Memory-Augmented Model-Driven Network for Pansharpening

Keyu Yan[1,2⋆], Man Zhou[1,2⋆✉], Li Zhang[1,2], and Chengjun Xie[1✉]

[1] Hefei Institute of Physical Science, Chinese Academy of Sciences, China
[2] University of Science and Technology of China, China
{keyu,manman,zanly20}@mail.ustc.edu.cn, cjxie@iim.ac.cn

Abstract. In this paper, we propose a novel memory-augmented model-driven deep unfolding network for pan-sharpening. First, we devise the maximal a posterior estimation (MAP) model with two well-designed priors on the latent multi-spectral (MS) image, i.e., global and local implicit priors to explore the intrinsic knowledge across the modalities of MS and panchromatic (PAN) images. Second, we design an effective alternating minimization algorithm to solve this MAP model, and then unfold the proposed algorithm into a deep network, where each stage corresponds to one iteration. Third, to facilitate the signal flow across adjacent iterations, the persistent memory mechanism is introduced to augment the information representation by exploiting the Long short-term memory unit in the image and feature spaces. With this method, both the interpretability and representation ability of the deep network are improved. Extensive experiments demonstrate the superiority of our method to the existing state-of-the-art approaches. The source code is released at `https://github.com/Keyu-Yan/MMNet`.

Keywords: Pan-sharpening, Maximal a posterior estimation model, Deep unfolding method, Memory mechanism

1 Introduction

Nowadays, Many researchers are concerned on pan-sharpening since it is a critical image processing technology in the domain of remote sensing. Current remote satellites are often outfitted with two types of imaging sensors: multi-spectral and panchromatic sensors, which separately produce low-spatial resolution multi-spectral (LRMS) and panchromatic (PAN) images. Nonetheless, due to technological limits of imaging equipment, providing high-spatial resolution multi-spectral (HRMS) images, which is needed in practice, is difficult. Pan-sharpening attempts to construct the HRMS version by fusing the LRMS image and PAN image.

The main challenge of the Pan-sharpening task is how to recover more spatial features and keep more comprehensive spectral information. In the past

⋆ Co-first authors contributed equally, ✉ corresponding authors.

decades, many traditional techniques, such as component substitution (CS), multi-resolution analysis (MRA), and variational optimization (VO) methodologies, have been put forth. When processing complex scenes, the limited representational capacity of traditional approaches yields disappointing results. Recently, Pan-sharpening techniques based on deep learning (DL) [34,18] have seen considerable success. However, the majority of DL-based approaches currently in use simply concentrate on constructing deeper and more sophisticated network topologies in a black-box fashion without considering the rationality of models, and ignore the inherent information that exists between various modalities.

To increase the interpretability, some model-driven CNN models with clear physical meaning have emerged. The basic idea is to adopt prior knowledge to formulate optimization problems for computer vision tasks such as denoising [4], image compressive [41], then unfold the optimization algorithms into deep neural modules. Motivated by such designs [8,10], Xu *et al.* [37] propose the first deep unfolding network for Pan-sharpening. It formulates Pan-sharpening as two separate optimization problems and each stage in the implementation corresponds to one iteration in optimization. However, the potential of cross-stages has not been fully explored and feature transformation between adjacent stages with reduced channel numbers leads to information loss, which further hinders their performance improvements.

We present a novel interpretable memory-augmented model-driven network for Pan-sharpening to address the aforementioned issues. To be more specific, we first build a variational model from a maximal a posterior (MAP) framework to define the Pan-sharpening problem with two well-designed priors, namely, local and global implicit priors. The local implicit prior implicitly models the relationship between the HRMS and the PAN image from a local perspective, and so can assist in capturing the local relevant information between the HRMS and the PAN image. The global implicit prior addresses the non-local auto-regression property between the two images from a global perspective, allowing for effective use of the global correlation between the two images. Because the scene in the HRMS and PAN images is nearly identical, both images include repetitively similar patterns, which corresponds to the motivation of the developed non-local auto-regression prior. Second, we present an alternating minimization algorithm to unfold the variational model with cascading stages. Each stage has three interrelated sub-problems, and each module corresponds to an iterative algorithm operator. Fig. 1 shows a specific flowchart with colored sub-problems. To facilitate signal flow between iterative stages, persistent memory enhances information representation by using the long-short term unit. In the three sub-problems, each iterative module's output feature maps are picked and combined for the next iterative step, promoting information fusion and decreasing information loss. With this method, both the interpretability and representation ability of the deep network are improved.

Our main contributions are summarized in the following three aspects:

1. By extending the iterative algorithm into a multistage solution that combines the benefits of both model-based and data-driven deep learning techniques,

we present a novel interpretable memory-augmented model-driven network (MMNet) for Pan-sharpening. The interpretation of the deep model is enhanced by such a design.
2. To address the significant information loss problem in the signal flow, we suggest a new memory mechanism which is orthogonal to signal flow and develop a non-local cross-modality module. Such a design enhances the deep model's capacity for representation.
3. Extensive experiments over three satellite datasets demonstrate that our proposed network outperforms other state-of-the-art algorithms both qualitatively and quantitatively.

2 Related Work

2.1 Traditional Methods

By observing the existing Pan-sharpening methods, we can roughly divide the traditional methods into three main categories: CS-based methods, MRA-based methods and VO-based methods.

The CS-based approaches separate spatial and spectral information from the LRMS image and replace spatial information with a PAN image. Intensity hue-saturation (IHS) fusion [5], the principal component analysis (PCA) methods [21,32], Brovey transforms [14], and the Gram-Schmidt (GS) orthogonalization method [22] are common CS-based approaches. These CS-based approaches are rapid since LRMS images simply need spectral treatment to remove and replace spatial components, but the resultant HRMS show severe spectral distortion.

The MRA-based methods inject high-frequency features of PAN derived by multi-resolution decomposition techniques into upsampled multi-spectral images. Decimated wavelet transform (DWT) [26], high-pass filter fusion (HPF) [31], Laplacian pyramid (LP) [34] , the smoothing filter-based intensity modulation (SFIM) [25] and atrous wavelet transform (ATWT) [28] are typical MRA-based methods that reduce spectral distortion and improve resolution, but they rely heavily on multi-resolution technique, which can cause local spatial artifacts.

