Deep Unfolding Network with Spatial-spectral Perception Enhanced for Pan-sharpening

Mengjiao Zhao[1,*]
mj.zhao@zju.edu.cn
Mengting Ma[2,*]
mtma@zju.edu.cn
Xiangdong Li[1]
xiangdong.li@zju.edu.cn
Ao Gao[1]
gaoao.olivia@zju.edu.cn
Siyang Song[3]
ss2796@cam.ac.uk
Wei Zhang[1,4,†]
cstzhangwei@zju.edu.cn

[1] School of Software Technology, Zhejiang University, Hangzhou, China
[2] College of Computer Science and Technology, Zhejiang University, Hangzhou, China
[3] School of Computing and Mathematical Sciences, University of Leicester, UK
[4] Innovation Center of Yangtze River Delta, Zhejiang University, Jiaxing, China

Abstract

Pan-sharpening aims to perform super-resolution processing on low-resolution multispectral (LR-MS) images guided by high-resolution panchromatic (PAN) images. Existing pan-sharpening methods have two main issues. Firstly, deep learning-based methods are mostly designed based on black-box principles, lacking sufficient interpretability. Secondly, model-based methods enhance interpretability but do not fully consider domain-specific prior knowledge, namely complex spatial and spectral relationships, which limits their performance. To address these challenges, we propose a novel deep unfolding network with spatial-spectral perception enhanced for pan-sharpening, namely SSPEDUN. Specifically, we model the pan-sharpening problem as the minimization of a variational model with spatial reconstruction priors and spectral modulation priors. The spatial reconstruction prior reconstructs high-quality spatial information based on observed image spatial relationships, while the spectral modulation prior accurately modulates the spectral relationships between images. Then, we design an efficient iterative proximal gradient descent algorithm to alternately solve the data subproblem and the prior subproblem of the model, and then unfold this algorithm into a deep network. In the deep unfolding network, we introduce a data projection module to address data mapping during the optimization process and carefully design a Perception Enhancement Module (PEM) as the prior module to precisely model spatial and spectral relationships. Extensive experiments on three satellite datasets demonstrate the superiority of our method. The source code is available in our supplementary material.

*These authors contribute equally.
†Corresponding author

1 Introduction

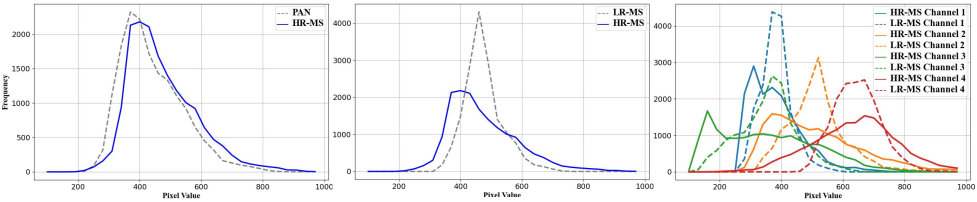
Recent advances in remote sensing technique boosts the demand for high-resolution multispectral (HR-MS) images, which have been widely applied in areas such as environmental monitoring and agricultural development [24, 26]. However, the physical limitations in existing satellite systems make it challenging to simultaneously achieve images of both high spatial and spectral resolutions with equipped sensors [23]. Consequently, pan-sharpening becomes an alternative solution to fuse high-resolution panchromatic (PAN) images with their corresponding low-resolution multispectral (LR-MS) images to generate the target HR-MS images with both high spatial and spectral resolution.

In the past decades, there has been an explosive growth of pan-sharpening solutions, with a focus on both traditional handcrafted methods and deep learning (DL)-based methods. Traditional handcrafted methods include Component Substitution (CS), Multi-resolution Analysis (MRA), and Variational Optimization (VO). However, these methods mostly rely on manually defined priors and constraints, with limited capability in representing image features, thus limiting their performance [12]. Inspired by the recent success of DL models in visual tasks, various DL-based pan-sharpening methods are proposed that can be divided into two types, i.e., single-branch-based and dual-branch-based, according to the reconstruction way of spatial-spectral information. The single-branch-based methods involve directly concatenating LR-MS images and PAN images, then applying CNN networks to extract spatial and spectral information to reconstruct HR-MS images [13, 21]. Meanwhile, the dual-branch-based methods utilize CNN networks to separately extract the spectral information from LR-MS images and the spatial information from PAN images, which are fused to generate HR-MS images [2, 29]. Although these methods demonstrate superiority in spatial-spectral information reconstruction, they mostly construct network typologies in a black-box manner, without considering the interpretability of the model.

Therefore, to improve interpretability, model-based pan-sharpening methods are proposed, where deep unfolding networks (DUNs) have been widely applied. However, current model-based methods still have flaws as they either only model the degradation process of HR-MS images without considering the complex spatial and spectral relationships between images (e.g., GPPNN [18]), or only consider a macroscopic perspective without detailed analysis from a spatial or spectral reconstruction viewpoint, leading to poor performance (e.g., MMNet [19]). In conclusion, although existing DUNs enhance the interpretability of pan-sharpening tasks, they neither fully consider domain-specific priors, i.e., complex spatial and spectral relationships, nor conduct cross-modal interactions, and thus limit their performances.

