This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Memory-augmented Deep Conditional Unfolding Network for Pan-sharpening

Gang Yang[1†], Man Zhou[2,1†], Keyu Yan[2,1], Aiping Liu[1*], Xueyang Fu[1], Fan Wang[2]
[1]University of Science and Technology of China, China
[2]Hefei Institute of Physical Science, Chinese Academy of Sciences, China
{yg1997, manman, keyu}@mail.ustc.edu.cn, {aipingl, xyfu}@ustc.edu.cn, wang_fan6@163.com

Abstract

Pan-sharpening aims to obtain high-resolution multi-spectral (MS) images for remote sensing systems and deep learning-based methods have achieved remarkable success. However, most existing methods are designed in a black-box principle, lacking sufficient interpretability. Additionally, they ignore the different characteristics of each band of MS images and directly concatenate them with panchromatic (PAN) images, leading to severe copy artifacts [9]. To address the above issues, we propose an interpretable deep neural network, namely Memory-augmented Deep Conditional Unfolding Network with two specified core designs. Firstly, considering the degradation process, it formulates the Pan-sharpening problem as the minimization of a variational model with denoising-based prior and non-local auto-regression prior which is capable of searching the similarities between long-range patches, benefiting the texture enhancement. A novel iteration algorithm with built-in CNNs is exploited for transparent model design. Secondly, to fully explore the potentials of different bands of MS images, the PAN image is combined with each band of MS images, selectively providing the high-frequency details and alleviating the copy artifacts. Extensive experimental results validate the superiority of the proposed algorithm against other state-of-the-art methods.

1. Introduction

With the rapid development of remote sensors, increasing satellite images are available for a wide range of applications such as mapping services, military systems, and environmental monitoring. Satellites capture multispectral (MS) and panchromatic (PAN) images simultaneously with complementary information for each modality that PAN images have a high spatial solution and MS images contain rich spectral information [15, 36]. In order to obtain the images with both high spectral and spatial resolution, the Pan-sharpening technique that aims to fuse the MS and PAN images has attracted increasing attention.

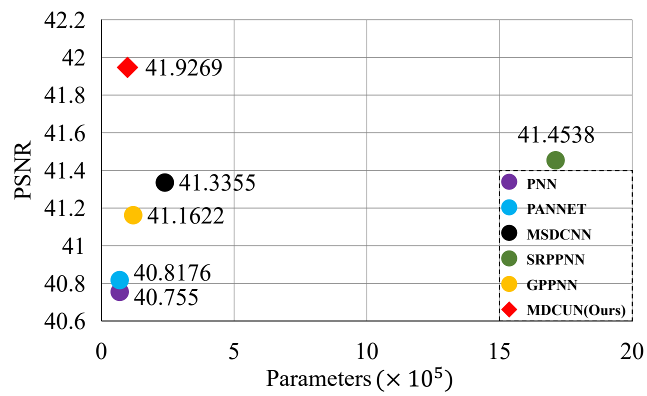


Figure 1. Trade-off between parameters and model performance for Pan-sharpening on WorldViewII dataset.

The past decades have witnessed the explosive growth of research works in the Pan-sharpening field, where the focuses include model-based and deep learning (DL)-based methods. Due to the ill-posed property of Pan-sharpening, the former usually requires hand-crafted priors to regularize the solution space of the latent high-resolution (HR) MS images. However, the limited representation ability of hand-crafted priors results in unsatisfactory performance when processing complex scenes. Besides, the traditional methods are challenging in optimization, limiting their practical applications. Inspired by the success of deep neural networks, various DL-based Pan-sharpening algorithms have been proposed. While they demonstrate superiority in feature representation and model generalization, the long-standing issue that existing DL-based Pan-sharpening methods suffer from is the lack of interpretability as most of them are designed in a black-box principle without considering the rationality of models. Integrating the domain knowledge with interpretable DL-based models is therefore promising for improving the Pan-sharpening performance. Additionally, existing methods ignore the different characteristics of each band of MS images and directly concatenate them with

† Co-first authors contributed equally,* corresponding author.
This work was supported by the USTC Research Funds of the Double First-Class Initiative (Grant YD2100002004, 2100002003).

PAN images as input along channel dimension, which may lead to the severe copy artifacts [9].

Very recently, a few models attempt to incorporate advantages of both model-based and DL-based methods in the image processing community [11, 32, 60]. Inspired by such designs, Xu *et.al* [52] propose the first deep unfolding network for Pan-sharpening. It formulates Pan-sharpening as two separate optimization problems regularized by a deep prior for both PAN and low-resolution (LR) MS images. Nevertheless, the designed implicit priors are still difficult to investigate thoroughly their influence and the potential of cross-stages has not been fully explored.

In summary, existing state-of-the-art (SOTA) methods suffer from two-fold issues: 1) lacking sufficient interpretability, and 2) ignoring the different characteristics of each band of MS images. To this end, in this paper, we propose an interpretable deep unfolding network by combining advantages of both the model-based and data-driven DL methods, namely Memory-augmented Deep Conditional Unfolding Network (**MDCUN**). Considering the degradation process and observing that MS images often contain repetitive structures, we formulate the Pan-sharpening problem as the minimization of a variational model with two newly-designed prior terms, including denoising-based prior and non-local auto-regression prior. Specifically, the former aims to reconstruct the latent MS images while the latter learns the similarities between long-range patches, benefiting the texture enhancement and reducing the aliasing artifacts. Then, a novel effective iteration algorithm with built-in CNNs is exploited for transparent model design to further increase the model interpretability. Moreover, to fully explore the potentials of different bands of MS images, we propose a band-aware PAN-guided high-frequency information extraction module. To be specific, the PAN image is combined with each band of MS image, selectively providing the high-frequency details and alleviating the copy artifacts. Additionally, the contextual memory mechanism is introduced to augment the capacity across iterative stages, therefore facilitating the information interaction. The proposed method is assessed with extensive experiments, and the results demonstrate its superiority qualitatively and quantitatively. The contributions of our work are summarized as follows:

- We formulate the Pan-sharpening as the minimization of a variational model and introduce the denoising-based prior and non-local auto-regression prior to improve the long-range coherence.
- We propose an interpretable deep network, namely Memory-augmented Deep Conditional Unfolding Network, which incorporates advantages of both the model-based and data-driven DL methods.
- A band-aware PAN-guided high-frequency information extraction module is devised to fully explore the potentials of different bands of MS images. The contextual memory mechanism is additionally introduced to augment the capacity across iterative stages, facilitating the information interaction.
- Extensive experiments over different satellite datasets demonstrate that our method outperforms state-of-the-art algorithms with fewer parameters.

