SSUN-Net: Spatial-Spectral Prior-Aware Unfolding Network for Pan-Sharpening

Shijie Fang, Hongping Gan*

School of Software, Northwestern Polytechnical University, China
fangshijie@mail.nwpu.edu.cn, ganhongping@nwpu.edu.cn

Abstract

Deep Unfolding Networks (DUNs), with their outstanding performance and partial interpretability, have revitalized the field of pan-sharpening. However, the current DUNs for pan-sharpening rely entirely on implicit deep priors, ignoring the intrinsic physical prior knowledge of MultiSpectral image (MS) and Panchromatic image (PAN) to guide the reconstruction process. Moreover, these methods often depend on single-scale prior features, failing to adequately capture multiscale information, resulting in spatial and spectral distortions in detail. In this paper, we introduce a spatial-spectral prior-aware framework for pan-sharpening, called SSPF, which formulates a constrained minimization problem integrating MS and PAN prior knowledge based on spatial and spectral domains. We further develop SSPF into a lightweight deep unfolding network, called SSUN-Net, which provides more efficient prior feature extraction and requires less computational cost. Additionally, we augment SSUN-Net's capabilities by integrating a customized Multi-scale Prior Structure (MPS). MPS imposes constraints on the solution space at various scales through regularization, which markedly enhances the reconstruction of intricate details. Extensive experiments demonstrate the significant advantages of our proposed SSUN-Net over the current SOTA methods.

Introduction

Pan-sharpening aims to reconstruct High-Resolution MultiSpectral images (HRMS) by utilizing both high resolution Panchromatic images (PAN) and Low-Resolution MultiSpectral images (LRMS), thereby endowing HRMS with spatial information from PAN and spectral information from LRMS (Meng et al. 2021; Vivone et al. 2021). As a result, pan-sharpening is a widely researched hot topic in fields such as remote sensing and computer science (Deng et al. 2022).

Let $\mathbf{H} \in \mathbb{R}^{C\times W\times H}$, $\mathbf{L} \in \mathbb{R}^{C\times w\times h}$, and $\mathbf{P} \in \mathbb{R}^{1\times H\times W}$ be HRMS, LRMS, and PAN, respectively, where C or 1 is the number of bands, H or h is the height size, and W or w is the width size. Mathematically, the quantified reconstruction process of HRMS can be modeled as the following degradation model (Ghamisi et al. 2019):

$$\mathbf{L} = \mathbf{SH} + \mathbf{N_L}, \mathbf{P} = \mathbf{HB} + \mathbf{N_P}, \quad (1)$$

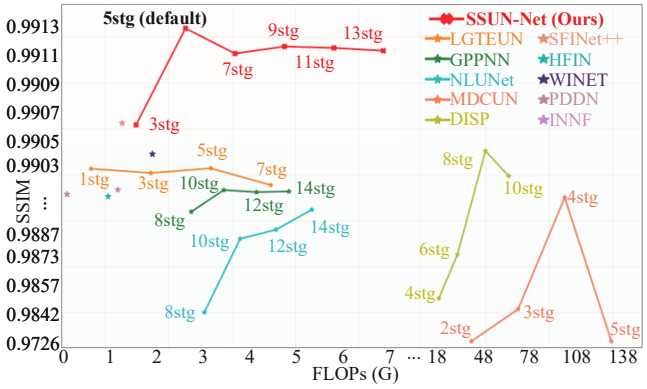


Figure 1: Trade-off SSIM-FLOPs for SSUN-Net and other representative methods. Polyline represent DUNs, and the star-shaped scatter points represent pure DL methods. Compared with other DUNs, our proposed SSUN-Net achieves the best performance with the least computational cost (FLOPs), compared to other DUNs.

where $\mathbf{S}$ and $\mathbf{B}$ denote the spatial and spectral degradation operators, respectively. $\mathbf{N_L}$ and $\mathbf{N_P}$ denote the spatial and spectral noise, respectively. Inverting HRMS from spectral and spatial degradation is a highly ill-posed problem; and many scholars have developed a variety of methods to solve this problem. According to the characteristics of feature extraction techniques, these pan-sharpening methods can be divided into the following three categories.

Traditional pan-sharpening methods involve using established prior knowledge to reconstruct HRMS from LRMS and PAN. Typical methods include Component Substitution (CS) (Aiazzi, Baronti, and Selva 2007), Multi-resolution Analysis (MRA) (Liu 2000), and Variational Optimization (VO) (Fang et al. 2013; Jiang et al. 2014). For example, CS and MRA are utilized for decomposing PAN's spatial features and integrating LRMS. VO treats Eq. (1) as an optimization problem to obtain HRMS. However, these traditional methods always suffer from issues like spectral and spatial distortions, and difficult parameter estimation, which limit their practical applicability.

Pure Deep Learning (DL) methods for pan-sharpening utilize versatile combinations of feature extraction compo-

*Corresponding Author.

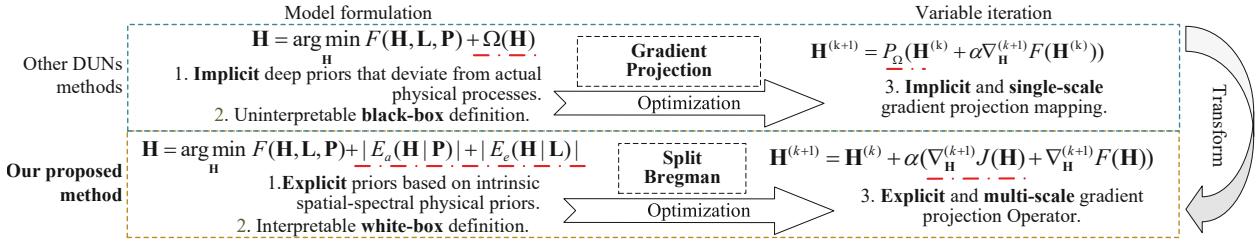


Figure 2: Comparison between our proposed SSUN-Net and previous unfolding framework.

nents to attain impressive outcomes by learning nonlinear correlations between input data and labels in the dataset. PNN (Masi et al. 2016) is one of the first methods to employ Convolutional Neural Networks (CNN) for pan-sharpening, demonstrating superior performance compared to traditional methods. Subsequently, deep learning technology becomes increasingly prominent in pan-sharpening. Recently, SFINet++ (Zhou et al. 2024) and HFIN (Tan et al. 2024) are two representative methods of pure DL methods. However, these methods not only heavily rely on the insights of exceptional researchers but also disregard the fundamental prior knowledge understanding of the underlying process of pan-sharpening. These attributes contribute to a deficiency in interpretability and theoretical robustness (Lempitsky, Vedaldi, and Ulyanov 2018; Zhang et al. 2021).