Motivation. We conduct a detailed analysis of the spatial and spectral relationships between PAN images, LR-MS images, and HR-MS (GT) images. As illustrated in Fig. 1, we find that the spatial information in PAN images is highly similar to that in HR-MS images, making it possible to restore the spatial details of HR-MS images by only extracting spatial information from PAN images. Additionally, the spectral trends of LR-MS images do not correspond to the bands in their corresponding HR-MS images, indicating that it is necessary to modulate the spectral information of LR-MS image with the spectral information from the PAN image to reconstruct the spectral information of their target HR-MS image.

In this paper, we propose a novel deep unfolding network with spatial-spectral perception enhanced for pan-sharpening, namely **SSPEDUN**. Unlike previous DUNs [10, 18, 19], our SSPEDUN deeply integrates domain-specific priors. Specifically, we model the degradation process of HR-MS images as a variational model optimization problem with spatial



(a) HR-MS vs. PAN (b) HR-MS vs. LR-MS (c) HR-MS channels vs. LR-MS channels

Figure 1: (a) Comparison of pixel distributions between HR-MS images and PAN images. (b) Comparison of pixel distributions between HR-MS images and LR-MS images. (c) Comparison of pixel distributions between HR-MS images and LR-MS images for all channels.

reconstruction prior and spectral modulation prior. The spatial reconstruction prior models the spatial correlation between the target HR-MS image and the input PAN image based on their spatial information relationship (as shown in Fig. 1 (a)). The spectral modulation prior modulates the spectral mapping relationship between the target HR-MS image, the input PAN and LR-MS images from the spectral perspective, aiming to achieve better spectral reconstruction results. Then, we design an efficient iterative proximal gradient descent algorithm to alternately solve the data subproblem and prior subproblem of the model. We unfold the proposed algorithm into the deep unfolding network, i.e., SSPEDUN, where each stage corresponds to an iteration. In the proposed SSPEDUN, we design a data projection module to realize the linear projection of inputs in each stage. Additionally, we design a Perception Enhancement Module (PEM) as the prior module. Unlike previous rudimentary prior modules, our PEM, based on two specific priors we propose, is capable of extracting high-quality spatial information of PAN images in the spatial domain and finely modulating the spectral relationships in the inputs in the frequency domain.

The main contributions of our work are summarized as follows:

- We model the pan-sharpening problem as the minimization of a variational model and introduce the spatial reconstruction prior and the spectral modulation prior based on practical analysis. In this way, we enhance the spatial and spectral quality of reconstructed HR-MS images.
- We devise a novel interpretable deep unfolding network with spatial-spectral perception enhanced, namely SSPEDUN. Within this framework, we tailor the PEM which simultaneously reconstructs high-quality spatial information and modulates the spectrum, improving the performance of pan-sharpening.
- Extensive experiments on various satellite datasets demonstrate that our method outperforms other state-of-the-art methods qualitatively and quantitatively.

2 Related Work

2.1 Pan-sharpening Methods

Traditional pan-sharpening methods can be classified into Component Substitution (CS), Multi-resolution Analysis (MRA), and Variational Optimization (VO) methods. CS methods [3, 5, 6] extract specific components from LR-MS and PAN images using reversible projection algorithms, and then replace or merge these components to restore the spectrally enhanced image. MRA methods [11, 14] inject spatial information extracted from PAN images into LR-MS images using multi-resolution decomposition techniques to enhance spatial textures. VO methods [15, 17] reformulate pan-sharpening as a variational optimization problem. However, due to their limited representation capacity, the spectral sharpening results of traditional methods often exhibit adverse spatial or spectral distortions. Recently, DL-based methods have been employed for pan-sharpening, significantly improving performance compared to traditional methods. According to the spatial-spectral information reconstruction approach, DL-based methods can be divided into two categories, i.e., single-branch-based and dual-branch-based. Single-branch-based methods directly concatenates LR-MS images and PAN images, then uses CNN-based networks to extract spatial and spectral information to reconstruct HR-MS images, e.g., PNN [13] and PANNet [21]. Another type utilizes CNN-based networks to separately extract spectral information from LR-MS images and spatial information from PAN images, finally fusing them to generate HR-MS images, such as SRPPNN [2], CTINN [28] and MutInf [29]. Although DL-based methods have made progress in spatial-spectral information reconstruction, they construct deep learning networks in a black-box manner, lacking interpretability.

