_λ takes the upsampled LRMS image as the spectral reference and D_s takes the filtered HRPAN image as the spatial reference for each band. Such limited assumptions cause inconsistency in both qualitative and quantitative aspects.

Impact of Intrinsic Loss Functions Another crucial aspect of our ablation study is to examine the influence of intrinsic supervision. The loss $\mathcal{L}_{intrinsic}$ provides supervision to the spatial component in each stage. Through visual comparisons, intrinsic supervision plays a key role in promoting the consistency of the spatial component, resulting in enhanced texture details in the fused results. This observation is further supported by the quantitative results presented in Table 3 (III-IV), where all metrics value exhibit significant improvements after the introduction of intrinsic supervision.

Conclusion

In this paper, we propose an interpretable deep unfolded network with intrinsic supervision for pan-sharpening. We formulate pan-sharpening as the minimization of a variational model with two novel priors, thereby providing guidance for the fusion process. In order to fully exploit the potential of these priors, we introduce intrinsic supervision to orchestrate the optimization of features within the intermediate layers of the network, enhancing pan-sharpening performance. Extensive experiments are conducted to showcase the superior performance of our proposed method compared to other state-of-the-art methods, both qualitatively and quantitatively. In the future, we plan to explore the versatility of our method in various remote sensing applications.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62276192).

References

Alparone, L.; Aiazzi, B.; Baronti, S.; Garzelli, A.; Nencini, F.; and Selva, M. 2008. Multispectral and Panchromatic Data Fusion Assessment Without Reference. *Photogrammetric Eng. Remote Sens.*, 74(2): 193–200.

Alparone, L.; Baronti, S.; Aiazzi, B.; and Garzelli, A. 2016. Spatial Methods for Multispectral Pansharpening: Multiresolution Analysis Demystified. *IEEE Trans. Geosci. Remote Sens.*, 54(5): 2563–2576.

Cao, X.; Fu, X.; Hong, D.; Xu, Z.; and Meng, D. 2021. PanCSC-Net: A Model-driven Deep Unfolding Method for Pansharpening. *IEEE Trans. Geosci. Remote Sens.*, 60: 1–13.

Casagli, N.; Intrieri, E.; Tofani, V.; Gigli, G.; and Raspini, F. 2023. Landslide Detection, Monitoring and Prediction with Remote-sensing Techniques. *Nature Reviews Earth and Environment*, 4(1): 51–64.

Ghahremani, M.; and Ghassemian, H. 2016. Nonlinear IHS: A Promising Method for Pan-sharpening. *IEEE Geosci. Remote Sens. Lett.*, 13(11): 1606–1610.

Haydn, R. 1982. Application of the IHS Color Transform to the Processing of Multisensor Data and Image Enhancement. In *Proc. Int. Symp. Remote Sens. Arid and Semi-Arid Lands*.

Javan, F. D.; Samadzadegan, F.; Mehravar, S.; Toosi, A.; Khatami, R.; and Stein, A. 2021. A Review of Image Fusion Techniques for Pan-sharpening of High-resolution Satellite Imagery. *ISPRS J. Photogramm. Remote Sens.*, 171: 101–117.

Jin, X.; Gu, Y.; and Xie, W. 2022. Intrinsic Hyperspectral Image Decomposition With DSM Cues. *IEEE Trans. Geosci. Remote Sens.*, 60: 1–13.

Laben, C. A.; and Brower, B. V. 2000. Process for Enhancing the Spatial Resolution of Multispectral Imagery Using Pan-sharpening. US Patent 6,011,875.

Liu, J. 2000. Smoothing Filter-based Intensity Modulation: A Spectral Preserve Image Fusion Technique for Improving Spatial Details. *Int. J. Remote Sens.*, 21(18): 3461–3472.

Lu, H.; Yang, Y.; Huang, S.; Tu, W.; and Wan, W. 2022. A Unified Pansharpening Model Based on Band-adaptive Gradient and Detail Correction. *IEEE Trans. Image Process.*, 31: 918–933.

Ma, J.; Yu, W.; Chen, C.; Liang, P.; Guo, X.; and Jiang, J. 2020. Pan-GAN: An Unsupervised Pan-sharpening Method for Remote Sensing Image Fusion. *Inf. Fusion*, 62: 110–120.

Meng, X.; Wang, N.; Shao, F.; and Li, S. 2022. Vision Transformer for Pansharpening. *IEEE Trans. Geosci. Remote Sens.*, 60: 1–11.

Ruder, S. 2016. An Overview of Gradient Descent Optimization Algorithms. *arXiv preprint arXiv:1609.04747*.

Sun, Y.; Yang, Y.; Liu, Q.; Chen, J.; Yuan, X.-T.; and Guo, G. 2020. Learning Non-locally Regularized Compressed Sensing Network with Half-quadratic Splitting. *IEEE Trans. Multimedia*, 22(12): 3236–3248.

Vicinanza, M. R.; Restaino, R.; Vivone, G.; Dalla Mura, M.; and Chanussot, J. 2014. A Pansharpening Method Based on the Sparse Representation of Injected Details. *IEEE Geosci. Remote Sens. Lett.*, 12(1): 180–184.

Vivone, G. 2019. Robust Band-dependent Spatial-detail Approaches for Panchromatic sharpening. *IEEE Trans. Geosci. Remote Sens.*, 57(9): 6421–6433.

Vivone, G.; Dalla Mura, M.; Garzelli, A.; Restaino, R.; Scarpa, G.; Ulfarsson, M. O.; Alparone, L.; and Chanussot, J. 2020. 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.

Wald, L. 2002. *Data Fusion: Definitions and Architectures: Fusion of images of different spatial resolutions*. Presses des MINES.

Wald, L.; Ranchin, T.; and Mangolini, M. 1997. Fusion of Satellite Images of Different Spatial Resolutions: Assessing the Quality of Resulting Images. *Photogrammetric Eng. Remote Sens.*, 63(6): 691–699.

Wang, X.; Hu, Q.; Cheng, Y.; and Ma, J. 2023. Hyperspectral Image Super-Resolution Meets Deep Learning: A Survey and Perspective. *IEEE/CAA J. Autom. Sinica*, 10(8): 1664–1687.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Image Process.*, 13(4): 600–612.

Xu, S.; Zhang, J.; Zhao, Z.; Sun, K.; Liu, J.; and Zhang, C. 2021. Deep Gradient Projection Networks for Pansharpening. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1366–1375.

Yang, G.; Zhou, M.; Yan, K.; Liu, A.; Fu, X.; and Wang, F. 2022. Memory-augmented Deep Conditional Unfolding Network for Pan-sharpening. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1788–1797.

Yang, J.; Fu, X.; Hu, Y.; Huang, Y.; Ding, X.; and Paisley, J. 2017. PanNet: A Deep Network Architecture for Pansharpening. In *Proc. IEEE Int. Conf. Comput. Vis.*, 5449–5457.

Zhang, B.; Wu, Y.; Zhao, B.; Chanussot, J.; Hong, D.; Yao, J.; and Gao, L. 2022. Progress and Challenges in Intelligent Remote Sensing Satellite Systems. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, 15: 1814–1822.

Zhang, H.; Wang, H.; Tian, X.; and Ma, J. 2023. P2Sharpen: A Progressive Pansharpening Network with Deep Spectral Transformation. *Inf. Fusion*, 91: 103–122.

Zhou, H.; Liu, Q.; Weng, D.; and Wang, Y. 2022. Unsupervised Cycle-consistent Generative Adversarial Networks for Pan-sharpening. *IEEE Trans. Geosci. Remote Sens.*, 60: 1–14.