Recent years, the VO-based methods are concerned because of the fine fusion effect on ill-posed problems. P+XS pan-sharpening approach [2] firstly assumes that PAN image is derived from the linear combination of various bands of HRMS, whereas the LRMS image is from the blurred HRMS. However, the conditions for the assumption linear combinatorial relationship are untenable. To avoid original drawbacks, constraints such as dynamic gradient sparsity property (SIRF) [6], local gradient constraint (LGC) [9], and group low-rank constraint for texture similarity (ADMM) [33] were introduced. These various constraints requiring the manual setting of parameters can only inadequately reflect the limited structural relations of the images, which can also result in degradation.

2.2 Deep Learning Based Methods

With the excellent capabilities for nonlinear mapping learning and feature extraction, (DL)-based methods have rapidly been widely used for Pan-sharpening.

PNN [27] tries to use three convolutional units to directly map the relationship between PAN, LRMS and HRMS images. Inspired by PNN, a large number of pan-sharpening studies based on deep learning have emerged. For example, PANNet [38] adopts the residual learning module in Resnet [16], MSDCNN [39] adds multi-scale modules on the basis of residual connection, and SRPPNN [3] refers to the design idea of SRCNN [7] in the field of image super-resolution. Fu *et al.* [11] introduced grouped multi-scale dilated convolutions to obtain multi-scale contextual features. The above networks roughly stack the existing CNN frameworks, but they don't effectively utilize the spatial and spectral information of PAN and LRMS images, resulting in large redundancy in structural design.

Recently, some model-driven CNN models with clear physical meaning emerged. The basic idea is to use prior knowledge to formulate optimization problems for computer vision tasks, then unfold the optimization algorithms and replace the steps in the algorithm with deep neural networks. For example, Lefkimmiatis [23] unrolled the proximal gradient method and used a limited number of the iterations to construct the denoising network. Zhang *et al.* [41] transformed the iterative shrinkage-thresholding algorithm into a deep network form for image compressive sensing. Recently, based on the observation models for LRMS and PAN images, Xie *et al.* [36] and Xu *et al.* [37] propose the model-based deep learning network MHNet and GPPNN for Pan-sharpening, respectively. Unfortunately, the cross-stages with reduced channel numbers lead to the loss of information, which hinders their performance improvements.

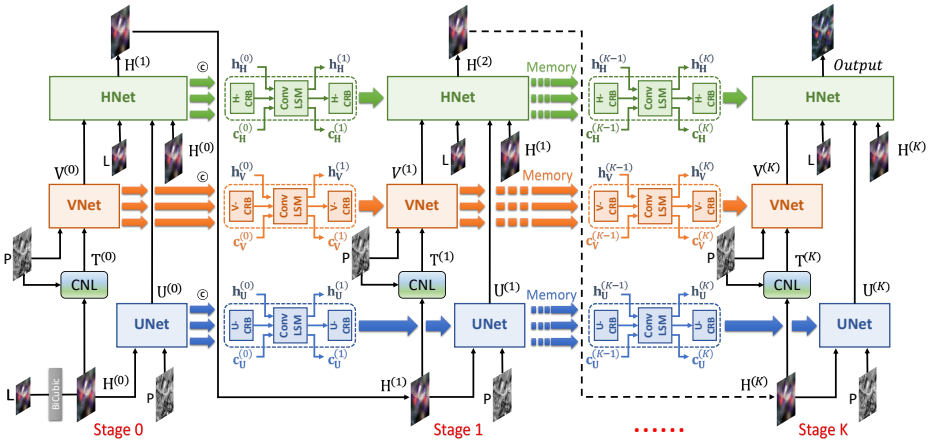

Fig. 1. The overall architecture of our proposed MMNet, consisting of information flow and memory flow. For information flow, LRMS is firstly up-sampled and then performs the stage-wise iteration updating in the overall K stages. CNL aims to explore the cross-modalities information to transfer the detailed structures of PAN to generate the immediate output N. The persistent memory mechanism (memory flow) is introduced to augment the information representation. (Best viewed in color.)

3 Proposed approach

3.1 Model formulation

In general, Pan-sharpening aims to obtain the HRMS image $\mathbf{H}$ from its degradation observation $\mathbf{L} = (\mathbf{H} \otimes \mathbf{K}) \downarrow_s + \mathbf{n_s}$, where $\mathbf{K}$ and $\downarrow_s$ denote blurring kernel and down-sampling operation, and $\mathbf{n_s}$ is usually assumed to be additive white Gaussian noise (AWGN) [36,37]. In formula, the degradation process by using the maximum a posterior (MAP) principle can be reformulated as (A detailed derivation process of the MAP model is provided in the supplementary material):

$$\max_{\mathbf{H}} \frac{1}{2} ||\mathbf{L} - \mathbf{DKH}||_2^2 + \eta \Omega_l(\mathbf{H}|\mathbf{P}) + \lambda \Omega_{NL}(\mathbf{H}|\mathbf{P}), \tag{1}$$

where $\mathbf{D}$ matrix denotes $\downarrow_s$, $\Omega_l(.)$ and $\Omega_{NL}(.)$ are the local and global implicit prior associated with $\mathbf{H}$. We solve the optimization problem using half-quadratic splitting (HQS) algorithm [12,20,17]. By introducing two auxiliary variables $\mathbf{U}$ and $\mathbf{V}$, Eq. 1 can be reformulated as a non-constrained optimization problem:

$$\begin{aligned} \min_{\mathbf{H},\mathbf{U},\mathbf{V}} \frac{1}{2} ||\mathbf{L} - \mathbf{DKH}||_2^2 + \frac{\eta_1}{2} ||\mathbf{U} - \mathbf{H}||_2^2 + \eta \Omega_l(\mathbf{U}|\mathbf{P}) \\ + \frac{\lambda_1}{2} ||\mathbf{V} - \mathbf{H}||_2^2 + \lambda \Omega_{NL}(\mathbf{V}|\mathbf{P}), \end{aligned} \tag{2}$$

where η_1 and λ_1 are penalty parameters. When η_1 and λ_1 approach infinity, Eq. 2 converges to Eq. 1. Minimizing Eq. 2 involves updating $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{H}$ alternately.