Deep Unfolding Networks (DUNs) for pan-sharpening are emerging techniques that combine the strengths of the VO and DL technology. GPPNN (Xu et al. 2021), a pioneering DUN for pan-sharpening, establishes a CNN-based deep prior module to obtain implicit prior. Following this, researchers have designed various DUNs that can utilize more complex deep priors, such as the Transformer-based denoising priors of LGTEUN (Li et al. 2023a), and non-local priors of MDCUN (Yang et al. 2022) and NLUNet (Li et al. 2023b). In addition, DISP (Wang et al. 2024) attempts to add intrinsic supervision to DUNs to strengthen modality relevance. These methods can be uniformly simplified into the following paradigm. Firstly, the optimization model based on implicit deep prior for DUNs can be established:

$$\mathbf{H} = \min_{\mathbf{H}} F(\mathbf{H},\mathbf{L},\mathbf{P}) + \Omega(\mathbf{H}), \tag{2}$$

where $F(\mathbf{H},\mathbf{L},\mathbf{P})$ is the fidelity term obtained from Eq. (1), and Ω is the used implicit deep prior. Secondly, Eq. (2) can be solved by projected gradient descent (Aggarwal, Mani, and Jacob 2019) as an iterative convergence problem:

$$\mathbf{H}^{(k+1)} = P_{\Omega}(\mathbf{H}^{(k)} + \nabla_{\mathbf{H}} F(\mathbf{H}^{(k)})), \tag{3}$$

where P_{Ω} is the implicit and single-scale projection operator, which projects $\mathbf{H}^{(k)}$ into a solution space constrained by Ω. This paradigm reveals that current DUN-based pan-sharpening methods still achieve performance by stacking DL-based operators and thus have two additional key limitations: **1) Limited Utilization of Prior Knowledge**: The existing DUNs rely on implicit deep priors, Ω, that deviate from the actual physical processes of pan-sharpening and ignores the intrinsic guidance of the physical prior knowledge of MS and PAN. As a result, their multimodal (spatial-spectral) feature extraction capability is limited and require significant computational cost. **2) Limited Multi-scale Capacity**: These DUNs neglect the constraints on the multi-scale solution space, thus becoming insensitive to subtle information during the reconstruction process.

In order to solve these limitations in DUNs, we present a spatial-spectral prior-aware framework for pan-sharpening, named SSPF, which approaches the reconstruction task as a constrained minimization problem and capitalizes on the intrinsic physical priors inherent to MS and PAN, integrating these prior knowledge across spatial and spectral domains. SSPF is then optimized through the Split Bregman algorithm (Goldstein and Osher 2009) and transformed into a lightweight DUN with an explicit and transparent design, termed SSUN-Net. SSUN-Net aims at adaptive learning feature representations of prior knowledge at a deeper level, facilitating a clearer understanding of the underlying processes and the extraction of features. Furthermore, we integrate a customized Multi-scale Prior Structure (MPS) in SSUN-Net. The structure strategically embeds spectral or spatial prior regularization, enabling the extraction and constraint of the solution space under multi-scale priors, significantly improving the quality of details in the reconstructed images. In summary, our approach diverges from existing DUNs across several pivotal dimensions, as depicted in Fig. 2. More specifically, our main contributions are as follows:

- We propose SSPF, a spatial-spectral prior-aware framework for pan-sharpening, guided by intrinsic physical priors inherent in MS and PAN, rather than implicit ones.
- We optimize the minimization problem in SSPF and unfold SSPF into a lightweight DUN, named SSUN-Net, which more effectively incorporates spatial-spectral prior knowledge to extract features and requires less computational cost.
- We customize MPS for SSUN-Net to extract and amalgamate prior regularization across various scales, thereby enhancing the reconstruction of intricate details.

In addition, the proposed SSUN-Net outperforms state-of-the-art (SOTA) methods in extensive experimental results, while also delivering superior visual effects.

Proposed Method

Model Formulation

Previous DUN-based pan-sharpening frameworks not fully utilize the intrinsic physical priors knowledge of MS and PAN to guide the HRMS reconstruction process. To alleviate

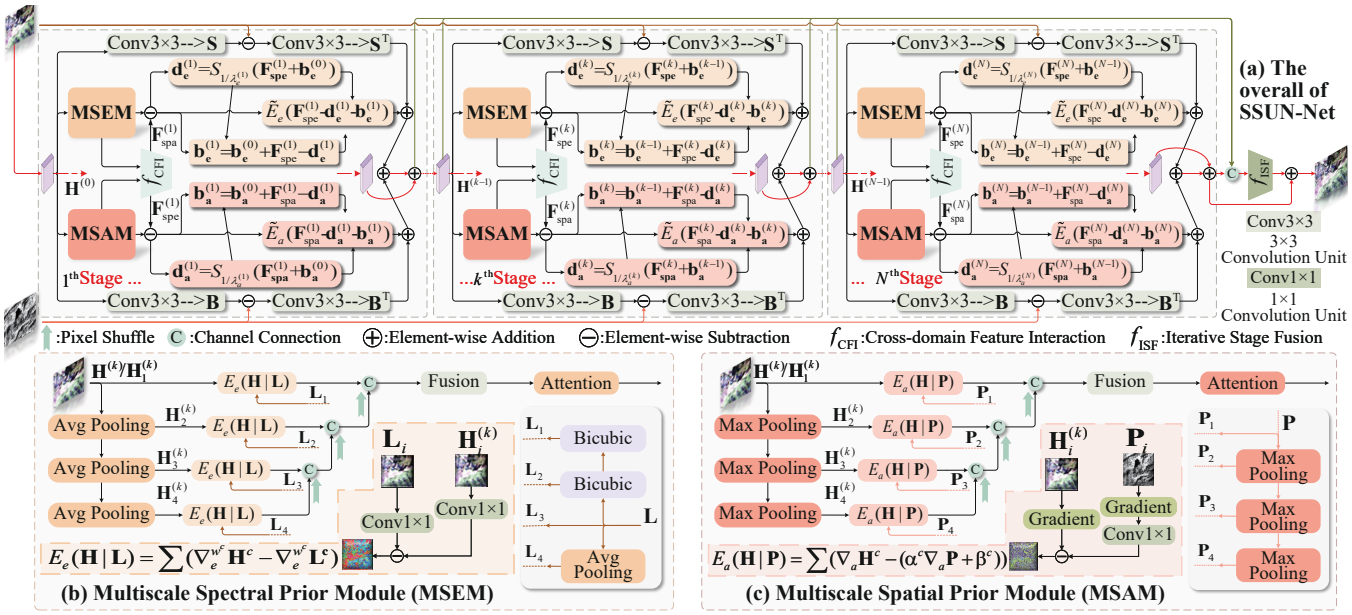


Figure 3: The proposed SSUN-Net, comprised N reconstruction stages, which utilizes multiscale priors from both the spectral and spatial domains to constrain the reconstruction outcomes.

this issue, we introduce SSPF as following:

$$\hat{\mathbf{H}} = \arg\min_{\mathbf{H}} F(\mathbf{H}, \mathbf{L}, \mathbf{P}) + |E_a(\mathbf{H}|\mathbf{P})| + |E_e(\mathbf{H}|\mathbf{L})|, \quad (4)$$

where $F(\mathbf{H}, \mathbf{L}, \mathbf{P}) = \frac{u}{2}||\mathbf{SH} - \mathbf{L}|| + \frac{v}{2}||\mathbf{HB} - \mathbf{P}||$ is constituted of the joint fidelity term for Eq. (1), which is subject to ℓ_2-norm constraint for smooth degradation perception, with u and v is used to balance the contributions of each term. $|E_e(\mathbf{H}|\mathbf{L})|$ and $|E_a(\mathbf{H}|\mathbf{P})|$ represent prior constraints in spectral and spatial domains, respectively, which are regularized by the ℓ_1-norm to improve the sparsity of the prior features for effective denoising feature selection.