2.2 Deep Unfolding Network

In recent years, many model-based methods have been proposed to enhance the interpretability of networks, with deep unfolding networks (DUNs) widely applied in model-based methods. Generally, DUNs unfold their optimization algorithms for the problem at hand and parameterize unfolding models for end-to-end training [20]. In pan-sharpening, GPPNN [18] applies DUNs for the first time. It models pan-sharpening as the spectral degradation process of LR-MS images and the spatial degradation process of PAN images, introducing two image priors to capture spatial and spectral information of PAN and LR-MS images. However, it only models the degradation process and does not consider the complex spatial and spectral relationships between images, thus limiting its performance. MMNet [19] and LGTEUN [10] introduce local priors and global priors during modeling to capture global and local information. However, they only consider the macroscopic perspective and do not conduct a detailed analysis from the spatial or spectral reconstruction viewpoint, resulting in poor performance.

3 Method

3.1 Model Formulation

In this paper, we model the pan-sharpening problem as a super-resolution problem of the LR-MS image guided by the PAN image. The degradation process of the HR-MS image $H \in \mathbb{R}^{H\times W\times C}$ can be expressed as:

$$L = DBH + N_L, \quad P = SH + N_P \tag{1}$$

where D and B represent the blur and downsampling operators respectively; S denotes the spectral response function of the panchromatic imaging sensor, and N_L and N_P represent the noise introduced during the capture process of the LR-MS image $L \in \mathbb{R}^{h\times w\times C}$ and the PAN image $P \in \mathbb{R}^{H\times W\times 1}$, respectively. Based on this degradation model, we introduce the image

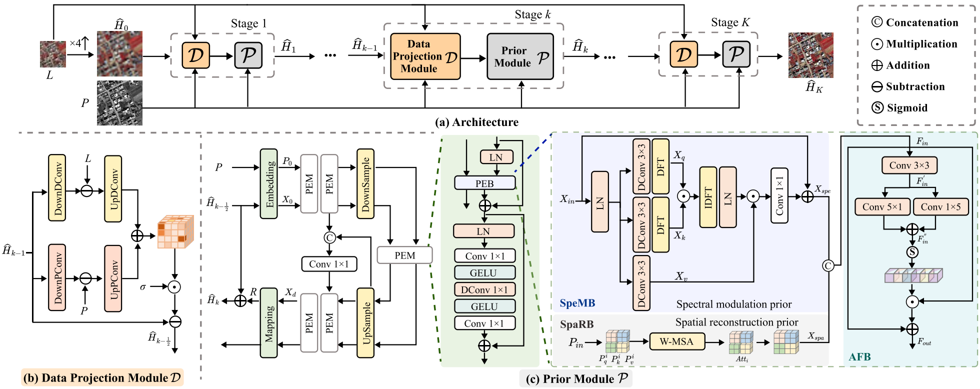


Figure 2: Illustration of our method. (a) Overall architecture of SSPEDUN. (b) Data Projection module $\mathcal{D}$. (c) Structure and components of the prior module $\mathcal{P}$.

space reconstruction prior $\Omega_P(P)$ and the spectral modulation prior $\Omega_L(H,P)$ according to our analysis in Section 1 to reconstruct high-quality spatial and spectral information. Then, the H can be obtained by solving the following minimization problem,

$$\widehat{H} = \underset{H}{argmin}\frac{1}{2}||L-DBH||^2 + \frac{1}{2}||P-SH||^2 + \lambda_1\Omega_P(P) + \lambda_2\Omega_L(H,P), \tag{2}$$

where $\frac{1}{2}||L-DBH||^2$ and $\frac{1}{2}||P-SH||^2$ are data fidelity terms and λ_1 and λ_2 are hyperparameters balancing their importance. We solve Eq. 2 as an iterative convergence problem using the proximal gradient descent (PGD) algorithm, i.e.,

$$\widehat{H}_k = \underset{H}{argmin}\frac{1}{2}||H-(\widehat{H}_{k-1}-\sigma\nabla f(\widehat{H}_{k-1})||^2 + \lambda_1\Omega_P(P) + \lambda_2\Omega_L(H,P), \tag{3}$$

where $\widehat{H}_k$ represents the output of the k-th iteration, σ denotes the step size for updates, and $\nabla f(\widehat{H}_{k-1})$ can be expressed in the following form,

$$\nabla f(\widehat{H}_{k-1}) = (DB)^T(DB\widehat{H}_{k-1}-L) + (S\widehat{H}_{k-1}-P)S^T. \tag{4}$$

We divide Eq. 3 into a data sub-problem (Eq. 5) and a prior sub-problem (Eq. 6), and solve them alternately,

$$\widehat{H}_{k-\frac{1}{2}} = \widehat{H}_{k-1} - \sigma\nabla f(\widehat{H}_{k-1}), \tag{5}$$

$$\widehat{H}_k = prox_{\Omega_L}(\widehat{H}_{k-\frac{1}{2}}) + prox_{\Omega_P}(P), \tag{6}$$

where $prox_{\Omega_L}$ and $prox_{\Omega_P}$ represent the proximal operators of the priors $\Omega_L(.)$ and $\Omega_P(.)$, respectively. Next, we use this iterative process to design our deep unfolding network, where we generalize the two iterative steps, Eq. 5 and Eq. 6, into network modules, namely the data projection module $\mathcal{D}$ and the prior module $\mathcal{P}$.