Updating U. Given the estimated HRMS image $\mathbf{H}^{(k)}$ at iteration k, the auxiliary variable $\mathbf{U}$ can be updated as:

$$\mathbf{U}^{(k)} = \arg\min_{\mathbf{U}} \frac{\eta_1}{2} \left|\left|\mathbf{U} - \mathbf{H}^{(k)}\right|\right|_2^2 + \eta_2 \Omega_l(\mathbf{U}|\mathbf{P}). \tag{3}$$

By applying the proximal gradient method [30] to Eq. (3), we can derive

$$\mathbf{U}^{(k)} = \text{prox}_{\Omega_l(.)}(\mathbf{U}^{(k-1)} - \delta_1 \nabla f_1(\mathbf{U}^{(k-1)})), \tag{4}$$

where $\text{prox}_{\Omega_l}(\cdot)$ is the proximal operator corresponding to the implicit local prior $\Omega_l(\cdot)$, δ_1 denotes the updating step size, and the gradient $\nabla f_1(\mathbf{U}^{(k-1)})$ is

$$\nabla f_1(\mathbf{U}^{(k-1)}) = \mathbf{U}^{(k-1)} - \mathbf{H}^{(k)}. \tag{5}$$

Updating V. Given $\mathbf{H}^{(k)}$, $\mathbf{V}$ can be updated as:

$$\mathbf{V}^{(k)} = \arg\min_{\mathbf{V}} \frac{\lambda_1}{2} \left|\left|\mathbf{V} - \mathbf{H}^{(k)}\right|\right|_2^2 + \lambda \Omega_{NL}(\mathbf{V}|\mathbf{P}). \tag{6}$$

Similarly, we can obtain

$$\mathbf{V}^{(k)} = \text{prox}_{\Omega_{NL}(.)}(\mathbf{V}^{(k-1)} - \delta_2 \nabla f_2(\mathbf{V}^{(k-1)})), \tag{7}$$

where $\text{prox}_{\Omega_{NL}}(\cdot)$ is the proximal operator corresponding to the non-local prior term $\Omega_{NL}(\cdot)$, δ_2 indicates the updating step size, and the gradient $\nabla f_2(\mathbf{V}^{(k-1)})$ is computed as

$$\nabla f_2(\mathbf{V}^{(k-1)}) = \mathbf{V}^{(k-1)} - \mathbf{H}^{(k)}. \tag{8}$$

Updating H. Given $\mathbf{U}^{(k)}$ and $\mathbf{V}^{(k)}$, $\mathbf{H}$ is updated as:

$$\begin{aligned}\mathbf{H}^{(k+1)} = \arg\min_{\mathbf{H}} \frac{1}{2} ||\mathbf{L} - \mathbf{DKH}||_2^2 + \frac{\eta_1}{2} \left|\left|\mathbf{U}^{(k)} - \mathbf{H}\right|\right|_2^2 \\ + \frac{\lambda_1}{2} \left|\left|\mathbf{V}^{(k)} - \mathbf{H}\right|\right|_2^2.\end{aligned} \tag{9}$$

Although we can derive the closed form update of $\mathbf{H}$ from Eq. 9, the updating equation requires computing the inverse of a large matrix, which is computationally inefficient. To solve this problem, we continue to update $\mathbf{U}$ and $\mathbf{V}$ according to the established updating rules, and we update $\mathbf{H}$ using the gradient decent method. Consequently, the updated equation for $\mathbf{H}$ is

$$\mathbf{H}^{(k+1)} = \mathbf{H}^{(k)} - \delta_3 \nabla f_3(\mathbf{H}^{(k)}), \tag{10}$$

where δ_3 is the step size, and the gradient $\nabla f_3(\mathbf{H}^{(k)})$ is

$$\begin{aligned}\nabla f_3(\mathbf{H}^{(k)}) = (\mathbf{DK})^T(\mathbf{DKH}^{(k)} - \mathbf{L}) + \eta_1(\mathbf{H}^{(k)} - \mathbf{U}^{(k)}) \\ + \lambda_1(\mathbf{H}^{(k)} - \mathbf{V}^{(k)}),\end{aligned} \tag{11}$$

where T is the matrix transpose operation.

3.2 Memory-Augmented Model-Driven Network

Based on the iterative algorithm, we construct a model-driven deep neural network for Pan-sharpening as shown in Fig. 1. This network is an implementation of the algorithm for solving Eq. (1). Since the regularization terms $\Omega_l(\cdot)$ and $\Omega_{NL}(\cdot)$ are not explicitly defined, the two proximal operators $\text{prox}_{\Omega_l}(\cdot)$ and $\text{prox}_{\Omega_{NL}}(\cdot)$ cannot be explicitly inferred in the proposed algorithm. Thus, we employ deep CNNs to learn the two proximal operators for updating $\mathbf{U}$ and $\mathbf{V}$.

However, there are still several problems of deep unfolding network need to be resolved. First, cross-stages, or short-term memory, hasn't been fully explored. Further limiting their advancements is the fact that the feature transformation with channel number reduction obscured the severe information loss between adjacent stages, which is acknowledged as the rarely realized long-term dependency. We integrate the memory mechanism into the UNet, VNet and HNet shown in

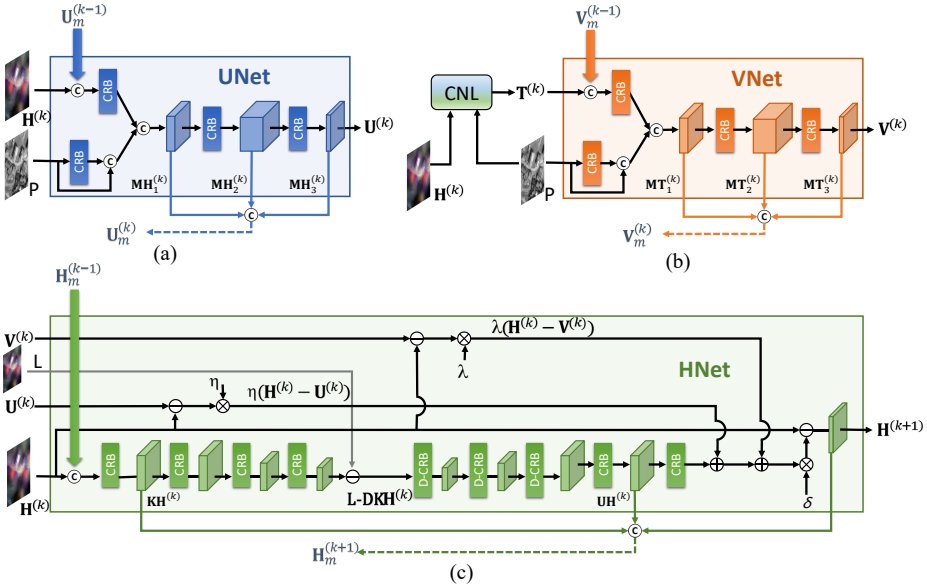

Fig. 2. The detailed structures of each sub-problems of U, V and H.