Spectral Prior, $E_e(\mathbf{H}|\mathbf{L})$, saves the spectral information in the recovery process from the LRMS modality to the HRMS one, which can be constrained by spectral gradient (Fang et al. 2013), $\nabla\gamma$, and expressed as follows:

$$\nabla\gamma^c = \{\gamma^c - \gamma^{c-1}, c = 1, 2, ..., C\}, \quad (5)$$

where γ^c represents c^{th} band of the image γ. Eq. (5) can be flexibly extended to a non-local spectral gradient:

$$\nabla_e^{w^c}\gamma^c = w_1^c\gamma^1 + w_2^c\gamma^2 + ... + w_C^c\gamma^C, \quad (6)$$

where w_i^c represents the linear weighting coefficients. It is easy to observe that Eq. (5) is a specific form of Eq. (6), where $w_i^c = \begin{cases} 1, & i = c \\ -1, & i = c+1 \\ 0, & \text{others} \end{cases}$. As a result, the spectral prior, $E_e(\mathbf{H}|\mathbf{L})$, can be expressed in SSPF as:

$$E_e(\mathbf{H}|\mathbf{L}) = \sum_{c\in C}(\nabla_e^{w^c}\mathbf{H}^c - \nabla_e^{w^c}\mathbf{L}^c). \quad (7)$$

Spatial Prior, $E_a(\mathbf{H}|\mathbf{P})$, saves the structural information in the recovery process from PAN modality to HRMS one. Since PAN and MS originate from the same ground information, their gradient information exhibits a high degree of consistency (Fu et al. 2019). Therefore, the relationship between $\mathbf{P}$ and $\mathbf{H}^c$ can be defined as:

$$\nabla_a\mathbf{H}^c = \alpha^c\nabla_a\mathbf{P} + \beta^c, \quad (8)$$

where α^c and β^c are to estimate the linear transformation from $\nabla_a\mathbf{P}$ to $\nabla_a\mathbf{H}^c$. And $\nabla_a\gamma$ denotes the image gradient of γ, which can be obtained based on the difference between local adjacent pixels:

$$\begin{aligned}\nabla_a\gamma_{(i,j)} = &8\gamma_{(i,j)} - \gamma_{(i+1,j-1)} - \gamma_{(i+1,j)} - \gamma_{(i+1,j+1)} \\ &- \gamma_{(i,j-1)} - \gamma_{(i,j+1)} - \gamma_{(i-1,j-1)} - \gamma_{(i-1,j)} - \gamma_{(i-1,j+1)},\end{aligned} \quad (9)$$

where $\gamma_{(i,j)}$ represents the pixel intensity of the γ at coordinates (i, j). Therefore, the spatial prior, $E_a(\mathbf{H}|\mathbf{P})$, can be expressed in SSPF as:

$$E_a(\mathbf{H}|\mathbf{P}) = \sum_{c\in C}(\nabla_a\mathbf{H}^c - (\alpha^c\nabla_a\mathbf{P} + \beta^c)). \quad (10)$$

Model Optimization

In this section, we outline the procedure of using Split Bregman algorithm to optimize SSPF.

Firstly, by introducing the auxiliary variables $\mathbf{d_e}$ and $\mathbf{d_a}$, we can separate the regularization term of Eq. (4), which can be expressed as:

$$\begin{aligned}&\min_{\mathbf{H}} F(\mathbf{H}, \mathbf{L}, \mathbf{P}) + |\mathbf{d_e}| + |\mathbf{d_a}|, \\ &\text{s.t. } \mathbf{d_e} = E_e(\mathbf{H}|\mathbf{L}), \mathbf{d_a} = E_a(\mathbf{H}|\mathbf{P}).\end{aligned} \quad (11)$$

Subsequently, we can add the penalty function term to Eq. (11) and convert it into an equivalent unconstrained problem as follows:

$$\begin{aligned}\min_{\mathbf{H},\mathbf{d_e},\mathbf{d_a}} &F(\mathbf{H}, \mathbf{L}, \mathbf{P}) + \lambda_e|\mathbf{d_e}| + ||\mathbf{d_e} - E_e(\mathbf{H}|\mathbf{L})|| \\ &+ \lambda_a|\mathbf{d_a}| + ||\mathbf{d_a} - E_a(\mathbf{H}|\mathbf{P})||,\end{aligned} \quad (12)$$

where λ_e and λ_a serve as penalty coefficients that regulate the extent to which the equation enforces the constraint. Moreover, we follow the principle of Bregman iteration (Bregman 1967) and introduce iteration variables $\mathbf{b_e}$ and $\mathbf{b_a}$ to modify Eq. (12) by iteratively solving as follows:

$$\begin{aligned}\min_{\mathbf{H},\mathbf{d_e},\mathbf{d_a}} F(\mathbf{H},\mathbf{L},\mathbf{P}) + \frac{\lambda_e}{2}||\mathbf{d_e} - E_e(\mathbf{H}|\mathbf{L}) - \mathbf{b_e}^{(k)}|| + |\mathbf{d_e}| \\ + \frac{\lambda_a}{2}||\mathbf{d_a} - E_a(\mathbf{H}|\mathbf{P}) - \mathbf{b_a}^{(k)}|| + |\mathbf{d_a}|,\end{aligned} \tag{13}$$

where $\mathbf{b_e}$ and $\mathbf{b_a}$ are iterative variables, updated in the k^{th} iteration by the following equation:

$$\begin{cases} \mathbf{b_e}^{(k+1)} = \mathbf{b_e}^{(k)} + E_e(\mathbf{H}^{(k+1)}|\mathbf{L}) - \mathbf{d_e}^{(k+1)}, & \text{(14a)} \\ \mathbf{b_a}^{(k+1)} = \mathbf{b_a}^{(k)} + E_a(\mathbf{H}^{(k+1)}|\mathbf{P}) - \mathbf{d_a}^{(k+1)}, & \text{(14b)} \end{cases}$$

where $k \in \{0, 1, ..., N\}$ and N is the maximum number of iterations. As a result, given that the mixed constraints of the energy equation in Eq. (4) are decoupled, and we can efficiently minimize Eq. (13) by iteratively minimizing $\mathbf{H}$ and the auxiliary variables $\mathbf{d_e}$ and $\mathbf{d_a}$, respectively, through the following steps:

$$\begin{cases} \mathbf{H}^{(k+1)} = \arg\min_{\mathbf{H}} F(\mathbf{H},\mathbf{L},\mathbf{P}) + \frac{\lambda_e}{2}||\mathbf{d_e}^{(k)} - E_e(\mathbf{H}|\mathbf{L}) - \mathbf{b_e}^{(k)}|| \\ \qquad\qquad + \frac{\lambda_a}{2}||\mathbf{d_a}^{(k)} - E_a(\mathbf{H}|\mathbf{P}) - \mathbf{b_a}^{(k)}||, & \text{(15)} \\ \mathbf{d_e}^{(k+1)} = \arg\min_{\mathbf{d_e}} \frac{\lambda_e}{2}||\mathbf{d_e} - E_e(\mathbf{H}^{(k+1)}|\mathbf{L}) - \mathbf{b_e}^{(k)}|| + |\mathbf{d_e}|, & \text{(16)} \\ \mathbf{d_a}^{(k+1)} = \arg\min_{\mathbf{d_a}} \frac{\lambda_a}{2}||\mathbf{d_a} - E_a(\mathbf{H}^{(k+1)}|\mathbf{P}) - \mathbf{b_a}^{(k)}|| + |\mathbf{d_a}|. & \text{(17)} \end{cases}$$