3.2 Deep Unfolding Network

Our deep unfolding network SSPEDUN is illustrated in Fig. 2 (a). First, it performs fourfold upsampling on the given LR-MS image $L \in \mathbb{R}^{h\times w\times C}$ to obtain the initialized input

$\widehat{H}_0 \in \mathbb{R}^{H\times W\times C}$. Subsequently, the $\widehat{H}_0$ and the given PAN image $P \in \mathbb{R}^{H\times W\times 1}$ are jointly fed to the network, where they undergo K stages of processing to reconstruct the required spatial and spectral information. These stages are intentionally designed to correspond to the K iterations in the optimization algorithm, where each stage comprises a data projection module $\mathcal{D}$ and a customized prior module $\mathcal{P}$ for spatial-spectral reconstruction. Finally, the reconstructed HR-MS image $\widehat{H}_K \in \mathbb{R}^{H\times W\times C}$ is obtained after K stages.

3.2.1 Data Projection Module $\mathcal{D}$

We design a data projection module $\mathcal{D}$ to simulate the linear mapping process in Eq. 5. As shown in Fig. 2 (b) which details the k-th iteration, the $\mathcal{D}$ takes the output of the $k-1$-th stage $\widehat{H}_{k-1}$, L, and P as inputs. The processing flow of $\mathcal{D}$ is as follows,

$$\widehat{H}_{k-\frac{1}{2}} = \widehat{H}_{k-1} - \sigma(DConv_{\uparrow}(DConv_{\downarrow}(\widehat{H}_{k-1}) - L) + PConv_{\uparrow}(PConv_{\downarrow}(\widehat{H}_{k-1}) - P)) \quad (7)$$

where $DConv_{\uparrow}$ and $DConv_{\downarrow}$ respectively represent upsampling and downsampling achieved by 3×3 depth convolutions. Similarly, $PConv_{\downarrow}$ and $PConv_{\uparrow}$ are implemented by point convolutions with the purpose of reducing the number of channels from C to 1 and increasing the number of channels from 1 to C.

3.2.2 Prior Module $\mathcal{P}$

When designing image denoising priors, previous deep unfolding network-based methods directly extract features from the input PAN image and LR-MS image, which neglect crucial domain knowledge and thus failed to comprehensively account for modal characteristics of LR-MS and PAN images as well as their complex spatial and spectral relationships. To address this issue, we introduce the spatial reconstruction prior and the spectral modulation prior when designing the prior module $\mathcal{P}$ to reconstruct the spatial and spectral properties of HR-MS images.

As shown in Fig. 2 (c), the $\mathcal{P}$ takes the PAN image P and the output $\widehat{H}_{k-\frac{1}{2}}$ of $\mathcal{D}$ as inputs. Specifically, it first utilizes embedding layers to map $\widehat{H}_{k-\frac{1}{2}}$ and P to features X_0 and P_0 respectively. Then, X_0 and P_0 are embedded into deep feature X_d through an encoder, bottleneck, and decoder. The encoder and decoder each contain two PEMs and a resized deep convolution module, and the bottleneck has a single PEM. Each PEM consists of two layer normalizations (LNs), a Perception Enhancement Block (PEB), and a feed-forward network composed of convolutional layers. Specifically, the PEB consists of a Spatial Reconstruction Block (SpaRB), a Spectral Modulation Block (SpeMB), and an Adaptive Fusion Block (AFB), as shown in Fig. 2 (c). SpaRB utilizes a customized self-attention operation based on local windowing in the spatial domain to extract high-quality spatial details, while SpeMB refines modulated spectral information using element-wise multiplication in the frequency domain. Then, the outputs of the two blocks are concatenated and inputted into AFB for adaptive fusion. After a series of processing steps, X_d is finally mapped to R by a convolutional layer. The output $\widehat{H}_k$ is obtained by adding $\widehat{H}_{k-\frac{1}{2}}$ and the reshaped R. Next, we will provide a detailed introduction to SpaRB, SpeMB and AFB.

Spatial Reconstruction Block. Considering the spatial reconstruction prior proposed in Section 3.1, we design the SpaRB, as illustrated in Fig. 2 (c). It takes $P_{in} \in \mathbb{R}^{H\times W\times C}$ as the input. It first divides P_{in} into non-overlapping windows of size $M\times M$, which are then reshaped into $\frac{HW}{M^2}\times M^2\times \frac{C}{2}$. Subsequently, SpaRB generates query P_q, key P_k, and value P_v through linear projection, and splits them into h heads along the channel dimension: $P_q = [P_q^1, P_q^2, \cdots, P_q^h]$, $P_k = [P_k^1, P_k^2, \cdots, P_k^h]$, and $P_v = [P_v^1, P_v^2, \cdots, P_v^h]$ where the dimension of each head is $d_h = \frac{C}{2h}$. Fig. 2 (c) only shows the scenario where $h = 1$ for simplicity. The computation of each local self-attention Att_i proceeds as,

$$\mathrm{Att}_i = softmax(\frac{P_q^i (P_k^i)^T}{\sqrt{d_h}} + \mathrm{Pos}_i) P_v^i, \quad i = 1, 2, \cdots, h, \tag{8}$$

where Pos_i is a learnable parameter used to embed positional information. Finally, we concatenate the heads along the channel dimension and merge the windows to obtain the output X_{spa}.