Fig. 2 in order to facilitate the signal flow between iterative stages. To be more precise, each iterative module's output intermediate images and different-layer feature maps are chosen, integrated for additional transformation, and then inserted into the following iterative stage for information interaction across stages, minimizing information loss. Next, we'll go into more detail about the upgraded versions of three sub-networks with embedded memory mechanism.

UNet. To increase the model capability, the memory of previous information at previous stages is introduced to the expressed module corresponding to $\text{prox}_{\Omega_l}(\cdot)$. As illustrated in Fig. 2 (a), the UNet is designed with the basic CRB unit which consists of the pure convolution layer and the effective residual blocks. Taking the k-th iteration for example, the computation flow of UNet is defined as

$$\mathbf{P}_1^{(k-1)} = \text{Cat}(\text{CRB}(\mathbf{P}), \mathbf{P}), \tag{12}$$

$$\mathbf{H}_1^{(k-1)} = \text{Cat}(\text{CRB}(\mathbf{H}^{(k-1)}), \mathbf{U}_m^{(k-1)}), \tag{13}$$

$$\mathbf{MH}_1^{(k-1)} = \text{Cat}(\mathbf{H}_1^{(k-1)}, \mathbf{P}_1^{(k-1)}), \tag{14}$$

$$\mathbf{MH}_2^{(k-1)} = \text{CRB}(\mathbf{MH}_1^{(k-1)}), \tag{15}$$

$$\mathbf{U}^{(k)} = \text{CRB}(\mathbf{MH}_2^{(k-1)}), \tag{16}$$

where $\text{Cat}(\cdot)$ represents the concatenation operation along the channel dimension and $\mathbf{U}_m^{(k-1)}$ is the high-throughput information from previous stage to reduce

the information loss. The updated memory $\mathbf{U}^{(k)}$ can be obtained by exploiting ConvLSTM unit to transform the different-layer's features $\mathbf{MH}_1^{(k-1)}$, $\mathbf{MH}_2^{(k-1)}$ and $\mathbf{U}^{(k)}$ as

$$\mathbf{MHU} = \mathrm{CRB}(\mathrm{Cat}(\mathbf{MH}_1^{(k-1)}, \mathbf{MH}_2^{(k-1)}, \mathbf{MH}_3^{(k-1)})), \tag{17}$$

$$\mathbf{h}_{\mathbf{U}}^{(k)}, \mathbf{c}_{\mathbf{U}}^{(k)} = \mathrm{ConvLSTM}(\mathbf{MHU}, \mathbf{h}_{\mathbf{U}}^{(k-1)}, \mathbf{c}_{\mathbf{U}}^{(k-1)}), \tag{18}$$

$$\mathbf{U}_m^{(k)} = \mathrm{CRB}(\mathbf{h}_{\mathbf{U}}^{(k)}), \tag{19}$$

where $\mathbf{h}_{\mathbf{U}}^{(k-1)}$ and $\mathbf{c}_{\mathbf{U}}^{(k-1)}$ denotes the hidden state and cell state in $ConvLSTM$ to augment the long-range cross stage information dependency. Furthermore, $\mathbf{h}_{\mathbf{U}}^{(k)}$ is directly fed into the CRB to generate the updated memory $\mathbf{U}_m^{(k)}$. The transition process of ConvLSTM is unfolded as

$$\mathbf{i}^{(k)} = \sigma(\mathbf{W}_{si} * \mathbf{MHU} + \mathbf{W}_{hi} * \mathbf{h}_{\mathbf{U}}^{(k-1)} + \mathbf{b}_i), \tag{20}$$

$$\mathbf{f}^{(k)} = \sigma(\mathbf{W}_{sf} * \mathbf{MHU} + \mathbf{W}_{hf} * \mathbf{h}_{\mathbf{U}}^{(k-1)} + \mathbf{b}_f), \tag{21}$$

$$\mathbf{c}^{(k)} = \mathbf{f}^{(k)} \odot \mathbf{c}_{\mathbf{U}}^{(k-1)} + \mathbf{i}^{(k)} \odot tanh(\mathbf{W}_{sc} * \mathbf{MHU}, \\ + \mathbf{W}_{hc} * \mathbf{h}_{\mathbf{U}}^{(k-1)} + \mathbf{b}_c), \tag{22}$$

$$\mathbf{o}^{(k)} = \sigma(\mathbf{W}_{so} * \mathbf{MHU} + \mathbf{W}_{ho} * \mathbf{h}_{\mathbf{U}}^{(k-1)} + \mathbf{b}_o), \tag{23}$$

$$\mathbf{h}_{\mathbf{U}}^{(k)} = \mathbf{o}^{(k)} \odot tanh(\mathbf{c}_{\mathbf{U}}^{(k)}), \tag{24}$$

where $*$ and $\odot$ denote the convolution operation and Hadamard product, respectively. $\mathbf{c}_{\mathbf{U}}^{(k)}$ and $\mathbf{h}_{\mathbf{U}}^{(k)}$ represent the cell state and hidden state, respectively. σ and $tanh$ denote the sigmoid and tanh function, respectively. In this way, not only the information loss of feature channel reduction is alleviated, but also the long-term cross-stage information dependency can be enhanced.

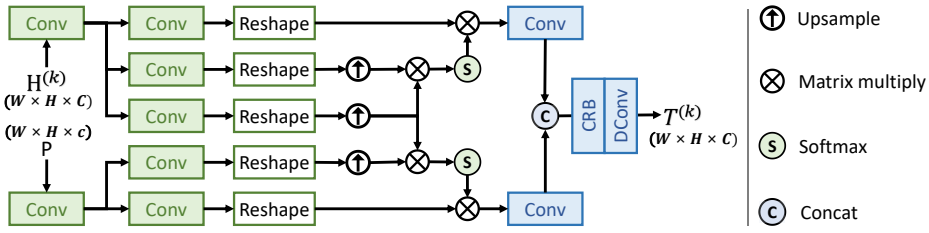

Fig. 3. The cross-modality non-local operation module. It takes the updated $\mathbf{H}^{(k)}$ and $\mathbf{P}$ as input and generates the refined image $\mathbf{T}^{(k)}$.