The update for $\mathbf{d_e}$ and $\mathbf{d_a}$ are evident from Eq. (16) and Eq. (17) that there is no coupling between the elements of $\mathbf{d_e}$, $\mathbf{d_a}$, and $\mathbf{H}$. We can compute the optimal values of $\mathbf{d_e}$ and $\mathbf{d_a}$ using the soft thresholding denoted as $S_\epsilon(\cdot)$ (Beck and Teboulle 2009), which can be expressed as follows:

$$\begin{cases} \mathbf{d_e}^{(k+1)} = S_{1/\lambda_e}(E_e(\mathbf{H}^{(k+1)}|\mathbf{L}) + \mathbf{b_e}^{(k)}), & \text{(18a)} \\ \mathbf{d_a}^{(k+1)} = S_{1/\lambda_a}(E_a(\mathbf{H}^{(k+1)}|\mathbf{P}) + \mathbf{b_a}^{(k)}), & \text{(18b)} \end{cases}$$

where:

$$S_\epsilon(\cdot) = sgn(\cdot) * max(|\cdot| - \epsilon, 0). \tag{19}$$

We only do one iteration in the update of $\mathbf{d_e}$ and $\mathbf{d_a}$ to maximize the efficiency. The update of $\mathbf{H}$ in Eq. (15) can be carried out using a gradient descent algorithm, and expressed as follows:

$$\mathbf{H}^{(k+1)} = \mathbf{H}^{(k)} + \alpha(\nabla_{\mathbf{H}}^{(k+1)} J(\mathbf{H}) + \nabla_{\mathbf{H}}^{(k+1)} F(\mathbf{H})), \tag{20}$$

where α denotes the iteration step size, and $\nabla_{\mathbf{H}} J(\mathbf{H})$ with respect to the regularization-based data term can be further computed as:

$$\begin{aligned}\nabla_{\mathbf{H}}^{(k+1)} J(\mathbf{H}) = \frac{\lambda_e}{2}\widetilde{E_e}(E_e(\mathbf{H}^{(k)}|\mathbf{L}) - \mathbf{d_e}^{(k+1)} - \mathbf{b_e}^{(k+1)}) \\ + \frac{\lambda_a}{2}\widetilde{E_a}(E_a(\mathbf{H}^{(k)}|\mathbf{P}) - \mathbf{d_a}^{(k+1)} - \mathbf{b_a}^{(k+1)}),\end{aligned} \tag{21}$$

where $\widetilde{\star}$ denotes the pseudo-inverse operator, and $\nabla_{\mathbf{H}} F(\mathbf{H})$ with respect to the degradation-based data term can be further computed as:

$$\nabla_{\mathbf{H}}^{(k+1)} F(\mathbf{H}) = \frac{u}{2}\mathbf{S}^\top(\mathbf{S}\mathbf{H}^{(k)} - \mathbf{L}) + \frac{v}{2}(\mathbf{B}\mathbf{H}^{(k)} - \mathbf{P})\mathbf{B}^\top, \tag{22}$$

where $\star^\top$ represents the transpose operator. In general, SSPF iterates according to the above equation until the variables converge or the maximum number of iterations is reached.

Deep Unfolding Network

We expand the iterations in SSPF into a deep unfolding network, SSUN-Net, as depicted in Fig. 3a. The k^{th} stage of SSUN-Net correspond to the k^{th} iteration in SSPF, enhancing the perception of scenes and data in a learnable manner.

Adaptive Degradation Perception. During the k^{th} stage, internal parameters ($\lambda_e^{(k)}$, $\lambda_a^{(k)}$, $u^{(k)}$, $v^{(k)}$, $\alpha^{(k)}$) and the operators $\mathbf{S}$ and $\mathbf{B}$ are not manually tuned. Instead, they are configured as learnable modules at the k^{th} stage, permitting SSUN-Net to estimate them adaptively. For $\mathbf{B}$ and $\mathbf{S}$, we employ convolutional units with the number of hidden layers being mid and a ReLU activation to estimate degradation projection. Similarly, the pseudo-inverse and transpose of these operators are implemented using specific transpose convolutions. Consequently, Eq. (22) can be reformulated as follows:

$$\begin{aligned}\nabla_{\mathbf{H}}^{(k)} F(\mathbf{H}) = \frac{u^{(k)}}{2} f_{\mathbf{S}^\top}^{mid}(f_{\mathbf{S}}^{mid}(\mathbf{H}^{(k-1)}) - \mathbf{L}) \\ + \frac{v^{(k)}}{2} f_{\mathbf{B}^\top}^{mid}(f_{\mathbf{B}}^{mid}(\mathbf{H}^{(k-1)}) - \mathbf{P}),\end{aligned} \tag{23}$$

where $f_\Theta^{mid}(\cdot)$, $\Theta \in \{\mathbf{B}, \mathbf{S}, \mathbf{B}^\top, \mathbf{B}^\top\}$, represents the convolution unit corresponding to the Θ operator.

Multi-scale Prior Regularization. We introduce a Multi-scale Prior Structure (MPS), which serves as a cornerstone of our approach. By integrating our bespoke priors, $E_e(\mathbf{H}|\mathbf{L})$ and $E_a(\mathbf{H}|\mathbf{P})$, into the MPS, we develop a Multi-scale Spatial Prior Module (MSAM) and a Multi-scale Spectral Prior Module (MSEM), as shown in Fig. 3b and Fig. 3c, respectively. The spectral prior feature $\mathbf{F}_{\text{spe}}^{(k)}$ and the spatial prior feature $\mathbf{F}_{\text{spa}}^{(k)}$ are derived by processing $\mathbf{H}^{(k-1)}$ through MSAM and MSEM, respectively. These features are then interacted in the customized Cross-domain Feature Interaction (CFI) mechanism to promote complementarity and reduce redundancy, expressed as follows:

$$\begin{aligned}\{\mathbf{F}_{\text{spe}}^{(k)}, \mathbf{F}_{\text{spa}}^{(k)}\} = \text{S}(f_{\text{CFI}}(\text{C}(\text{MSEM}(\mathbf{H}^{(k-1)}, \mathbf{L}), \\ \text{MSAM}(\mathbf{H}^{(k-1)}, \mathbf{P})))),\end{aligned} \tag{24}$$

where, $f_{\text{CFI}}(\cdot)$ signifies the convolutional unit dedicated to feature interaction, $\text{C}(\cdot)$ denotes the concatenation process along the channel dimension, and $\text{S}(\cdot)$ refers to the channel-wise split operation. Subsequently, the regularization denoted by $\mathbf{F}_{\text{spe}}^{(k)}$ and $\mathbf{F}_{\text{spa}}^{(k)}$ are applied to project the solution space onto the multiscale spatial-spectral prior and thus the equation corresponding to Eq. (21) is expressed as follows:

$$\begin{aligned}\nabla_{\mathbf{H}}^{(k)} J(\mathbf{H}) = \frac{\lambda_e^{(k)}}{2}\widetilde{E_e}(\mathbf{F}_{\text{spe}}^{(k)} - \mathbf{d_e}^{(k)} - \mathbf{b_e}^{(k)}) \\ + \frac{\lambda_a^{(k)}}{2}\widetilde{E_a}(\mathbf{F}_{\text{spa}}^{(k)} - \mathbf{d_a}^{(k)} - \mathbf{b_a}^{(k)}).\end{aligned} \tag{25}$$