Spectral Modulation Block. Based on the spectral modulation prior, we design the SpeMB. As shown in Fig. 2 (c), it takes $X_{in} \in \mathbb{R}^{H \times W \times C}$ as input, which first passes through a LN. Then, it utilizes three 3×3 depth convolutions to generate feature embeddings X_q, X_k, and X_v. To better modulate the relationship between spectra, we transform X_q and X_k into the frequency domain through the Discrete Fourier Transform (DFT). Additionally, according to the convolution theorem, the convolution of two signals in the spatial domain is equivalent to their element-wise multiplication in the frequency domain. Therefore, we can obtain the weight matrices X_q and X_k in the frequency domain through element-wise multiplication with low computational cost, namely,

$$W = \mathcal{F}(X_q) \odot \mathcal{F}(X_k), \quad X_{att} = LN(\mathcal{F}^{-1}(W)) \odot X_v, \tag{9}$$

where $\mathcal{F}(.)$ represents the DFT, $\mathcal{F}^{-1}(.)$ represents the Inverse DFT (IDFT), and $\odot$ denotes element-wise multiplication. Then, we obtain refined modulation of the reconstructed spectral information X_{spe} through residual connections.

Adaptive Fusion Block. We propose the AFB to better integrate spatial and spectral information. As illustrated in Fig. 2 (c), we first concatenate the outputs of SpaRB and SpeMB along the channel dimension, i.e., F_{in}, and compress it into a lower-dimensional embedding space $F_{in}^{'} \in \mathbb{R}^{2 \times H \times W}$ through a 3×3 convolutional layer. Then, we extend the receptive field of the embedding space along the vertical and horizontal directions using 5×1 and 1×5 convolutional layers, allowing AFB to focus on learning complementary features with fewer costs. Subsequently, we aggregate the embedding spaces from both directions using element-wise addition to obtain gating controls. Finally, we derive high-quality fusion features F_{out} through the gating controls,

$$F_{in}^{''} = Conv_{5\times1}(F_{in}^{'}) + Conv_{1\times5}(F_{in}^{'}), \quad F_{out} = sigmoid(F_{in}^{''}) \odot F_{in} + F_{in}. \tag{10}$$

4 Experiments

4.1 Datasets and Evaluation Metrics

We conduct extensive experiments on three datasets from the satellites GaoFen-2, WorldView-2, and WorldView-3 to validate the effectiveness of the proposed method. Following Wald's protocol [16], we employ downsampling operations to generate reduced-resolution datasets for each satellite sensor. For each dataset, we create training and testing datasets in a ratio of 70% to 30%.

To assess the performance of our model, we select five evaluation metrics, including peak signal-to-noise ratio (PSNR), structural similarity (SSIM), Q4, spectral angle mapper