VNet. In Eq. 7, global implicit prior aims to measure the non-local cross-modality similarity and then aggregates the semantically-related and structure-consistent content from long-range patches in HRMS images and across the

modalities of HRMS and PAN images. To this end, we devise a novel cross-modality non-local operation module (denoted as CNL). Fig. 3 illustrates the CNL module, which receives the updated HRMS image $\mathbf{H}^{(k)}$ and PAN image $\mathbf{P}$ as input and generates the refined image $\mathbf{T}^{(k)}$.

With the output of CNL module $\mathbf{T}^{(k)}$, the previous output $\mathbf{V}^{(k-1)}$ and the accumulated memory state $\mathbf{V}_m^{(k-1)}$, we can obtain the updated $\mathbf{V}^{(k)}$ as shown in Fig. 2 (b). It can be clearly seen that the VNet has a similar architecture with that of UNet, which is consistent with their similar updating rules. Additionally, the memory transmission of VNet is also the same as that of UNet.

HNet. To transform the update process of $\mathbf{H}^{(k+1)}$ in Eq. 10 into a network. Firstly, we need to implement the two operations, i.e., *Down* $\downarrow_s$ and *Up* $\uparrow_s$, using the network. Specifically, *Down* $\downarrow_s$ is implemented by a CRB module with spatial identify transformation, and an additional s-strides followed CRB module with spatial resolution reduction:

$$\mathbf{KH}^{(k)} = \mathrm{CRB}(\mathrm{Cat}(\mathbf{H}^{(k)}, \mathbf{H}_m^{(k)})), \tag{25}$$

$$\mathbf{DKH}^{(k)} = \mathrm{CRB}^{(s)} \downarrow (\mathbf{KH}^{(k)}), \tag{26}$$

where $\mathrm{CRB}^{(s)} \downarrow$ aims to perform the s times down-sampling. The latter operation *Up* $\uparrow_s$ is implemented by a deconvolution layer containing the s-strides CRB module with spatial resolution expansion and a CRB module with spatial identify transformation:

$$\mathbf{UH}^{(k)} = \mathrm{CRB}^{(s)} \uparrow (\mathbf{L} - \mathbf{DKH}^{(k)}), \tag{27}$$

where $\mathrm{CRB}^{(s)} \uparrow$ aims to perform the s times up-sampling. Further, in context of Eq. 10, Eq. 26 and Eq. 27, the updated $\mathbf{H}^{(k+1)}$ and the updated memory $\mathbf{H}_m^{(k+1)}$ can be obtained as follows:

$$\mathbf{MH}^{(k+1)} = \mathrm{CRB}(\mathrm{Cat}(\mathbf{KH}^{(k)}, \mathbf{UH}^{(k)}, \mathbf{H}^{(k+1)})), \tag{28}$$

$$\mathbf{h}_{\mathbf{H}}^{(k+1)}, \mathbf{c}_{\mathbf{H}}^{(k+1)} = \mathrm{ConvLSTM}(\mathbf{MH}^{(k+1)}, \mathbf{h}_{\mathbf{H}}^{(k)}, \mathbf{c}_{\mathbf{H}}^{(k)}), \tag{29}$$

$$\mathbf{H}_m^{(k+1)} = \mathrm{CRB}(\mathbf{h}_{\mathbf{H}}^{(k+1)}), \tag{30}$$

where ConvLSTM performs similar functions as aforementioned. The features $\mathbf{KH}^{(k)}$, $\mathbf{UH}^{(k)}$ and $\mathbf{H}^{(k+1)}$ are obtained by different locations, thus possessing more adequate information and alleviate the information loss. Finally, with the updated $\mathbf{V}^{(k)}$, $\mathbf{U}^{(k)}$ and the accumulated memory state $\mathbf{H}_m^{(k)}$, we can obtain the updated $\mathbf{H}^{(k+1)}$ as illustrated in Fig. 2 (c).

3.3 Network Training

The distance between the estimated HRMS image from our proposed MMNet and the ground truth HRMS image is defined as the training loss for each training

pair. Mean squared error (MSE) loss is the most commonly used loss function for calculating distance. MSE loss, on the other hand, usually produces over-smoothed results. Therefore, we construct our training objective function using the mean absolute error (MAE) loss, which is defined as

$$\mathcal{L} = \sum_{i=1}^{N} \left\| \mathbf{H}_i^{(K+1)} - \mathbf{H}_{gt,i} \right\|_1, \tag{31}$$

where $\mathbf{H}_i^{(K+1)}$ denote the i-th estimated HRMS image, $\mathbf{H}_{gt,i}$ is i-th ground truth HRMS image and N is the number of training pairs.

4 Experiments

4.1 Datasets and Evaluation Metrics

We use the Wald protocol tool [35] to produce the training set since ground-truth pan-sharpened images are not available. To be more precise, the PAN image $\mathbf{P} \in R^{rM \times rN}$ and the MS image $\mathbf{H} \in R^{M \times N \times C}$ are both down-sampled with a ratio of r and then are represented by $\mathbf{L} \in R^{M/r \times N/r \times C}$ and $\mathbf{p} \in R^{M \times N}$, respectively. The LRMS image is thus considered to be $\mathbf{L}$, the PAN image to be $\mathbf{p}$, and the ground truth HRMS picture to be $\mathbf{H}$. In our experiment, the WorldViewII, WorldViewIII, and GaoFen2 satellites' remote sensing images are used for evaluation. There are hundreds of image pairs in each database, which are split into training, validation, and testing sets using the 7:2:1 ratio. Each training pair in the training set consists of one PAN image measuring 128 by 128 pixels, one LRMS patch measuring 32 by 32 pixels, and one ground truth HRMS patch measuring 128 by 128 pixels.

We use the the peak signal-to-noise ratio (PSNR), the structural similarity (SSIM), the relative dimensionless global error in synthesis (ERGAS) [1], the spectral angle mapper (SAM) [40], the correlation coefficient (SCC), and the Q index [34] to measure how well all the methods work on the test data compared to the ground truth.

In order to assess the generalizability of our approach, we generate an additional real-world, full-resolution dataset of 200 samples over the chosen GaoFen2 satellite for evaluation. Specifically, the additional dataset is created using the full-resolution setting, which produces the MS and PAN images in the way described above without downsampling. As a result, the MS image is $128 \times 128 \times 4$ and the PAN image is $32 \times 32 \times 1$. Since a ground-truth HRMS image is not available, we adopt three widely-used non-reference image quality assessment (IQA) metrics for evaluation: the spectral distortion index D_λ, the spatial distortion index D_S, and the quality without reference (QNR).