	Metrics	Pure Deep Learning Methods					Deep Unfolding Networks					
		SFINet++ (TPAMI'24)	HFIN (CVPR'24)	WINET (TGRS'24)	PDDN (ICCV'23)	INNF (AAAI'22)	DISP (AAAI'24)	NLUNet (TGRS'23)	LGTEUN (IJCAI'23)	MDCUN (CVPR'22)	GPPNN (CVPR'21)	**SSUN-Net (Ours)**
WV-II	ERGAS↓	0.9538	0.9906	0.9024	0.9889	0.9176	0.8759	1.0034	0.8968	1.0060	0.9248	**0.8349**
	SSIM↑	0.9675	0.9694	0.9714	0.9702	0.9706	0.9720	0.9644	0.9734	0.9635	0.9702	**0.9740**
	PSNR↑	41.587	41.903	42.048	41.374	41.895	42.253	41.064	42.677	41.130	41.838	**42.705**
	SCC↑	0.5147	0.5697	0.5753	0.5819	0.5756	0.5837	0.5413	0.5832	0.5274	0.5807	**0.5942**
	SAM↓	0.0236	0.0226	0.0226	0.0240	0.0222	0.0215	0.0246	0.0211	0.0249	0.0222	**0.0202**
WV-III	ERGAS↓	3.0217	3.1523	3.1753	3.3839	3.1000	3.0096	3.2981	3.0151	3.3531	3.0923	**2.9054**
	SSIM↑	0.9261	0.9192	0.9187	0.9088	0.9216	0.9267	0.9140	0.9246	0.9111	0.9226	**0.9292**
	PSNR↑	30.767	30.481	30.327	29.844	30.542	30.835	29.972	30.788	29.834	30.577	**31.116**
	SCC↑	0.8333	0.8156	0.8205	0.8095	0.8250	0.8315	0.8162	0.8291	0.8142	0.8290	**0.8361**
	SAM↓	0.0720	0.0766	0.0802	0.0852	0.0745	0.0720	0.0811	0.0723	0.0828	0.0738	**0.0690**
GF-2	ERGAS↓	0.4376	0.4907	0.4472	0.5098	0.4831	0.4493	0.4976	0.4798	0.4830	0.4812	**0.4286**
	SSIM↑	0.9906	0.9891	0.9904	0.9892	0.9894	0.9904	0.9885	0.9894	0.9890	0.9893	**0.9913**
	PSNR↑	49.358	48.311	49.107	48.947	48.520	49.090	48.229	48.529	48.414	48.496	**49.481**
	SCC↑	0.5897	0.5583	0.5630	0.5440	0.5664	0.5679	0.5458	0.5728	0.5461	0.5596	**0.5996**
	SAM↓	0.0087	0.0090	0.0090	0.0102	0.0095	0.0097	0.0099	0.0097	0.0098	0.0096	**0.0087**
FLOPs (G)		1.3112	1.0104	1.9597	0.1284	1.2201	55.334	4.6099	3.2113	118.30	4.1901	2.6698
Params (M)		0.0848	0.0773	0.3336	0.0395	0.0613	3.0872	0.3062	0.3004	0.1538	0.3594	0.2934

Table 1: Comparison of SSUN-Net with other methods on simulated data. The symbol ↑ or ↓ is used to indicate that a higher or lower value corresponds to a better result. Refer to the *Supplementary Material* for more details on the comparative evaluation.

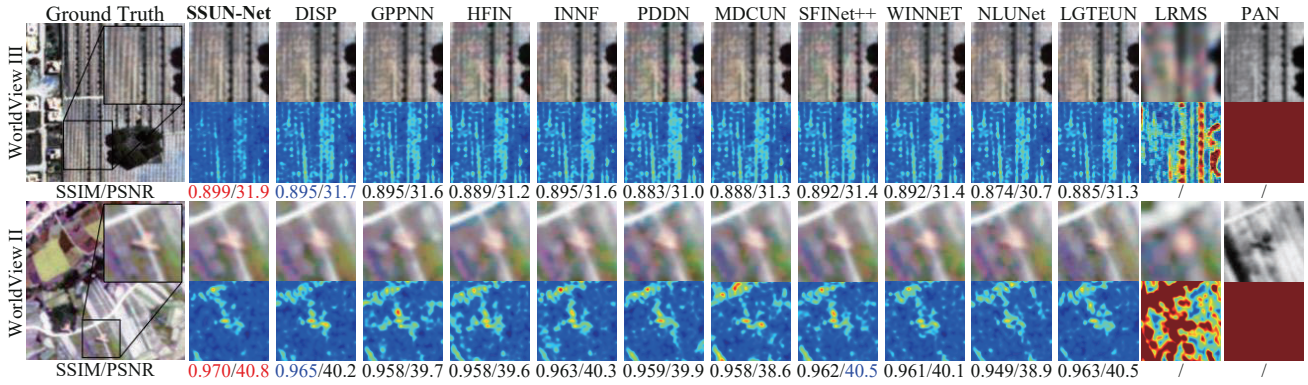


Figure 4: Visual comparison of SSUN-Net with other methods on simulated data from the WorldView III and WorldView II.

Iterative Stage Fusion (ISF). Notably, to prevent the loss of image-level information, we design the Iterative Stage Fusion (ISF) mechanism, which concatenates the outputs from each stage, denoted as $\{\mathbf{H}^{(k)}\}_{k=1}^{N}$. Subsequently, a nonlinear weighting is applied to these concatenated outputs to generate the final output $\mathbf{X}$:

$$\mathbf{X} = \mathbf{H}^{(N)} + f_{\text{ISF}}(\text{C}(\{\mathbf{H}^{(k)}\}_{k=1}^{N})), \tag{26}$$

where $f_{\text{ISF}}(\cdot)$ represents the nonlinear weighting operator, which consists of convolutions to reduce the channel dimension and ReLU activation.