Methods	Params(M)	Flops(G)	GaoFen-2					WorldView-2					WorldView-3				
			PSNR ↑	SSIM ↑	Q4 ↑	SAM ↓	ERGAS ↓	PSNR ↑	SSIM ↑	Q4 ↑	SAM ↓	ERGAS ↓	PSNR ↑	SSIM ↑	Q4 ↑	SAM ↓	ERGAS ↓
SFIM [■]	–	–	34.7715	0.8572	0.4584	0.0657	4.2073	32.6334	0.8728	0.5159	0.0597	3.1919	21.4154	0.5415	0.4525	0.1147	8.8553
BICUBIC [■]	–	–	35.3075	0.8514	0.3630	0.0597	4.0382	32.2961	0.8238	0.328	0.0552	3.5407	20.5048	0.3572	0.2355	0.1157	9.8285
Wavelet [■]	–	–	33.9208	0.8197	0.4033	0.0695	4.6445	32.1992	0.8500	0.4577	0.0638	3.3799	21.4464	0.5656	0.5271	0.1503	9.1545
IHS [■]	–	–	35.2315	0.8837	0.5217	0.0661	3.9912	32.8250	0.8775	0.5305	0.0637	3.0585	22.3452	0.6133	0.5714	0.5714	7.9444
SRPPNN [■]	0.8977	21.1069	45.5621	0.9824	0.8927	0.0256	1.2327	42.4197	0.9777	0.8366	0.0214	0.9567	31.5376	0.9474	0.9409	0.0674	2.8496
MSDCNN [■]	0.2390	3.9112	43.9541	0.9768	0.8640	0.0303	1.4720	41.0747	0.9725	0.8148	0.025	1.0915	32.0549	0.9531	0.9473	0.0631	2.6779
GPPNN [■]	0.1198	1.3967	43.5980	0.9764	0.8663	0.0326	1.5218	40.5086	0.9698	0.8009	0.0275	1.206	31.698	0.9508	0.9445	0.0653	2.8112
MutInf [■]	0.1855	2.46	44.8303	0.9800	0.8835	0.0277	1.3394	41.9530	0.9760	0.8259	0.0227	1.0152	31.8294	0.9522	0.9468	0.0636	2.7528
HSIT [■]	42.8274	54.9373	44.2385	0.9717	0.8642	0.0310	1.4743	41.6025	0.9735	0.8232	0.0239	1.0434	31.1346	0.9510	0.9468	0.071	2.965
Panformer [■]	1.5251	2.9426	45.1304	0.9811	0.8859	0.0263	1.2929	41.5788	0.9739	0.8245	0.0237	1.0502	31.1338	0.9437	0.938	0.067	2.9619
MSDDN [■]	0.6701	3.1521	45.7893	0.9836	0.8949	0.025	1.1924	42.1967	0.9768	0.8339	0.0221	0.9766	32.0545	0.9535	0.9483	0.063	2.6826
MDCUN [■]	0.0984	118.2970	45.4785	0.9822	0.8913	0.0255	1.2354	42.3217	0.9773	0.8373	0.0217	0.9618	31.9930	0.9532	0.9481	0.0630	2.7030
LGTEUN [■]	0.2022	1.2845	45.9539	0.9840	0.8990	0.0246	1.1712	42.5427	0.9781	0.8384	0.0211	0.9407	32.0793	0.9536	0.9488	0.0602	2.6629
WINet [■]	2.1736	7.7704	45.9538	0.9842	0.8987	0.0244	1.1676	42.4081	0.9778	0.8371	0.0215	0.9570	32.2596	0.9550	0.9497	0.0618	2.6244
Ours	0.1712	1.8275	**46.5204**	**0.9858**	**0.9055**	**0.0229**	**1.0930**	**42.7301**	**0.9787**	**0.8427**	**0.0206**	**0.9213**	**32.3005**	**0.9555**	**0.9505**	**0.0598**	**2.6047**

Table 1: The experimental results of all competing methods on three benchmark datasets.The best and second best values are highlighted in **bold** and underline, respectively.

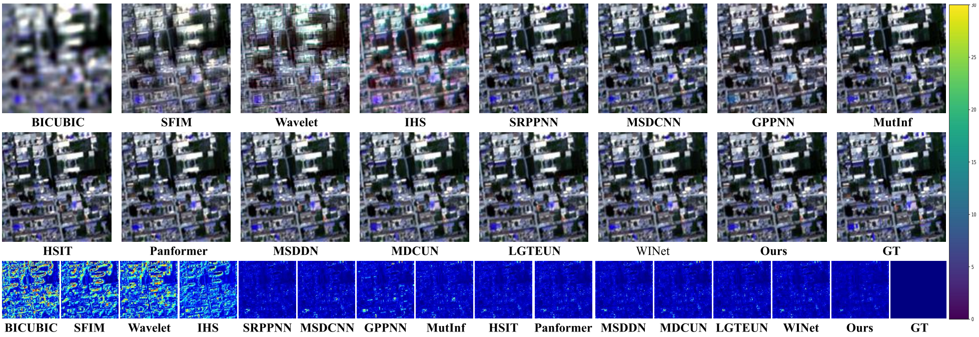


Figure 3: Visual comparison and absolute errors of our method versus other representative pan-sharpening methods on the WorldView-2 dataset.

(SAM), and relative dimensionless global error in synthesis (ERGAS). Additionally, to further compare the generalization capability of the model, we employ three non-reference metrics to evaluate its performance, including spectral distortion index (D_λ), spatial distortion index (D_s), and quality without reference (QNR).

4.2 Implementation Details

During training, our SSPEDUN is supervised by the L_1 loss between the output $\widehat{H}_k$ and the GT image. We set the number of stages K to 2. The data projection module $\mathcal{D}$ shares parameters across stages, while the prior module $\mathcal{P}$ does not share parameters. In each SpaRB, we set the size of the local window M to 8 and the number of heads h to 2. We train on three satellite datasets for 500 epochs using the Adam optimizer with fixed hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The batch size is set to 16. The learning rate is initialized to 1.5×10^{-3} and decayed by a factor of 0.85 every 100 epochs for effective convergence. All experiments are conducted using a single NVIDIA RTX A5000 GPU within the PyTorch framework.

4.3 Comparison with SOTA methods

To validate the effectiveness of our method in the pan-sharpening task, we conduct extensive experiments on benchmark datasets, comparing it with four traditional methods, i.e.,

Metric	SRPPNN	MSDCNN	GPPNN	MutInf	HSIT	Panformer	MSDDN	MDCUN	LGTEUN	WINet	Ours
$D_\lambda \downarrow$	0.0640	0.0639	0.0670	0.0638	0.0640	0.0627	0.0639	0.0661	0.0645	0.0644	**0.0622**
$D_s \downarrow$	0.0858	0.0783	0.0785	0.0794	0.0862	0.0825	0.0758	0.0834	0.0771	0.0761	**0.0757**
QNR↑	0.8567	0.8637	0.8607	0.8644	0.8557	0.8609	**0.8659**	0.8568	0.8644	0.8627	0.8652

Table 2: Quantitative results at full resolution on the WorldView-2 dataset.