4.2 Implementation Details

In our experiments, all our designed networks are implemented in PyTorch [29] framework and trained on the PC with a single NVIDIA GeForce GTX 3060

GPU. In the training phase, these networks are optimized by the Adam optimizer [19] over 1000 epochs with a mini-batch size of 4. The learning rate is initialized with 8×10^{-4}. When reaching 200 epochs, the learning rate is decayed by multiplying 0.5. Furthermore, all the hidden and cell states of ConvLSTM are initialized as zero and the input $\mathbf{H}^{(0)}$ of our unfolding network is obtained by applying Bibubic up-sampling over LRMS image $\mathbf{L}$.

4.3 Comparison with SOTA methods

To verify the effectiveness of our proposed method on the Pan-sharpening task, we conduct several experiments on the benchmark datasets compared with several representative pan-sharpening methods: 1) five commonly-recognized state-of-the-art deep-learning based methods, including PNN [27], PANNET [38], MSDCNN [39], SRPPNN [3], and GPPNN [37]; 2) five promising traditional methods, including SFIM [25], Brovey [13], GS [22], IHS [15], and GFPCA [24].

Quantitative and qualitative results. The average quantitative performance between our method and aforementioned competitive algorithms on the three satellite datasets are tabulated in Table 1. It is clearly figured out that deep DL-based methods surpass the traditional methods and our proposed method can significantly outperform other state-of-the-art competing methods in terms of all the metrics. The qualitative comparison of the visual results over the representative sample from the WorldView-II dataset is in Figure 4. To highlight the differences in detail, we select the Red, Green and Blue bands of the generated HRMS images to better visualize the qualitative comparison. As can be seen, our method can obtain a better visual effect since it accurately enhances the spatial details and preserves the spectral information, which is consistent with quantitative results shown in Table 1. More experimental results on the three datasets are included in the supplementary material.

Table 1. Quantitative comparison with the state-of-the-art methods. The best results are highlighted by **bold**. The ↑ or ↓ indicates higher or lower values correspond to better results.

Method	WordView II				GaoFen2				WordView III			
	PSNR ↑	SSIM↑	SAM↓	ERGAS↓	PSNR ↑	SSIM↑	SAM↓	ERGAS↓	PSNR ↑	SSIM↑	SAM↓	ERGAS↓
SFIM	34.1297	0.8975	0.0439	2.3449	36.9060	0.8882	0.0318	1.7398	21.8212	0.5457	0.1208	8.9730
Brovey	35.8646	0.9216	0.0403	1.8238	37.7974	0.9026	0.0218	1.3720	22.506	0.5466	0.1159	8.2331
GS	35.6376	0.9176	0.0423	1.8774	37.2260	0.9034	0.0309	1.6736	22.5608	0.5470	0.1217	8.2433
IHS	35.2962	0.9027	0.0461	2.0278	38.1754	0.9100	0.0243	1.5336	22.5579	0.5354	0.1266	8.3616
GFPCA	34.5581	0.9038	0.0488	2.1411	37.9443	0.9204	0.0314	1.5604	22.3344	0.4826	0.1294	8.3964
PNN	40.7550	0.9624	0.0259	1.0646	43.1208	0.9704	0.0172	0.8528	29.9418	0.9121	0.0824	3.3206
PANNet	40.8176	0.9626	0.0257	1.0557	43.0659	0.9685	0.0178	0.8577	29.684	0.9072	0.0851	3.4263
MSDCNN	41.3355	0.9664	0.0242	0.9940	45.6874	0.9827	0.0135	0.6389	30.3038	0.9184	0.0782	3.1884
SRPPNN	41.4538	0.9679	0.0233	0.9899	47.1998	0.9877	0.0106	0.5586	30.4346	0.9202	0.0770	3.1553
GPPNN	41.1622	0.9684	0.0244	1.0315	44.2145	0.9815	0.0137	0.7361	30.1785	0.9175	0.0776	3.2596
Ours	**41.8577**	**0.9697**	**0.0229**	**0.9420**	**47.2668**	**0.9890**	**0.0102**	**0.5472**	**30.5451**	**0.9214**	**0.0769**	**3.1032**

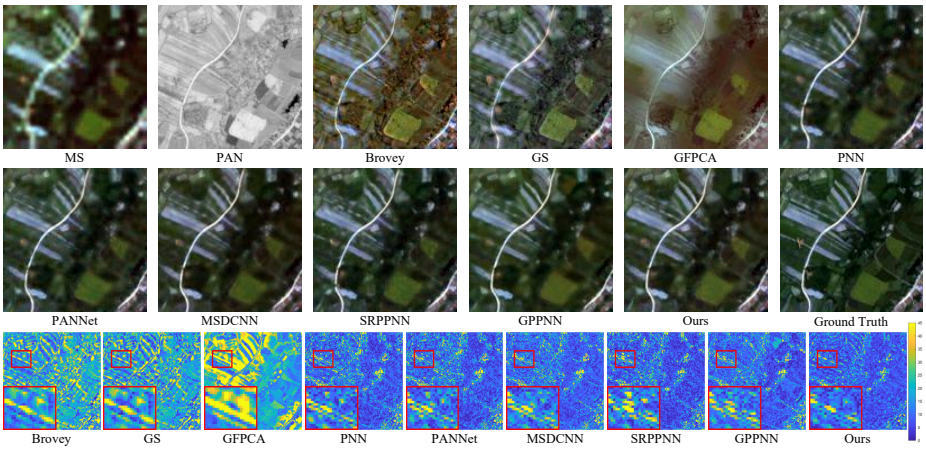

Fig. 4. Visual comparisons of the fused HRMS image for all the methods on one WorldView-II dataset. Images in the last row visualizes the MSE between the pan-sharpened results and the ground truth.

Effect on full-resolution scenes. To assess the performance of our network in the full resolution case, we apply a pre-trained model built on GaoFen2 data to the unseen GaoFen2 satellite datasets are constructed using the full-resolution setting in the preceding Section 4.1. The experimental results of the all the methods are summarized in Table 2. Additionally, we also show visual comparisons for all the methods on a full-resolution sample in Fig. 5, from which we can observe that our proposed network obtains better visual fused effect both spatially and spectrally than other competing approaches.

Complexity Analysis. Comparisons on parameter numbers, actual inference speed on GPU and model performance (as measured by PSNR) are provided in Fig. 6. The most comparable solution to ours, GPPNN [37], is organized around the model-based unfolding principle and has comparable model parameters and flops reductions but inferior performance. This is due to the cross-stages with reduced channel numbers leading to the information loss and each stage of the model without fully exploring the potential of different modalities.