Loss Function

We introduce pixel loss and structural loss to jointly penalize the difference between the reconstructed image $\mathbf{X}$ and the Ground-Truth image (GT). The pixel loss, $L_{Pixel}(\theta)$, is defined as the ℓ_1 distance between $\mathbf{X}$ and GT:

$$L_{Pixel}(\theta) = |\mathbf{X} - \text{GT}|, \tag{27}$$

where θ represents a set of learnable parameters of SSUN-Net. In addition, we establish the structural loss, $L_{\nabla_a}(\theta)$, based on the spatial gradient defined in Eq. (9), as follows:

$$L_{\nabla_a}(\theta) = log|(\nabla_a \mathbf{X}) - (\nabla_a \text{GT})|. \tag{28}$$

Finally, the overall loss function of SSUN-Net is formulated as follows:

$$L(\theta) = L_{Pixel}(\theta) + \lambda_{loss} L_{\nabla_a}(\theta), \tag{29}$$

where λ_{loss} is the weight factor, which is set to 0.1 to optimize the performance of SSUN-Net.

Experiment

Experiment settings. We adopt WorldView II (WV-II), WorldView III (WV-III), GaoFen 2 (GF-2) satellite datasets, and generate reduced resolution simulated data through the Wald (Wald, Ranchin, and Mangolini 1997) protocol for simulation testing that follow the previous works. For each data pair, the space sizes of PAN and LRMS are 128×128 and 32×32, respectively. We use image quality assessment metrics for simulation testing, including: Peak Signal-to-Noise Ratio (PSNR) (Huynh-Thu and Ghanbari

Metrics	Pure Deep Learning Methods					Deep Unfolding Methods					
	SFINet++	HFIN	WINET	PDDN	INNF	DISP	NLUNet	LGTEUN	MDCUN	GPPNN	SSUN-Net
$D_\lambda \downarrow$	0.0784	0.0984	0.1187	0.0798	<u>0.0708</u>	0.0807	0.0763	0.1134	0.0765	0.0912	**0.0658**
$D_s \downarrow$	0.0886	0.0849	0.0827	0.1063	0.1076	<u>0.0822</u>	0.1074	0.1716	0.0846	0.0948	**0.0804**
$QNR \uparrow$	0.8400	0.8251	0.8084	0.8223	0.8292	<u>0.8437</u>	0.8245	0.7345	0.8454	0.8226	**0.8591**

Table 2: Comparison of SSUN-Net with other methods on real data from the GaoFen 2.

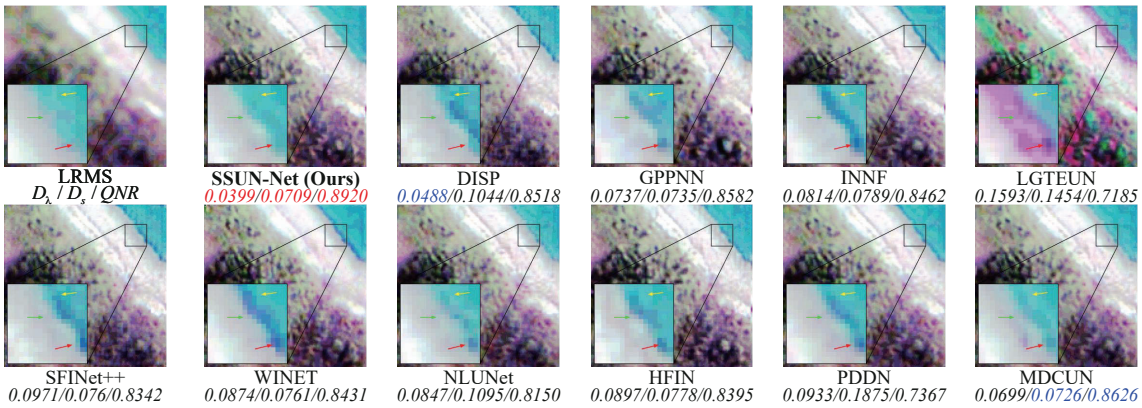


Figure 5: Visual comparison of SSUN-Net with other methods on real data from the GaoFen 2.

2008), Structural Similarity Index Measure (SSIM) (Wang et al. 2004), Spectral Angle Mapper (SAM) (Yuhas, Goetz, and Boardman 1992), Spatial Correlation Coefficient (SCC) (J. Zhou and Silander 1998), and Error Relative Global Dimension Synthesis (ERGAS) (Alparone et al. 2007).

In addition, we employ the additional GaoFen 2 dataset as full resolution real data for real testing. As there are no high-resolution multispectral ground-truth images as a reference in reality, no-reference metrics are used, including Spatial Distortion Index (D_s), Spectral Distortion Index (D_λ), and No-Reference Quality (QNR) (Alparone et al. 2008).

Comparison

The SOTA methods compared include pure DL methods (SFINet++, PDDN (He et al. 2023), INNF (Zhou et al. 2022)), WINET (Zhang et al. 2024), HFIN and DUNs (LGTEUN, MDCUN, GPPNN, NLUNet, DISP).

Evaluation of simulated data. Tab. 1 presents quantitative comparisons of our proposed SSUN-Net with other methods on simulated data, highlighting the best and second-best results in **Bold** and <u>Underline</u>, respectively. SSUN-Net demonstrates significant advantages. For example, compared with the classic DUN-based method GPPNN, SSUN-Net reduces ERGAS↓ by 0.0899 (9.7%), 0.1869 (6.1%), and 0.0526 (10.9%), on three datasets, respectively. Moreover, even compared with the SOTA pure DL-based method SFINet++ and DUN-based method DISP across three datasets, SSUN-Net reduces ERGAS↓ by 0.1189 (14.2%) and 0.0410 (4.9%); 0.1163 (4.0%) and 0.1042 (3.6%); and 0.0089 (2.0%) and 0.2065 (4.8%), respectively.

Furthermore, Fig. 4 provides a qualitative evaluation of visual comparison on simulated data, including visualization of RGB bands and mean square error, and zooming in on the region of interest for better observation. According to Fig. 4, we can see that SSUN-Net exhibits the finest texture, the clearest contour, and the least error.

It is worth noting that SSUN-Net achieves better results with less computational cost (FLOPs) than other DUNs, which may benefit from the sufficient guidance of prior knowledge on reconstruction.

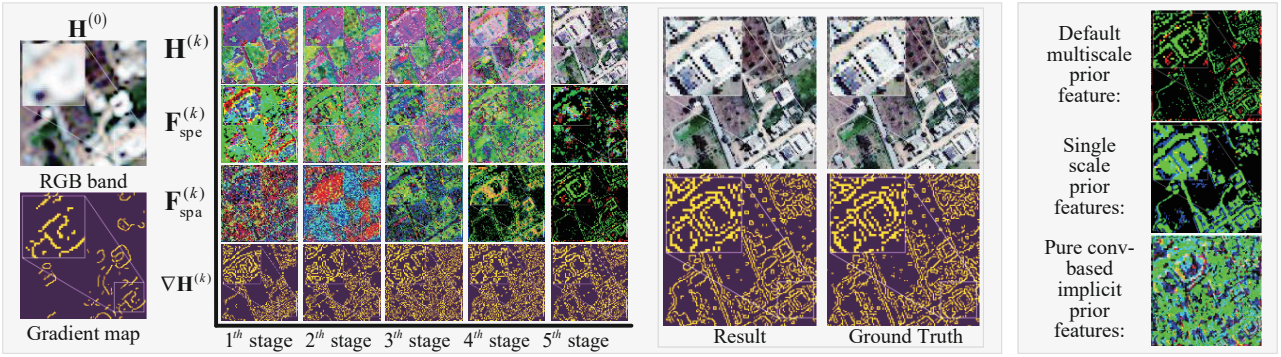
Evaluation of real data. Tab. 2 presents quantitative comparisons on real data. Compared with the GPPNN, SSUN-Net decreases $D_\lambda \downarrow$ and $D_s \downarrow$ by 0.0254 (38.6%) and 0.0144 (17.9%), respectively, and improves $QNR \uparrow$ by 0.0364 (4.4%). Similarly, SSUN-Net improves $QNR \uparrow$ by 0.0190 (2.3%) and 0.0154 (1.8%) compared to the SOTA method SFINet++ and DISP, respectively. Furthermore, Fig. 5 provides a qualitative comparison. The area marked by arrows highlights significant artifacts in the color transition areas of other models, while SSUN-Net closely matches the objective reality. These results demonstrate that SSUN-Net is more effective in applications of real-world tasks.