2. Related work

2.1. Classic pan-sharpening methods

The classic pan-sharpening methods can be classified into three broad categories, including Component Substitution (CS) [5, 16, 39], Multi-resolution Analysis (MRA) [37, 42], and Variational Optimization (VO) [8, 13, 41]. The common CS methods [5, 16, 39] separate spatial and spectral information from MS images by specific transformations, and then replace the separated spatial components with PAN images. The typical MRA methods [34, 38] complement the high-frequency details extracted by the multi-resolution decomposition techniques from PAN images to the up-sampled MS images. The VO methods [2, 6] are concerned because of the fine fusion effects on Pan-sharpening. They assume that there are certain constraints or prior conditions between the HR MS and PAN images, and establish specific optimization functions based on the proposed conditions, so as to well balance spectral and spatial quality by optimizing the above problems.

2.2. Deep learning based methods

With the highly nonlinear mapping capability of a convolutional neural network, PNN [35] utilizes three convolutional units to map the relationship between PAN, LR MS, and HR MS images, which achieves a significant improvement compared with other classical methods. Inspired by PNN, a large number of DL-based Pan-sharpening studies [4, 49] have emerged recently. For instance, PANNet [53] adopts the residual learning module as in ResNet [18], MSDCNN [54] adds multi-scale modules on the basis of residual connection, and SRPPNN [3] refers to the design idea of SRCNN [10]. Observing that the same object in MS and PAN is not always aligned, Li *et al.* [24] design a SIPSA-Net [24] with a feature alignment module which can align features from PAN and LR MS images. Wu *et al.* [47] utilize multiple parallel branches to integrate features of different scales into the backbone network to improve performance. Aiming at satellite image analysis, Ma *et al.* [33] propose an unsupervised framework based on generative confrontation networks. Additionally, some model-driven CNN models that are similar to our works with clear physical meaning emerge, such as MHNet [48] and GPPNN [52].

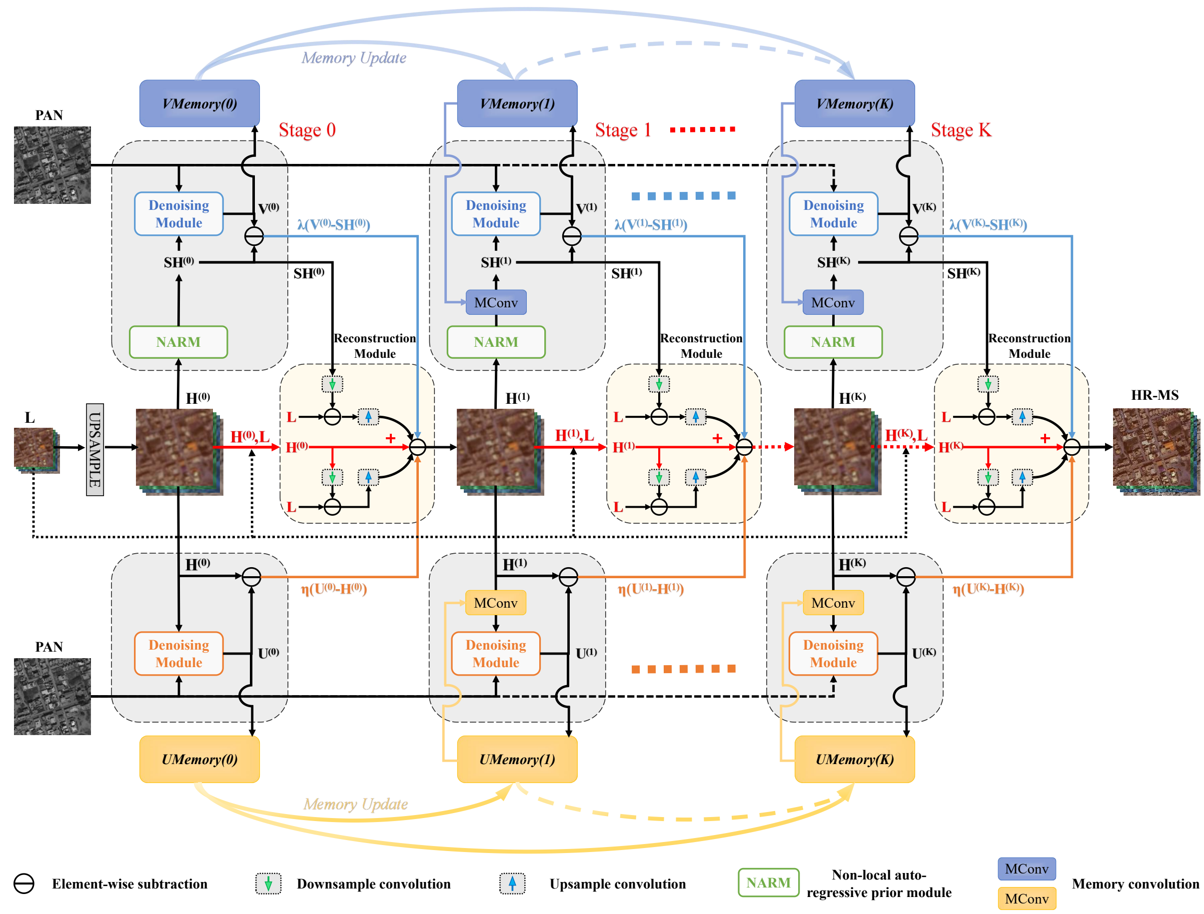


Figure 2. The overall architecture of **MDCUN**.

2.3. Deep unfolding network

In recent years, many researchers [4, 11, 20, 27, 30, 60] attempt to combine domain knowledge with deep neural networks to propose deep unfolding networks which take advantages of the model-based methods' interpretability and learning-based methods' strong mapping ability. Specifically, the deep unfolding network firstly unfolds certain optimization algorithms [1, 7, 14, 28, 29, 31, 32, 40, 50, 51, 58] and utilizes deep neural network to parameterize the unfolding model, then minimizes the loss function on a large training dataset and optimizes the parameters in an end-to-end manner. For example, Zhang *et al.* [56] transform the iterative shrinkage-thresholding algorithm into a deep network form for image compressive sensing. To effectively solve the JPEG compression artifacts removal problem, Fu *et al.* [14] design an alternating minimization algorithm and unfold it into the deep network architecture. Additionally, deep unfolding networks are also proposed in image super-resolution [59], image deblurring [22], snapshot compressive sensing [57, 61] and image demosaicking [21].

3. Methods

3.1. Motivation

In this paper, we formulate the Pan-sharpening as a PAN-guided MS super-resolution problem, in which the process of Pan-sharpening can be denoted as $L = DKH + e_h$, where L denotes the LR MS image through performing the blurring and down-sampling by K and D matrix over the HR MS version H respectively, and e_h denotes the noise. Referring the above observation model, HR MS images can be obtained by solving the minimization problem as:

$$\arg\min_{H} \frac{1}{2} ||L - DKH||_2^2 + \eta\Omega(H, P) \tag{1}$$

where P indicates the PAN images and provides the supplementary information for restoring the HR MS images H. And η is the Lagrange multiplier and $\Omega(H, P)$ describes the regularization function.