4.4 Ablation Study

Ablation studies are implemented on the WorldView-II dataset to explore the contribution of different hyper-parameters and model components to the performance of our proposed model.

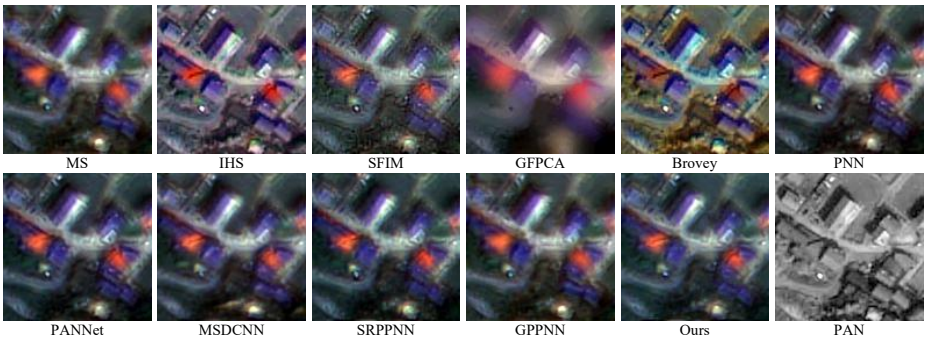

Fig. 5. Visual comparisons of the fused HRMS image on a full resolution sample.

Table 2. The average quantitative results on the GaoFen2 datasets in the full resolution case (boldface highlights the best).

Metrics	SFIM	GS	Brovey	IHS	GFPCA	PNN	PANNET	MSDCNN	SRPPNN	GPPNN	**Ours**
$D_\lambda \downarrow$	0.0822	0.0696	0.1378	0.0770	0.0914	0.0746	0.0737	0.0734	0.0767	0.0782	**0.0695**
$D_s \downarrow$	**0.1087**	0.2456	0.2605	0.2985	0.1635	0.1164	0.1224	0.1151	0.1162	0.1253	0.1139
$QNR \uparrow$	0.8214	0.7025	0.6390	0.6485	0.7615	0.8191	0.8143	0.8251	0.8173	0.8073	**0.8235**

Number of Stages. To investigate the impact of the number of unfolded stages on the performance, we experiment the proposed MMNet with varying numbers of stages K. Observing the results from Table 3, the model's performance has obtained considerable improvement as the number of stages increases until reaching to 4. When further increasing the K, the results show a decreasing trend, which may be caused by the difficulty of gradient propagation. We set $K = 4$ as the default stage number based on this experiment to balance the performance and computational complexity.

Table 3. The average results of MMNet with different number of stages.

Stage Number (K)	PSNR↑	SSIM↑	SAM↓	ERGAS↓	SCC↑	Q ↑	$D_\lambda \downarrow$	$D_S \downarrow$	QNR ↑
1	41.2772	0.9653	0.0249	1.0114	0.9664	0.7556	0.0616	0.1145	0.8319
2	41.4274	0.9673	0.0242	0.9834	0.9696	0.7650	0.0595	0.1106	0.8375
3	41.8058	0.9697	0.0224	0.9306	0.9737	0.7698	0.0622	0.1128	0.8329
4	**41.8577**	**0.9697**	**0.0229**	**0.9420**	**0.9745**	**0.7740**	**0.0629**	**0.1154**	**0.8299**
5	41.7545	0.9690	0.0226	0.9431	0.9729	0.7699	0.0600	0.1166	0.8315
6	41.4274	0.9673	0.0242	0.9834	0.9696	0.7650	0.0595	0.1106	0.8375

Effects of Different Components. To investigate the contribution of the devised modules in our network, we take the model with $K = 4$ as the baseline

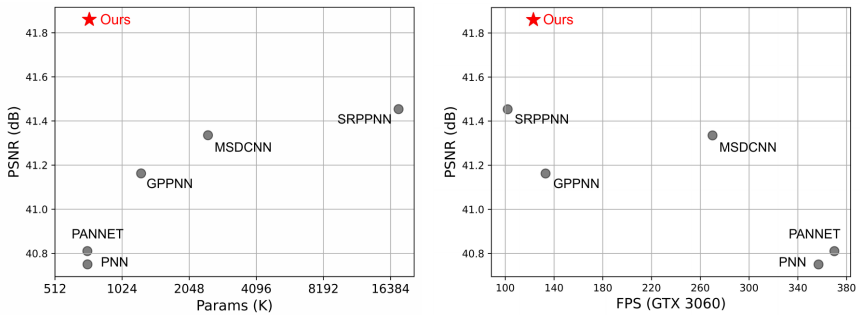

Fig. 6. Comparisons of model performance, number of parameters and FPS.

and then conduct the comparison by observing the difference before and after deleting the components. The corresponding quantitative comparison reported in Table 4, where SM represents memorizing information from a single layer and MM represents memorizing information from different-layer information at three locations. As can be observed, adding CNL and the memorized information from different locations will significantly improve the model performance.

Table 4. The results of MMNet with different components.

CNL	SM	MM	PSNR↑	SSIM↑	SAM↓	ERGAS↓	SCC↑	Q↑	$D_\lambda \downarrow$	$D_S \downarrow$	QNR↑
			41.4325	0.9668	0.0240	0.9933	0.9722	0.7579	0.0647	0.1178	0.8251
✓			41.6287	0.9683	0.0237	0.9653	0.9727	0.7673	0.0641	0.1154	0.8287
	✓		41.6476	0.9680	0.0237	0.9648	0.9729	0.7686	0.0639	0.1171	0.8277
		✓	41.7665	0.9697	0.0233	0.9437	0.9742	0.7731	0.0636	0.1168	0.8279
✓	✓		41.7199	0.9688	0.0235	0.9461	0.9734	0.7707	0.0618	0.1174	0.8289
✓		✓	**41.8577**	**0.9697**	**0.0229**	**0.9420**	**0.9745**	**0.7740**	**0.0629**	**0.1154**	**0.8299**

5 Conclusions

In this work, we propose a Memory-augmented Model-driven Network (MMNet) with interpretable structures for Pan-sharpening by unfolding the iterative algorithm into a multistage implementation. To augment the information representation across iterative stages, the persistent memory module is introduced. In this way, both the interpretability and representation ability of the deep network are improved. Extensive experiments demonstrate the superiority of the proposed method against other state-of-the-art models qualitatively and quantitatively. Additionally, compared to the other state-of-the-art models, MMNet also has competitive model parameters and running time.