SFIM [11], BICUBIC [8], Wavelet [9], and IHS [4], and ten DL-based methods, including SRPPNN [2], MSDCNN [22], GPPNN [18], MutInf [29], HSIT [1], Panformer [27], MSDDN [7], MDCUN [20], LGTEUN [10], and WINet [25]. More results regarding the following experiments can be found in the supplementary materials.

Quantitative Comparison. The quantitative comparison results of our method against the aforementioned competitive methods on the three datasets are presented in Tab. 1 . It is evident that DL-based methods surpass traditional model-based methods, and our method outperforms other comparison methods in terms of all evaluation metrics across the three datasets, demonstrating significant performance improvement. Specifically, on the GaoFen-2, WorldView-2, and WorldView-3 datasets, our SSPEDUN achieves improvements of 0.5665 dB, 0.1847 dB, and 0.0409 dB in PSNR, respectively, compared to the second-best method. It is worth noting that our model achieves a favorable balance with fewer parameters and computational requirements, and it outperforms other methods in terms of performance.

Qualitative Comparison. In Fig. 3, we present the qualitative results using typical samples from the WorldView-2 dataset. The top two rows of images depict the results of each method for pan-sharpening, while the bottom row of images illustrates the residual of mean squared error between the pan-sharpening results and GT images. Compared to other competing methods, our method exhibits smaller spectral and spatial distortions while preserving reasonable spectral distributions and accurate texture details. This visually pleasing effect further substantiates the superiority of our method.

Effect on full-resolution scenes. To further evaluate the performance of the model in full-resolution scenarios, we test our method against DL-based methods on the full-resolution WorldView-2 dataset. According to the Tab. 2, the proposed method achieves competitive results in terms of D_λ and D_s, and ranks second in terms of QNR. This indicates that our method exhibits superior generalization capability in full-resolution scenarios.

4.4 Ablation Study

Effects of the number of stages. We present in Tab. 3 the performance of our model under different numbers of stages to investigate the influence of stage count on model performance. As K increases from 1 to 2, the performance of our model reaches its peak. As K continues to increase, the model's performance shows a decreasing trend. In this paper, we balance the performance and computational complexity of our method by setting $K = 2$.

Influence of the two priors proposed. To investigate the impact of the proposed spatial reconstruction prior and spectral modulation prior, we conduct ablation experiments as shown in Tab. 4 (I-III). It can be observed that all metrics experience varying degrees of decline when there is no spatial reconstruction prior or spectral modulation prior. When both priors are modeled simultaneously, optimal model performance can be achieved, demonstrating the effectiveness of our two proposed priors.

Stage Number	PSNR ↑	SSIM ↑	Q4 ↑	SAM ↓	ERGAS ↓
$K = 1$	42.6860	0.9785	0.8415	0.0207	0.9243
$K = 2$	**42.7301**	0.9787	**0.8427**	**0.0206**	0.9213
$K = 3$	42.6785	**0.9788**	0.8426	0.0206	**0.9191**
$K = 4$	42.6747	0.9781	0.8420	0.0208	0.9267

Table 3: Quantitative results of our method with different number of stages on WorldView-2.

Config	Ω_P	Ω_L	PSNR ↑	SSIM ↑	Q4 ↑	SAM ↓	ERGAS ↓
I	✗	✗	46.3370	0.9856	0.9056	0.0231	1.1096
II	✗	✓	46.4867	0.9858	0.9055	0.0230	1.0998
III	✓	✗	46.3982	0.9852	0.9034	0.0233	1.1171
Ours	✓	✓	**46.5204**	**0.9858**	**0.9065**	**0.0229**	**1.0930**

Table 4: Quantitative results of ablation experiments on the GaoFen-2 datasets.

5 Conclusions

In this paper, we propose a highly interpretable deep unfolding network with precise spatial and spectral priors (SSPEDUN) for pan-sharpening. To fully exploit the potential of the designed priors in reconstructing high-quality spatial and spectral information, we devise the PEM capable of extracting high-quality spatial textures from PAN images and precisely modulating the spectral information of both PAN and LR-MS images to reconstruct pleasing spectral details. Additionally, we customize a data projection module to resolve the data mapping during the optimization process. In this way, both the interpretability and representational capacity of our model are enhanced. Extensive experimental results on three datasets demonstrate the superiority of the proposed SSPEDUN compared to other methods.

Acknowledgment

This work was supported by the Science and Technology Innovation 2025 Major Project of Ningbo City (Grant No. 2022Z032).

References

[1] Wele Gedara Chaminda Bandara and Vishal M Patel. Hypertransformer: A textural and spectral feature fusion transformer for pansharpening. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1767–1777, 2022.

[2] Jiajun Cai and Bo Huang. Super-resolution-guided progressive pansharpening based on a deep convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(6):5206–5220, 2020.