Motivated by the observation that remote sensing images contain rich repetitive structures, we utilize a well-established image prior (N prior) obtained from non-local auto-regressive prior model (NARM) to constraint above optimization. Given the MS patches, NARM seeks its sparse linear decomposition over a set of non-local (instead of local) neighborhoods. The NARM can be represented as:

$$H = SH + e_s \tag{2}$$

where the matrix S represents the autoregressive matrix of NARM, e_s is the modeling error of NARM.

By introducing the above NARM, the observation model is rewritten as:

$$L = DK(SH + e_s) = DKSH + n \tag{3}$$

where $n = DKe_s$ is a new modeling error. Therefore, the minimization problem of Eq. 1 is reformulated with:

$$\begin{aligned} \arg\min_{H} \frac{1}{2}\,||L \;-\; DKH||_2^2 \;+\; \frac{\mu}{2}\,||L \;-\; DKSH||_2^2 \\ +\; \eta\Omega_1(H|P) +\; \lambda\Omega_2(SH|P) \end{aligned} \tag{4}$$

where the last two terms correspond to the denoising prior (D prior) and the N prior, respectively.

3.2. Optimization

Following the framework of half-quadratic splitting (HQS) to introduce two auxiliary parameters U and V for H and SH respectively, Eq. 4 can be formulated as a non-constrained optimization problem:

$$\begin{aligned} \arg\min_{H,\,U,\,V} \frac{1}{2}\,||L - DKH||_2^2 + \frac{\mu}{2}\,||L - DKSH||_2^2 \\ + \frac{\eta_1}{2}\,||U - H||_2^2 + \eta_2\Omega_1(U|P) \\ + \frac{\lambda_1}{2}\,||V - SH||_2^2 + \lambda_2\Omega_2(V|P) \end{aligned} \tag{5}$$

where $\eta_1, \eta_2, \lambda_1$ and λ_2 are penalty parameters. To obtain an unrolling inference, Eq. 5 can be divided into the following three sub-problems and solved alternatively:

$$U^{(k)} = \arg\min_{U} \frac{\eta_1}{2}\left|\left|U - H^{(k)}\right|\right|_2^2 + \eta_2\Omega_1(U|P) \tag{6}$$

$$V^{(k)} = \arg\min_{V} \frac{\lambda_1}{2}\left|\left|V - SH^{(k)}\right|\right|_2^2 + \lambda_2\Omega_2(V|P) \tag{7}$$

$$\begin{aligned} H^{(k+1)} = \arg\min_{H} \frac{1}{2}\,||L - DKH||_2^2 \\ + \frac{\mu}{2}\,||L - DKSH||_2^2 \\ + \frac{\eta_1}{2}\left|\left|U^{(k)} - H\right|\right|_2^2 + \frac{\lambda_1}{2}\left|\left|V^{(k)} - SH\right|\right|_2^2 \end{aligned} \tag{8}$$

here, k denotes the HQS iteration index.

Moreover, we employ the proximal gradient projection method to solve the above three sub-problems:

$$U^{(k)} = prox_{\Omega_1}(U^{(k-1)} \;-\; \delta_1\nabla f_1(U^{(k-1)})) \tag{9}$$

$$V^{(k)} = prox_{\Omega_2}(V^{(k-1)} \;-\; \delta_2\nabla f_2(V^{(k-1)})) \tag{10}$$

$$H^{(k+1)} = H^{(k)} \;-\; \delta_3\nabla f_3(H^{(k)}) \tag{11}$$

where $prox_{\Omega_1}(\cdot)$ and $prox_{\Omega_2}(\cdot)$ are proximal operators corresponding to penalty $\Omega_1(\cdot)$ and $\Omega_2(\cdot)$. And the gradient related notations are detailed as:

$$\nabla f_1(U^{(k-1)}) = \eta_1(U^{(k-1)} \;-\; H^{(k)}) \tag{12}$$

$$\nabla f_2(V^{(k-1)}) = \lambda_1(V^{(k-1)} \;-\; SH^{(k)}) \tag{13}$$

$$\begin{aligned} \nabla f_3(H^{(k)}) &= (DK)^T(DKH^{(k)} \;-\; L) \\ &+\; \mu(DK)^T(DKSH^{(k)} \;-\; L) \\ &+\; \eta_1(H^{(k)} \;-\; U^{(k)}) \\ &+\; \lambda_1(SH^{(k)} \;-\; V^{(k)}) \end{aligned} \tag{14}$$

3.3. Deep unfolding network

Inspired by the principle of model-driven deep learning, our deep unfolding network contains K stages, which are intentionally designed to correspond to K iterations in the optimization algorithm as shown in Figure 2. In each network, two auxiliary variables (U and V) are updated firstly, and then the restored image is calculated to update the memory components ($UMemory$ and $VMemory$). To construct a step-by-step corresponding deep unfolding network architecture, we generalize the above iterative step as specified network modules, containing PAN-guided conditional band-aware MS denoise module, non-local auto-regressive prior module, memory-augmented information module, and reconstruction module.

In Figure 2, k-th iteration of HQS is cast to k-th stage of the model, which includes denoise modules (DMs), NARM module, and reconstruction module, as shown below:

$$U^{(k)} = U^{(k-1)} + DM(U^{(k-1)}, H^{(k)}|P) \tag{15}$$

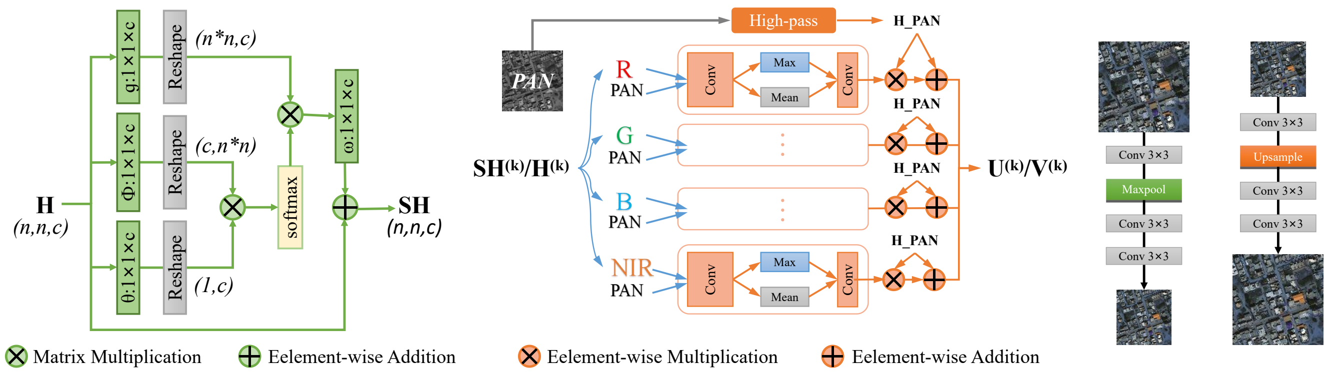
$$SH^{(k)} = NARM(H^{(k)}) \tag{16}$$

$$V^{(k)} = V^{(k-1)} + DM(V^{(k-1)}, SH^{(k)}|P) \tag{17}$$

$$\begin{aligned} H^{(k+1)} = H^{(k)} - \delta[Up(Down(H^{(k)}) - L) \\ + \mu Up(Down(SH^{(k)}) - L) \\ + \eta_1(H^{(k)} - U^{(k)}) \\ + \lambda_1(SH^{(k)} - V^{(k)})] \end{aligned} \tag{18}$$

where **Down** and **Up** represent the down-sampling and up-sampling functions in spatial resolution respectively. The **DM** and **NARM** denoted the Denoise Module and Non-local Auto-Regressive prior Module respectively. Besides, it can be noted that each denoise stage involves the PAN image while depending on previous states. Naturally, the design of the denoise module needs to consider the memory mechanism and condition-served PAN image.

To be specific, inspecting the k-th stage, the PAN-guided module is responsible for updating the two auxiliary variables $U^{(k)}$ and $V^{(k)}$ while the non-local auto-regressive



(a) (b) (c) (d)

Figure 3. Architectures of MDCUN's submodules. (a) The architecture of the **non-local auto-regressive prior module (NARM)**, (b) The inner structure of the **PAN-guided conditional band-aware MS denoise module**, (c) The inner structure of the **down-sampling-blocks (Down)** in reconstruction module, and (d) The inner structure of the **Up-sampling-blocks (Up)** in reconstruction module.

prior module aims to calculate the NARM matrix S for updating the corresponding $SH^{(K)}$. The memory-augmented information module takes the outputs $U^{(0)}, ..., U^{(k-1)}$ and $V^{(0)}, ..., V^{(k-1)}$ of denoise modules as input across long-range stages to facilitate the information flow. The reconstruction module corresponds to Eq. 18 to update the restored $H^{(k)}$. The updated $H^{(k)}$ is fed into the next stage and performs the repetitive operation until the stage number reaches K. We will elaborate each module next.

Non-local auto-regressive prior module

As we discussed in Sec. 3.1, NARM seeks sparse linear decomposition over a set of non-local neighborhoods. Follow [12], the pixel H_i can be approximately weighted by its nonlocal neighbors (including itself):

$$H_i \approx \sum_j \boldsymbol{\omega}_i^j H_i^j \tag{19}$$

where H_i^j represents the j-th nonlocal neighbor of H_i. And w_i^j is solved by the following optimization problems:

$$\widetilde{\omega_i} = \underset{\boldsymbol{\omega}_i}{argmin} ||H_i - H\boldsymbol{\omega}_i||_2^2 + \gamma||\boldsymbol{\omega}_i||_2^2 \tag{20}$$

where $H = [H_i^1, H_i^2, ..., H_i^J], \boldsymbol{\omega}_i = [\boldsymbol{\omega}_i^1, \boldsymbol{\omega}_i^2, ..., \boldsymbol{\omega}_i^J]$, and J represents the first J most similar nonlocal neighbors to H_i are chosen. γ represents the regularization parameter.

Based on determined coefficients ω_i, the formula of NARM matrix S in Eq. 2 is expressed by:

$$S_{i,j} = \begin{cases} \omega_i^j, \ H_j \ \ is \ a \ nonlocal \ neighbor \ of \ H_i \\ 0, otherwise \end{cases} \tag{21}$$

Calculating similarity among the nonlocal neighbors in Eq. 2 can be implemented by nonlocal networks [12, 45]. The output of NARM (SH) is expressed by:

$$SH_i = \frac{\sum_{\forall j} f(H_i, H_j) g(H_j)}{\sum_{\forall j} f(H_i, H_j)} \tag{22}$$

where similarity function $f(\cdot, \cdot)$ calculates the relationship between H_i and H_j. And the architecture of NARM is shown in Figure 3(a).

PAN-guided band-aware MS denoise module

As for the MS image enhancement problem, it is crucial to effectively exploit the intrinsic relations between the high-pass PAN images and all bands of the MS image with different bands. As shown in Figure 3(b), we introduce a high-pass modification block to learn the high-pass information, which can be used to enhance the spatial information of each band in MS, so as to achieve the purpose of denoising.

With the output of k-th stage network $H^{(k)}$ and the output of NARM $SH^{(k)}$, we consider the D prior and the N prior and take PAN images as the condition in Eq. 4. PAN-guided band-aware MS denoise module can be implemented by the denoising module (DM) guided by Eq. 6 and Eq. 7, where the output of DM ($U^{(k-1)}$ or $V^{(k-1)}$) in the previous stage, $H^{(k)}$ and condition P are used as the input of k-th stage of MDCUN, as shown in Eq. 15 and Eq. 17.

Memory-augmented information module

In this paper, considering the memory information in Eq. 15 and Eq. 17 and making full use of memory information generated by the model, we introduce memory components to store the memory information and keep the memory information updated. The memory components mainly store memory information of two kinds of priors.

As shown in Figure 2, in the input of k-th stage of PAN-guided band-aware MS denoise module, the outputs of DM ($U^{(k-1)}$ and $V^{(k-1)}$) in the previous stage will be replaced by memory components ($UMemory$ and $VMemory$), so the inputs of DM are the memory components, $H^{(k)}$ and condition P, so we have:

$$U^{(k)} = DM(UMemory, H^{(k)}, P) \tag{23}$$

$$V^{(k)} = DM(VMemory, H^{(k)}, P) \tag{24}$$

Table 1. The four evaluation metrics on the test 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.