Ablation Study

Impact of key components. To evaluate the effectiveness of CFI, ISF, and the Attention Mechanism (ATT) in MSAM and MSEM, we replace them with Dense Block (Huang et al. 2017), which have equivalent parameters. Net1, Net2, and Net3 in Tab 3 correspond to the ablation of CFI, ISF, and ATT, respectively. Part 1 in Tab 3 shows that the default SSUN-Net reduces ERGAS↓ by 0.0277 (6.1%), 0.0121 (2.7%), and 0.0152 (3.4%) compared to Net1, Net2, and Net3, respectively. These results substantiate the effectiveness of the components and synergistic integration of these components confers optimal performance upon the model.

Case	MODEL	Configuration of key components						Simulated Test			Real Test			FLOPs (G)	Params (M)
		CFI	ISF	ATT	MSM	$E_a(\mathbf{H}\|\mathbf{P})$	$E_e(\mathbf{H}\|\mathbf{L})$	ERGAS↓	SSIM↑	PSNR↑	$D_\lambda\downarrow$	$D_s\downarrow$	$QNR\uparrow$		
Default	**SSUN-Net**	✔	✔	✔	✔	✔	✔	**0.4286**	**0.9913**	**49.481**	**0.0658**	**0.0804**	**0.8591**	2.6698	0.2934
Part1:	Net1	✘	✔	✔	✔	✔	✔	0.4563	0.9902	48.944	0.0788	0.1971	0.7396	1.7261	0.2358
	Net2	✔	✘	✔	✔	✔	✔	0.4407	0.9908	49.227	0.0560	0.1247	0.8263	2.5565	0.2865
	Net3	✔	✔	✘	✔	✔	✔	0.4330	0.9911	49.357	0.0717	0.1053	0.8306	3.6380	0.2935
Part2:	Net4	✔	✔	✔	✔	✘	✔	0.4438	0.9906	49.202	0.0789	0.1028	0.8264	3.2322	0.3127
	Net5	✔	✔	✔	✔	✔	✘	0.4436	0.9906	49.278	0.0720	0.1250	0.8120	3.8797	0.2697
	Net6	✔	✔	✔	✘	✘	✘	0.4672	0.9895	48.726	0.0635	0.0810	0.8607	4.7964	0.2947
Part3:	Net7	✔	✔	✔	✘	✔	✔	0.4438	0.9902	49.209	0.0640	0.0841	0.8573	3.0442	0.2930
	Net8	✘	✘	✘	✔	✘	✘	0.4697	0.9897	48.935	0.0743	0.0905	0.8419	3.6730	0.3383
	GPPNN	-	-	-	-	-	-	0.4812	0.9893	48.496	0.0712	0.0949	0.8407	4.1901	0.3594

Table 3: Comparative results of ablation studies on simulated data and real data from the GaoFen 2.



(a) Intermediate reconstruction process of SSUN-Net. **(b) Prior features.**

Figure 6: Visual analysis on the iteration mechanism of SSUN-Net, MSAM and MSEM: (a) the change of $\mathbf{H}^{(k)}$, $\nabla\mathbf{H}^{(k)}$, $\mathbf{F}^{(k)}_{spe}$ and $\mathbf{F}^{(k)}_{spa}$ during the reconstruction process; (b) comparison of different prior features.

Impact of physical prior knowledge. To verify the awareness of physical prior knowledge, we design three variants of SSUN-Net, that is, SSUN-Net without using $E_a(\mathbf{H}|\mathbf{P})$ (Net4), SSUN-Net without using $E_e(\mathbf{H}|\mathbf{L})$ (Net5), and SSUN-Net without using $E_a(\mathbf{H}|\mathbf{P})$ and $E_e(\mathbf{H}|\mathbf{L})$ (Net6). As shown in Part 2 of Tab. 3, default SSUN-Net reduces ERGAS↓ by 0.0152 (3.42%), 0.0150 (3.37) and 0.038 (8.25%) compared to Net4, Net5, and Net6, respectively. Furthermore, the iterative process of $\mathbf{H}^{(k)}$, $\nabla\mathbf{H}^{(k)}$ and two prior features are visualized in Fig. 6 (a). According to Fig. 6 (a), we can obtain that MSEM focuses on capturing continuous spectral features, MSAM focuses on discrete structural features, and $\mathbf{H}^{(k)}$ and $\nabla\mathbf{H}^{(k)}$ become closer to GT with iteration progresses.

Impact of MPS. To assess the effectiveness of MPS, we design Net8, which incorporates GPPNN with MPS; and Net7, which is SSUN-Net without MPS. Part 3 in Tab. 3 indicates that compared to GPPNN and the default SSUN-Net, Net8 reduces ERGAS↓ by 0.0115 (2.5%) and increases it by 0.0511 (10.6%), respectively. Furthermore, SSUN-Net reduces ERGAS↓ by 0.0152 (3.42%) compared to Net7. This improvement in quality is primarily due to the focus on multiscale information enabled by MPS. In addition, we present the visualization of the features of single-scale and multi-scale $E_a(\mathbf{H}|\mathbf{P})$-based prior feature and implicit prior based on Dense Block in Fig. 6 (b). According to Fig. 6 (b), we can observe that the implicit prior feature, which deviates from physical prior knowledge, makes it difficult to focus on the structural information. Additionally, the single-scale prior feature is not sensitive to fine-grained information. Instead, our approach addresses these issues.

Moreover, refer to the *Supplementary Material* for additional experiments.

Conclusion

In this paper, we propose SSPF, a pan-sharpening framework that fully exploits the physical prior knowledge of PAN and MS. Then, we use Split Bregman algorithm to optimize and expand SSPF into SSUN-Net, which has a more transparent design, deeper prior feature extraction, and less computational cost. In addition, we customize MPS for SSUN-Net to address the lack of multiscale prior constraints in DUNs, allowing SSUN-Net to recover fine information with high quality. Comprehensive experiments show that our SSUN-Net significantly outperforms SOTA methods.