[3] Wjoseph Carper, Thomasm Lillesand, Ralphw Kiefer, et al. The use of intensity-hue-saturation transformations for merging spot panchromatic and multispectral image data. *Photogrammetric Engineering and remote sensing*, 56(4):459–467, 1990.

[4] Osama S. Faragallah. Enhancing multispectral imagery spatial resolution using optimized adaptive pca and high-pass modulation. *International Journal of Remote Sensing*, 39(20):6572–6586, Apr 2018. doi: 10.1080/01431161.2018.1463112. URL http://dx.doi.org/10.1080/01431161.2018.1463112.

[5] Morteza Ghahremani and Hassan Ghassemian. Nonlinear ihs: A promising method for pan-sharpening. *IEEE Geoscience and Remote Sensing Letters*, 13(11):1606–1610, 2016.

[6] Alan R Gillespie, Anne B Kahle, and Richard E Walker. Color enhancement of highly correlated images. ii. channel ratio and "chromaticity" transformation techniques. *Remote Sensing of Environment*, 22(3):343–365, 1987.

[7] Xuanhua He, Keyu Yan, Jie Zhang, Rui Li, Chengjun Xie, Man Zhou, and Danfeng Hong. Multi-scale dual-domain guidance network for pan-sharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[8] Zhengzhong Huang and Liangcai Cao. Bicubic interpolation and extrapolation iteration method for high resolution digital holographic reconstruction. *Optics and Lasers in Engineering*, 130:106090, Mar 2020. doi: 10.1016/j.optlaseng.2020.106090. URL http://dx.doi.org/10.1016/j.optlaseng.2020.106090.

[9] Roger 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)*, volume 2, pages 849–851. IEEE, 2001.

[10] Mingsong Li, Yikun Liu, Tao Xiao, Yuwen Huang, and Gongping Yang. Local-global transformer enhanced unfolding network for pan-sharpening. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 1071–1079. ijcai.org, 2023.

[11] JG 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.

[12] Hangyuan Lu, Yong Yang, Shuying Huang, Wei Tu, and Weiguo Wan. A unified pansharpening model based on band-adaptive gradient and detail correction. *IEEE Transactions on Image Processing*, 31:918–933, 2022. doi: 10.1109/TIP.2021.3137020.

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

[14] Robert A Schowengerdt. Reconstruction of multispatial, multispectral image data using spatial frequency content. *Photogrammetric Engineering and Remote Sensing*, 46(10): 1325–1334, 1980.

[15] Maria Rosaria Vicinanza, Rocco Restaino, Gemine Vivone, Mauro Dalla Mura, and Jocelyn Chanussot. A pansharpening method based on the sparse representation of injected details. *IEEE Geoscience and Remote Sensing Letters*, 12(1):180–184, 2014.

[16] Lucien Wald, Thierry Ranchin, and Marc Mangolini. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogrammetric engineering and remote sensing*, 63(6):691–699, 1997.

[17] Tingting Wang, Faming Fang, Fang Li, and Guixu Zhang. High-quality bayesian pansharpening. *IEEE Transactions on Image Processing*, 28(1):227–239, 2018.

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

[19] Keyu Yan, Man Zhou, Li Zhang, and Chengjun Xie. Memory-augmented model-driven network for pansharpening. In *European Conference on Computer Vision*, pages 306–322. Springer, 2022.

[20] Gang Yang, Man Zhou, Keyu Yan, Aiping Liu, Xueyang Fu, and Fan Wang. Memory-augmented deep conditional unfolding network for pan-sharpening. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1788–1797, 2022.

[21] Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, and John Paisley. Pannet: A deep network architecture for pan-sharpening. In *Proceedings of the IEEE international conference on computer vision*, pages 5449–5457, 2017.

[22] Qiangqiang Yuan, Yancong Wei, Xiangchao Meng, Huanfeng Shen, and Liangpei Zhang. A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(3):978–989, 2018.

[23] 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:1814–1822, 2022.

[24] Hao Zhang, Han Xu, Xin Tian, Junjun Jiang, and Jiayi Ma. Image fusion meets deep learning: A survey and perspective. *Information Fusion*, 76:323–336, 2021.

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

[26] Zhang Bing Zhang Bing, Wu Di Wu Di, Zhang Li Zhang Li, Jiao QuanJun Jiao QuanJun, and Li QingTing Li QingTing. Application of hyperspectral remote sensing for environment monitoring in mining areas. 2012.

[27] Huanyu Zhou, Qingjie Liu, and Yunhong Wang. Panformer: A transformer based model for pan-sharpening. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022.

[28] Man Zhou, Jie Huang, Yanchi Fang, Xueyang Fu, and Aiping Liu. Pan-sharpening with customized transformer and invertible neural network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 3553–3561, 2022.

[29] Man Zhou, Keyu Yan, Jie Huang, Zihe Yang, Xueyang Fu, and Feng Zhao. Mutual information-driven pan-sharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1798–1808, 2022.