Methods	Params	WorldView II				WorldView III				GaoFen2			
		PSNR ↑	SSIM ↑	SAM ↓	ERGAS ↓	PSNR ↑	SSIM ↑	SAM ↓	ERGAS ↓	PSNR ↑	SSIM ↑	SAM ↓	ERGAS ↓
SFIM	-	34.1297	0.8975	0.0439	2.3449	21.8212	0.5457	0.1208	8.973	36.906	0.8882	0.0318	1.7398
Brovey	-	35.8646	0.9216	0.0403	1.8238	22.5060	0.5466	0.1159	8.2331	37.7974	0.9026	0.0218	1.372
GS	-	35.6376	0.9176	0.0423	1.8774	22.5608	0.547	0.1217	8.2433	37.226	0.9034	0.0309	1.6736
IHS	-	32.1601	**0.9812**	10.3010	26.40	22.5579	0.5354	0.1266	8.3616	38.1754	0.9100	0.0243	1.5336
GFPCA	-	34.5581	0.9038	0.0488	2.1411	22.3344	0.4826	0.1294	8.3964	37.9443	0.9204	0.0314	1.5604
PNN	0.689	40.7550	0.9624	0.0259	1.0646	29.9418	0.9121	0.0824	3.3206	43.1208	0.9704	0.0172	0.8528
PANNET	0.688	40.8176	0.9626	0.0257	1.0557	29.6840	0.9072	0.0851	3.4263	43.0659	0.9685	0.0178	0.8577
MSDCNN	2.390	41.3355	0.9664	0.0242	0.994	30.3038	0.9184	0.0782	3.1884	45.6874	0.9827	0.0135	0.6389
SRPPNN	17.114	41.4538	0.9679	0.0233	0.9899	30.4346	0.9202	0.0770	3.1553	47.1998	0.9877	0.0106	0.5586
GPPNN	1.198	41.1622	0.9684	0.0244	1.0315	30.1785	0.9175	0.0776	3.2593	44.2145	0.9815	0.0137	0.7361
Ours	0.983	**41.9269**	0.9722	**0.0215**	**0.9050**	**30.5668**	**0.9227**	**0.0744**	**3.0987**	**47.2023**	**0.9879**	**0.0105**	**0.5533**

With the outputs $U^{(k)}$ and $V^{(k)}$ of PAN-guided band-aware MS denoise module, we input them into two different memory components respectively and complete the update of memory information in the memory components. In k-th stage, taking into account the two outputs $U^{(k)}$ and $V^{(k)}$ of PAN-guided band-aware MS denoise module, the element in $UMemory$ is $\{U^{(0)},\ U^{(1)}, ...,\ U^{(k)}\}$, and the element in $VMemory$ is $\{V^{(0)},\ V^{(1)}, ...,\ V^{(k)}\}$.

Reconstruction module

With $H^{(k)}$, SH^k, U^k and V^k, we can iteratively reconstruct the value of $H^{(k+1)}$ according to Eq. 11 and Eq. 14.

The operators $(DK)^T$ and DK are simulated using a convolution network layer respectively. Specifically, DK is simulated by a network call down-sampling-blocks ($Down$) consisting of a convolutional layer with 3×3 kernels and 64 channels, a maxpool layer to decrease the spatial resolution and two convolutional layers with 3×3 kernels for reprojection to the original dimension as shown in Figure 3(c). Similarly, the $(DK)^T$ is simulated by a network call Upsampling-blocks (Up) consisting of a convolutional layer with 3×3 kernels and 64 channels, a upsample layer to increase the spatial resolution, and two convolutional layers with 3×3 kernels for reprojection to the original dimension as shown in Figure 3(d).

4. Experiments

4.1. Datasets and evaluation metrics

In our experiments, remote sensing images obtained on three satellites are used, including WorldViewII, WorldViewIII, and GaoFen2. For each dataset, we have hundreds of image pairs, and the MS images are cropped into patches with the size of 32×32, and the size of corresponding PAN images is 128×128. For numerical stability, each patch is normalized by dividing the maximum value to make the pixels range from 0 to 1. Four widely used image quality assessment metrics are used to evaluate the performance, including the peak signal-to-noise ratio (PSNR) [19], Structural similarity (SSIM) [46], Erreur Relative Globale Adimensionnelle de Synthese (ERGAS) [43], Spectral angle mapper (SAM) [55], etc. The first three metrics measure the spatial distortion and the fourth one measures the spectral distortion. An image is better if its PSNR and SSIM are higher, and SAM and ERGAS are lower.

4.2. Implementation details

MDCUN is supervised by the l_1 loss between the output $H^{(K)}$ of MDCUN and the ground truth H. As the paired training samples are not available, we construct the training datasets using the Wald protocol [44] to generate paired images. Thanks to the parameter sharing across K stages, the overall model can be trained in an end-to-end manner. To further reduce the number of parameters and avoid over-fitting, we enforce two PAN-guided band-aware MS denoise modules to share the same parameters.

Training Setting: The implementation is based on Pytorch framework. For optimization, we employ an ADAM optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$ to update the network parameters for 1000 epochs with a batch size of 4. The initial learning rate is set to be $5e-04$ and decreases by half for every 200 epochs.

Reproducibility: All experiments are conducted on a TITAN RTX GPU with 24GB memory. And code is available in `https://github.com/yggame/MDCUN`.

4.3. Comparison with SOTA methods

We compare MDCUN with ten competitive methods, which include five classical methods (SFIM [26], Brovey [16], GS [23], IHS [17], and GFPCA [25]) and five DL-based methods (PNN [35], PANNET [53], MSDCNN [54], SRPPNN [3], and GPPNN [52]).

Quantitative results: The evaluation metrics on three datasets of 10 benchmark methods are reported in Table 1 where the best and the second best values are highlighted by bold and underline, respectively. It is clear to see that our method achieves the best performance on three satellites. This substantiates the effectiveness and flexibility of our method with a certain degree of generalization.