Code and Supplementary Materials —
https://github.com/ICSResearch/SSUN-Net

Acknowledgments

The work was supported by the National Natural Science Foundation of China under Grants 62101455 and 62471395.

References

Aggarwal, H. K.; Mani, M. P.; and Jacob, M. 2019. MoDL: Model-Based Deep Learning Architecture for Inverse Problems. *IEEE Transactions on Medical Imaging*, 38(2): 394–405.

Aiazzi, B.; Baronti, S.; and Selva, M. 2007. Improving Component Substitution Pansharpening Through Multivariate Regression of MS +Pan Data. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10): 3230–3239.

Alparone, L.; Aiazzi, B.; Baronti, S.; Garzelli, A.; Nencini, F.; and Selva, M. 2008. Multispectral and Panchromatic Data Fusion Assessment Without Reference. *ASPRS Journal of Photogrammetric Engineering and Remote Sensing*, 74: 193–200.

Alparone, L.; Wald, L.; Chanussot, J.; Thomas, C.; Gamba, P.; and Bruce, L. M. 2007. Comparison of Pansharpening Algorithms: Outcome of the 2006 GRS-S Data-Fusion Contest. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10): 3012–3021.

Beck, A.; and Teboulle, M. 2009. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, 2(1): 183–202.

Bregman, L. 1967. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3): 200–217.

Deng, L.-j.; Vivone, G.; Paoletti, M. E.; Scarpa, G.; He, J.; Zhang, Y.; Chanussot, J.; and Plaza, A. 2022. Machine Learning in Pansharpening: A benchmark, from shallow to deep networks. *IEEE Geoscience and Remote Sensing Magazine*, 10(3): 279–315.

Fang, F.; Li, F.; Shen, C.; and Zhang, G. 2013. A Variational Approach for Pan-Sharpening. *IEEE Transactions on Image Processing*, 22(7): 2822–2834.

Fu, X.; Lin, Z.; Huang, Y.; and Ding, X. 2019. A Variational Pan-Sharpening With Local Gradient Constraints. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10257–10266.

Ghamisi, P.; Rasti, B.; Yokoya, N.; Wang, Q.; Hofle, B.; Bruzzone, L.; Bovolo, F.; Chi, M.; Anders, K.; Gloaguen, R.; Atkinson, P. M.; and Benediktsson, J. A. 2019. Multisource and Multitemporal Data Fusion in Remote Sensing: A Comprehensive Review of the State of the Art. *IEEE Geoscience and Remote Sensing Magazine*, 7(1): 6–39.

Goldstein, T.; and Osher, S. 2009. The Split Bregman Method for L1-Regularized Problems. *SIAM Journal on Imaging Sciences*, 2(2): 323–343.

He, X.; Yan, K.; Li, R.; Xie, C.; Zhang, J.; and Zhou, M. 2023. Pyramid Dual Domain Injection Network for Pansharpening. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 12862–12871.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269.

Huynh-Thu, Q.; and Ghanbari, M. 2008. Scope of validity of PSNR in image/video quality assessment. *Electronics Letters*, 44: 800–801.

J. Zhou, D. L. C.; and Silander, J. A. 1998. A wavelet transform method to merge Landsat TM and SPOT panchromatic data. *International Journal of Remote Sensing*, 19(4): 743–757.

Jiang, Y.; Chen, L.; Wang, W.; Ding, X.; and Huang, Y. 2014. A compressed sensing-based pan-sharpening using joint data fidelity and blind blurring kernel estimation. In *2014 IEEE International Conference on Image Processing (ICIP)*, 5042–5046.

Lempitsky, V.; Vedaldi, A.; and Ulyanov, D. 2018. Deep Image Prior. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9446–9454.

Li, M.; Liu, Y.; Xiao, T.; Huang, Y.; and Yang, G. 2023a. Local-Global Transformer Enhanced Unfolding Network for Pan-sharpening. In Elkind, E., ed., *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, 1071–1079. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Li, X.; Li, Y.; Shi, G.; Zhang, L.; Li, W.; and Lei, D. 2023b. Pansharpening Method Based on Deep Nonlocal Unfolding. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–11.

Liu, J. G. 2000. Smoothing Filter-based Intensity Modulation: A spectral preserve image fusion technique for improving spatial details. *International Journal of Remote Sensing*, 21(18): 3461–3472.

Masi, G.; Cozzolino, D.; Verdoliva, L.; and Scarpa, G. 2016. Pansharpening by Convolutional Neural Networks. *Remote Sensing*, 8(7).

Meng, X.; Xiong, Y.; Shao, F.; Shen, H.; Sun, W.; Yang, G.; Yuan, Q.; Fu, R.; and Zhang, H. 2021. A Large-Scale Benchmark Data Set for Evaluating Pansharpening Performance: Overview and Implementation. *IEEE Geoscience and Remote Sensing Magazine*, 9(1): 18–52.

Tan, J.; Huang, J.; Zheng, N.; Zhou, M.; Yan, K.; Hong, D.; and Zhao, F. 2024. Revisiting Spatial-Frequency Information Integration from a Hierarchical Perspective for Panchromatic and Multi-Spectral Image Fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 25922–25931.

Vivone, G.; Dalla Mura, M.; Garzelli, A.; and Pacifici, F. 2021. A Benchmarking Protocol for Pansharpening: Dataset, Preprocessing, and Quality Assessment. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14: 6102–6118.

Wald, L.; Ranchin, T.; and Mangolini, M. 1997. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogrammetric Engineering and Remote Sensing*, 63: 691–699.

Wang, H.; Gong, M.; Mei, X.; Zhang, H.; and Ma, J. 2024. Deep Unfolded Network with Intrinsic Supervision for Pan-Sharpening. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6): 5419–5426.

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.

Xu, S.; Zhang, J.; Zhao, Z.; Sun, K.; Liu, J.; and Zhang, C. 2021. Deep Gradient Projection Networks for Pan-sharpening. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1366–1375.

Yang, G.; Zhou, M.; Yan, K.; Liu, A.; Fu, X.; and Wang, F. 2022. Memory-augmented Deep Conditional Unfolding Network for Pansharpening. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1778–1787.

Yuhas, R. H.; Goetz, A. F. H.; and Boardman, J. W. 1992. Discrimination among semi-arid landscape endmembers using the Spectral Angle Mapper (SAM) algorithm. In *JPL, Summaries of the Third Annual JPL Airborne Geoscience Workshop. Volume 1: AVIRIS Workshop*.

Zhang, J.; He, X.; Yan, K. R.; Cao, K.; Li, R.; Xie, C.; Zhou, M.; and Hong, D. 2024. Pan-Sharpening With Wavelet-Enhanced High-Frequency Information. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–14.

Zhang, Y.; Tiňo, P.; Leonardis, A.; and Tang, K. 2021. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5): 726–742.

Zhou, M.; Huang, J.; Fang, Y.; Fu, X.; and Liu, A. 2022. Pan-Sharpening with Customized Transformer and Invertible Neural Network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3): 3553–3561.

Zhou, M.; Huang, J.; Yan, K.; Hong, D.; Jia, X.; Chanussot, J.; and Li, C. 2024. A General Spatial-Frequency Learning Framework for Multimodal Image Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–18.