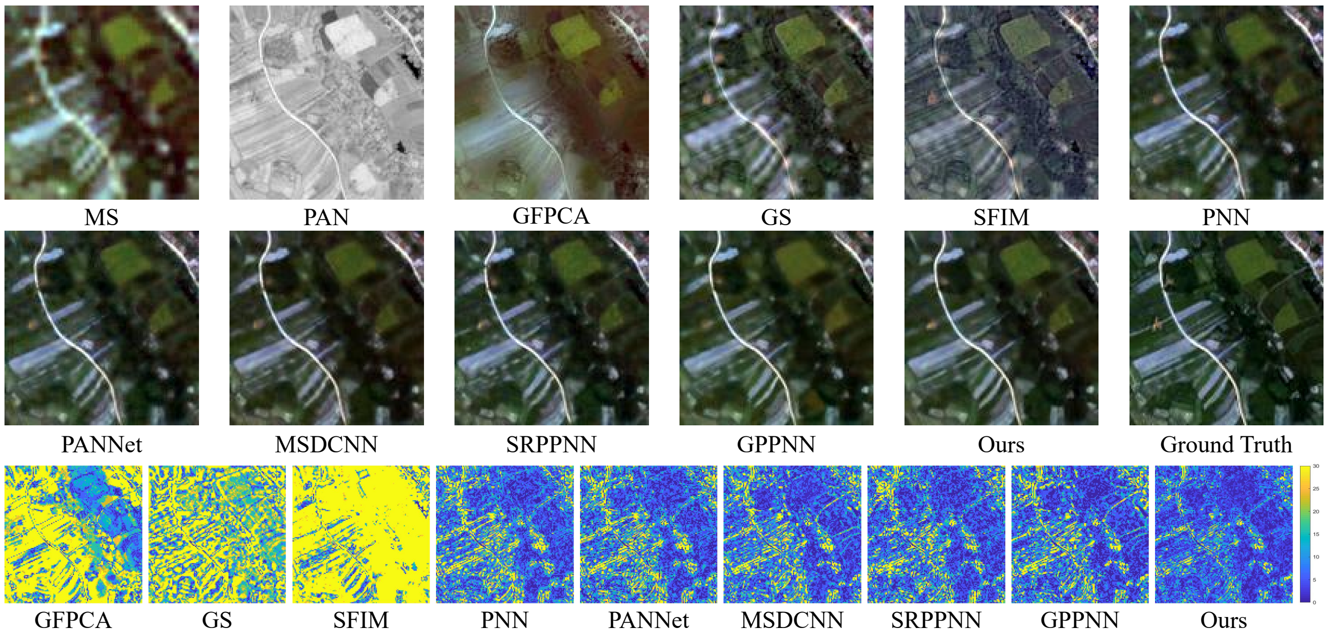


Figure 4. Qualitative comparison of all methods on WorldViewII. The last row visualizes the MSE residues between the Pan-sharpening results and the ground truth.

Table 2. The results of different configurations on WorldViewII. 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. (PS: Parameters Sharing)

Configuration	PS (inter stage)	PS (intra stage)	Memory	D prior	N prior	PSNR ↑	SSIM ↑	SAM ↓	ERGAS ↓
I	×	×	✓	✓	✓	41.9165	0.9719	0.0215	0.9062
II	✓	×	✓	✓	✓	**42.0412**	**0.9728**	**0.0212**	**0.8925**
III	×	✓	✓	✓	✓	41.8951	0.9722	0.0215	0.9082
IV	✓	✓	×	✓	✓	41.8464	0.9716	0.0217	0.9127
V	✓	✓	✓	×	×	36.2105	0.9056	0.0317	1.6121
VI	✓	✓	✓	✓	×	41.8036	0.9717	0.0217	0.9187
VII	✓	✓	✓	×	✓	41.8156	0.9721	0.0215	0.9050
MDCUN(Ours)	✓	✓	✓	✓	✓	41.9269	0.9722	0.0215	0.9050

Qualitative results: The qualitative results are demonstrated in Figure 4. It can be seen that our model recovers the images with fewer visible artifacts. The quality improvement achieved by MDCUN may be due to the fully usage of the feature maps from former stages to refine the final results. Additionally, the intermediate visual results of MDCUN with different stages are shown in Figure 5, from which we can observe that more detailed information is recovered along with greater number of stages.

4.4. Ablation study

To further verify the performance of our proposed method under different configurations, a series of ablation studies are carried out, including 1) Effects of the number of stages; 2) Reasonability of parameter sharing; 3) Effectiveness of memory, and 4) Influence of different priors.

Effects of the number of stages: To explore the impact of the number of unfolded stages on the performance, we experiment with varying numbers of stages K. Table 3 shows the results of different K from 1 to 6. It can be seen that the PSNR performance increases as the number of stages increases. We choose $K = 4$ in our implementation to balance the performance and computational complexity.

Reasonability of parameter sharing: We evaluate the scenario where the parameters are not shared when $K = 4$. In other words, MDCUN only contains a denoising module, a NARM, and a reconstruction module. The reasonability of parameter sharing is verified by the comparative experiments of the following two cases: 1) Parameters sharing in inter-stage; 2) Parameters sharing in intra-stage. As shown in Table 2(I-III), disabling parameter sharing in intra-stage improves performance to some extent, but parameter shar-

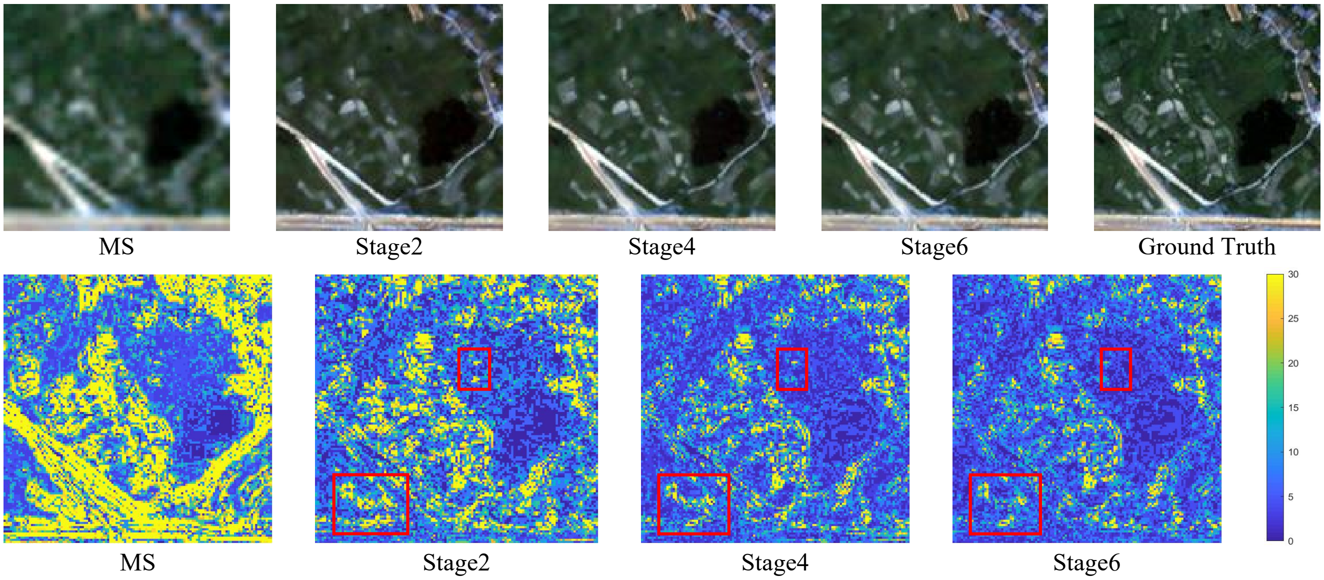


Figure 5. Intermediate visual results of different stages of MDCUN on WorldViewII. The last row visualizes the MSE residues between the Pan-sharpening results and the ground truth.

Table 3. The PSNR values of MDCUN with different number of stages on WorldViewII. 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.

Stages (K)	PSNR ↑	SSIM ↑	SAM ↓	ERGAS ↓
1	41.6093	0.9689	0.0229	0.9518
2	41.7395	0.9696	0.0225	0.9462
3	41.8234	0.9716	0.0217	0.9086
4	41.9269	0.9722	0.0215	0.9050
5	42.1424	0.9723	**0.0213**	**0.9014**
6	**42.1512**	**0.9724**	0.0214	0.9042

ing is a good strategy compared with the cost of more parameters. While disabling parameter sharing in inter-stage will weaken our network's performance.

Effectiveness of memory: We additionally perform a comparative experiment to verify the effectiveness of memory components. In our ablation study, the input of the k-th stage of DM is the DM output of the previous stage, rather than being replaced by memory components. As shown in Table 2(IV), the memory component is an effective strategy for improving performance.

Influence of different priors: Two different priors, denoising-based prior (D prior) and non-local auto-regression prior (N prior) are utilized in the proposed model. We therefore conduct ablation studies to investigate the influence of different priors. As demonstrated in Table 2(V-VII), the best performance is achieved when utilizing both two priors.

4.5. Cost-performance trade-off

To evaluate the trade-off between the cost (in terms of the number of parameters) and the performance (represented by PSNR), we compare the proposed method against five deep learning methods in Figure 1. The results demonstrate that our method can achieve better PSNR performance and a good trade-off between cost and performance compared to those of other deep learning-based methods.

4.6. Limitation

There are still several limitations. Due to the variability of different satellites, our method may not completely guarantee the superior performance over other methods on all datasets. Meanwhile, we need to train the model on each dataset individually, without examining the generalization ability when directly applying the trained model to another dataset. Additionally, we choose the number of stages in our model as 4. There is a large number of flops, which increases with the increase of the number of stages.

5. Conclusion and future work

In this paper, we propose a Memory-augmented Deep Conditional Unfolding Network that is both explainable and efficient. We formulate the Pan-sharpening problem as the minimization of a variational model with two beneficial priors. Extensive experiments demonstrate the superiority of the proposed method against other state-of-the-art models qualitatively and quantitatively. In future, we will apply our framework to more image tasks and achieve the generalization on more datasets.

References

[1] Manya V Afonso, José M Bioucas-Dias, and Mário AT Figueiredo. Fast image recovery using variable splitting and constrained optimization. *IEEE TIP*, 19(9):2345–2356, 2010. 3

[2] Coloma Ballester, Vicent Caselles, Laura Igual, Joan Verdera, and Bernard Rougé. A variational model for p+ xs image fusion. *IJCV*, 69(1):43–58, 2006. 2

[3] Jiajun Cai and Bo Huang. Super-resolution-guided progressive pansharpening based on a deep convolutional neural network. *IEEE TGRS*, 2020. 2, 6

[4] Xiangyong Cao, Xueyang Fu, Danfeng Hong, Zongben Xu, and Deyu Meng. Pancsc-net: A model-driven deep unfolding method for pansharpening. *IEEE TGRS*, pages 1–13, 2021. 2, 3

[5] Wjoseph Carper, Thomasm Lillesand, and Ralphw Kiefer. 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. 2

[6] Chen Chen, Yeqing Li, Wei Liu, and Junzhou Huang. Sirf: Simultaneous satellite image registration and fusion in a unified framework. *IEEE TIP*, 24(11):4213–4224, 2015. 2

[7] Liang Chen, Jiawei Zhang, Jinshan Pan, Songnan Lin, Faming Fang, and Jimmy S. Ren. Learning a non-blind deblurring network for night blurry images. In *CVPR*, pages 10542–10550, June 2021. 3

[8] Liang-Jian Deng, Gemine Vivone, Weihong Guo, Mauro Dalla Mura, and Jocelyn Chanussot. A variational pansharpening approach based on reproducible kernel hilbert space and heaviside function. *IEEE TIP*, 27(9):4330–4344, 2018. 2

[9] Xin Deng and Pier Luigi Dragotti. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE TPAMI*, 43(10):3333–3348, 2021. 1, 2

[10] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE TPAMI*, 38(2):295–307, 2015. 2

[11] Weisheng Dong, Peiyao Wang, Wotao Yin, Guangming Shi, Fangfang Wu, and Xiaotong Lu. Denoising prior driven deep neural network for image restoration. *IEEE TPAMI*, 41(10):2305–2318, 2018. 2, 3

[12] Weisheng Dong, Lei Zhang, Rastislav Lukac, and Guangming Shi. Sparse representation based image interpolation with nonlocal autoregressive modeling. *IEEE TIP*, 22(4):1382–1394, 2013. 5

[13] Xueyang Fu, Zihuang Lin, Yue Huang, and Xinghao Ding. A variational pan-sharpening with local gradient constraints. In *CVPR*, pages 10265–10274, 2019. 2

[14] Xueyang Fu, Zheng-Jun Zha, Feng Wu, Xinghao Ding, and John Paisley. Jpeg artifacts reduction via deep convolutional sparse coding. In *ICCV*, pages 2501–2510, 2019. 3

[15] Pedram Ghamisi, Behnood Rasti, Naoto Yokoya, Qunming Wang, Bernhard Hofle, Lorenzo Bruzzone, Francesca Bovolo, Mingmin Chi, Katharina Anders, Richard Gloaguen, Peter M. Atkinson, and Jon Atli Benediktsson. Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art. *IEEE GRSM*, 7(1):6–39, 2019. 1

[16] 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. 2, 6

[17] R Haydn. Application of the ihs color transform to the processing of multisensor data and image enhancement. In *Proc. of the International Symposium on Remote Sensing of Arid and Semi-Arid Lands, Cairo, Egypt, 1982*, 1982. 6

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2

[19] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008. 6

[20] Jian Zhang Jiechong Song, Bin Chen. Memory-augmented deep unfolding network for compressive sensing. In *ACM MM*, 2021. 3

[21] Filippos Kokkinos and Stamatios Lefkimmiatis. Deep image demosaicking using a cascade of convolutional residual denoising networks. In *ECCV*, September 2018. 3

[22] Jakob Kruse, Carsten Rother, and Uwe Schmidt. Learning to push the limits of efficient fft-based image deconvolution. In *ICCV*, pages 4596–4604, 2017. 3

[23] Craig A Laben and Bernard V Brower. Process for enhancing the spatial resolution of multispectral imagery using pansharpening, Jan. 4 2000. US Patent 6,011,875. 6

[24] Jaehyup Lee, Soomin Seo, and Munchurl Kim. Sipsa-net: Shift-invariant pan sharpening with moving object alignment for satellite imagery. In *CVPR*, pages 10166–10174, 2021. 2

[25] Wenzhi Liao, Xin Huang, Frieke Van Coillie, Guy Thoonen, Aleksandra Pižurica, Paul Scheunders, and Wilfried Philips. Two-stage fusion of thermal hyperspectral and visible rgb image by pca and guided filter. In *2015 7th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pages 1–4. Ieee, 2015. 6

[26] 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. 6

[27] Risheng Liu, Zhiying Jiang, Xin Fan, and Zhongxuan Luo. Knowledge-driven deep unrolling for robust image layer separation. *IEEE TNNLS*, 31(5):1653–1666, 2019. 3

[28] Risheng Liu, Zhiying Jiang, Xin Fan, and Zhongxuan Luo. Knowledge-driven deep unrolling for robust image layer separation. *IEEE Trans. Neural Networks Learn. Syst.*, 31(5):1653–1666, 2020. 3

[29] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *CVPR*, pages 10561–10570, June 2021. 3

[30] Risheng Liu, Long Ma, Yuxi Zhang, Xin Fan, and Zhongxuan Luo. Underexposed image correction via hybrid priors navigated deep propagation. *IEEE TNNLS*, 2021. 3

[31] Risheng Liu, Yuxi Zhang, Shichao Cheng, Zhongxuan Luo, and Xin Fan. A deep framework assembling principled modules for cs-mri: Unrolling perspective, convergence behaviors, and practical modeling. *IEEE TMI*, 39(12):4150–4163, 2020. 3

[32] Yang Liu, Jinshan Pan, Jimmy S. J. Ren, and Zhixun Su. Learning deep priors for image dehazing. In *ICCV*, pages 2492–2500. IEEE, 2019. 2, 3

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

[34] SG Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE TPAMI*, 11(7):674–693, 1989. 2

[35] Giuseppe Masi, Davide Cozzolino, Luisa Verdoliva, and Giuseppe Scarpa. Pansharpening by convolutional neural networks. *Remote Sensing*, 8(7):594, 2016. 2, 6

[36] Xiangchao Meng, Yiming Xiong, Feng Shao, Huanfeng Shen, Weiwei Sun, Gang Yang, Qiangqiang Yuan, Randi Fu, and Hongyan Zhang. A large-scale benchmark data set for evaluating pansharpening performance: Overview and implementation. *IEEE GRSM*, 9(1):18–52, 2021. 1

[37] Jorge Nunez, Xavier Otazu, Octavi Fors, Albert Prades, Vicenc Pala, and Roman Arbiol. Multiresolution-based image fusion with additive wavelet decomposition. *IEEE TGRS*, 37(3):1204–1211, 1999. 2

[38] Robert A Schowengerdt. Reconstruction of multispatial, multispectral image data using spatial frequency content. *Photogrammetric Engineering and Remote Sensing*, 46(10):1325–1334, 1980. 2

[39] Vijay P Shah, Nicolas H Younan, and Roger L King. An efficient pan-sharpening method via a combined adaptive pca approach and contourlets. *IEEE TGRS*, 46(5):1323–1335, 2008. 2

[40] Dror Simon and Michael Elad. Rethinking the CSC model for natural images. In *NeurIPS*, pages 2271–2281, 2019. 3

[41] Xin Tian, Yuerong Chen, Changcai Yang, and Jiayi Ma. Variational pansharpening by exploiting cartoon-texture similarities. *IEEE TGRS*, pages 1–16, 2021. 2

[42] Gemine Vivone, Luciano Alparone, Jocelyn Chanussot, Mauro Dalla Mura, Andrea Garzelli, Giorgio A Licciardi, Rocco Restaino, and Lucien Wald. A critical comparison among pansharpening algorithms. *IEEE TGRS*, 53(5):2565–2586, 2014. 2

[43] Lucien Wald. *Data fusion: definitions and architectures: fusion of images of different spatial resolutions*. Presses des MINES, 2002. 6

[44] 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. 6

[45] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the CVPR*, pages 7794–7803, 2018. 5

[46] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 6

[47] Xiao Wu, Ting-Zhu Huang, Liang-Jian Deng, and Tian-Jing Zhang. Dynamic cross feature fusion for remote sensing pansharpening. In *Proceedings of the IEEE/CVF ICCV*, pages 14687–14696, October 2021. 2

[48] Qi Xie, Minghao Zhou, Qian Zhao, Deyu Meng, Wangmeng Zuo, and Zongben Xu. Multispectral and hyperspectral image fusion by ms/hs fusion net. In *CVPR*, pages 1585–1594, 2019. 2

[49] Han Xu, Jiayi Ma, Zhenfeng Shao, Hao Zhang, Junjun Jiang, and Xiaojie Guo. Sdpnet: A deep network for pan-sharpening with enhanced information representation. *IEEE TGRS*, 59(5):4120–4134, 2021. 2

[50] Shuang Xu, Ouafa Amira, Junmin Liu, Chun-Xia Zhang, Jiangshe Zhang, and Guanghai Li. Ham-mfn: Hyperspectral and multispectral image multiscale fusion network with rap loss. *IEEE TGRS*, 58(7):4618–4628, 2020. 3

[51] Shuang Xu, Lizhen Ji, Zhe Wang, Pengfei Li, Kai Sun, Chunxia Zhang, and Jiangshe Zhang. Towards reducing severe defocus spread effects for multi-focus image fusion via an optimization based strategy. *IEEE Transactions on Computational Imaging*, 6:1561–1570, 2020. 3

[52] Shuang Xu, Jiangshe Zhang, Zixiang Zhao, Kai Sun, Junmin Liu, and Chunxia Zhang. Deep gradient projection networks for pan-sharpening. In *CVPR*, pages 1366–1375, June 2021. 2, 6

[53] Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, and John Paisley. Pannet: A deep network architecture for pan-sharpening. In *ICCV*, pages 5449–5457, 2017. 2, 6

[54] 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. 2, 6

[55] Roberta H Yuhas, Alexander FH Goetz, and Joe W Boardman. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. In *Proc. Summaries 3rd Annu. JPL Airborne Geosci. Workshop*, volume 1, pages 147–149, 1992. 6

[56] Jian Zhang and Bernard Ghanem. Ista-net: Interpretable optimization-inspired deep network for image compressive sensing. In *CVPR*, pages 1828–1837, 2018. 3

[57] Jian Zhang and Bernard Ghanem. Ista-net: Interpretable optimization-inspired deep network for image compressive sensing. In *CVPR*, pages 1828–1837, 2018. 3

[58] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *CVPR*, pages 3217–3226, 2020. 3

[59] K. Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. *CVPR*, pages 3214–3223, 2020. 3

[60] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *CVPR*, pages 3929–3938, 2017. 2, 3

[61] Chong Mou Zhuoyuan Wu, Jian Zhang. Dense deep unfolding network with 3d-cnn prior for snapshot compressive sensing. In *ICCV*, 2021. 3