References

1. Alparone, L., Wald, L., Chanussot, J., Thomas, C., Gamba, P., Bruce, L.M.: Comparison of pansharpening algorithms: Outcome of the 2006 grs-s data fusion contest. IEEE Transactions on Geoscience and Remote Sensing **45**(10), 3012–3021 (2007)
2. Ballester, C., Caselles, V., Igual, L., Verdera, J., Rougé, B.: A variational model for p+ xs image fusion. International Journal of Computer Vision **69**(1), 43–58 (2006)
3. Cai, J., Huang, B.: Super-resolution-guided progressive pansharpening based on a deep convolutional neural network. IEEE Transactions on Geoscience and Remote Sensing (2020)
4. Cao, X., Fu, X., Xu, C., Meng, D.: Deep spatial-spectral global reasoning network for hyperspectral image denoising. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–14 (2022)
5. CARPER, W., LILLESAND, T., KIEFER, R.: 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)
6. Chen, C., Li, Y., Liu, W., Huang, J.: Sirf: Simultaneous satellite image registration and fusion in a unified framework. IEEE Transactions on Image Processing **24**(11), 4213–4224 (2015)
7. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE image prior migration on pattern analysis and machine intelligence **38**(2), 295–307 (2015)
8. Dong, W., Wang, P., Yin, W., Shi, G., Wu, F., Lu, X.: Denoising prior driven deep neural network for image restoration. IEEE TPAMI **41**(10), 2305–2318 (2018)
9. Fu, X., Lin, Z., Huang, Y., Ding, X.: A variational pan-sharpening with local gradient constraints. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10265–10274 (2019)
10. Fu, X., Wang, M., Cao, X., Ding, X., Zha, Z.J.: A model-driven deep unfolding method for jpeg artifacts removal. IEEE Transactions on Neural Networks and Learning Systems pp. 1–15 (2021)
11. Fu, X., Wang, W., Huang, Y., Ding, X., Paisley, J.: Deep multiscale detail networks for multiband spectral image sharpening. IEEE Transactions on Neural Networks and Learning Systems **32**(5), 2090–2104 (2021)
12. Geman, D., Yang, C.: Nonlinear image recovery with half-quadratic regularization. IEEE transactions on Image Processing **4**(7), 932–946 (1995)
13. Gillespie, A.R., Kahle, A.B., Walker, R.E.: Color enhancement of highly correlated images. ii. channel ratio and "chromaticity" transformation techniques - sciencedirect. Remote Sensing of Environment **22**(3), 343–365 (1987)
14. Gillespie, A.R., Kahle, A.B., Walker, R.E.: Color enhancement of highly correlated images. ii. channel ratio and "chromaticity" transformation techniques. Remote Sensing of Environment **22**(3), 343–365 (1987)
15. Haydn, R., Dalke, G.W., Henkel, J., Bare, J.E.: Application of the ihs color transform to the processing of multisensor data and image enhancement. National Academy of Sciences of the United States of America **79**(13), 571–577 (1982)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)

17. He, R., Zheng, W.S., Tan, T., Sun, Z.: Half-quadratic-based iterative minimization for robust sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(2), 261–275 (2014)
18. Hu, J., Hong, D., Wang, Y., Zhu, X.: A comparative review of manifold learning techniques for hyperspectral and polarimetric sar image fusion. Remote Sensing **11**, 681 (03 2019)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017)
20. Krishnan, D., Fergus, R.: Fast image deconvolution using hyper-laplacian priors. Advances in neural information processing systems **22**, 1033–1041 (2009)
21. Kwarteng, P., Chavez, A.: Extracting spectral contrast in landsat thematic mapper image data using selective principal component analysis. Photogrammetric Engineering and remote sensing **55**(339-348), 1 (1989)
22. Laben, C.A., Brower, B.V.: Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening (2000), uS Patent 6,011,875
23. Lefkimmiatis, S.: Non-local color image denoising with convolutional neural networks. In: CVPR (July 2017)
24. Liao, W., Xin, H., Coillie, F.V., Thoonen, G., Philips, W.: Two-stage fusion of thermal hyperspectral and visible rgb image by pca and guided filter. In: Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (2017)
25. Liu, J.G.: 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)
26. Mallat, S.: A theory for multiresolution signal decomposition: The wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence **11**(7), 674–693 (1989)
27. Masi, G., Cozzolino, D., Verdoliva, L., Scarpa, G.: Pansharpening by convolutional neural networks. Remote Sensing **8**(7), 594 (2016)
28. Nunez, J., Otazu, X., Fors, O., Prades, A., Pala, V., Arbiol, R.: Multiresolution-based image fusion with additive wavelet decomposition. IEEE Transactions on Geoscience and Remote sensing **37**(3), 1204–1211 (1999)
29. Paszke, A., Gross, S., Massa, F., Lerer, A., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library (2019)
30. Rockafellar, R.T.: Monotone operators and the proximal point algorithm. Siam J Control Optim **14**(5), 877–898 (1976)
31. Schowengerdt, R.A.: Reconstruction of multispatial, multispectral image data using spatial frequency content. Photogrammetric Engineering and Remote Sensing **46**(10), 1325–1334 (1980)
32. Shah, V.P., Younan, N.H., King, R.L.: An efficient pan-sharpening method via a combined adaptive pca approach and contourlets. IEEE image prior migration on geoscience and remote sensing **46**(5), 1323–1335 (2008)
33. Tian, X., Chen, Y., Yang, C., Ma, J.: Variational pansharpening by exploiting cartoon-texture similarities. IEEE Transactions on Geoscience and Remote Sensing pp. 1–16 (2021)
34. Vivone, G., Alparone, L., Chanussot, J., Dalla Mura, M., Garzelli, A., Licciardi, G.A., Restaino, R., Wald, L.: A critical comparison among pansharpening algorithms. IEEE Transactions on Geoscience and Remote Sensing **53**(5), 2565–2586 (2014)
35. Wald, L., Ranchin, T., Mangolini, M.: Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. Photogrammetric Engineering and Remote Sensing **63**, 691–699 (11 1997)

36. Xie, Q., Zhou, M., Zhao, Q., Meng, D., Zuo, W., Xu, Z.: Multispectral and hyperspectral image fusion by ms/hs fusion net. In: CVPR. pp. 1585–1594 (2019)
37. Xu, S., Zhang, J., Zhao, Z., Sun, K., Liu, J., Zhang, C.: Deep gradient projection networks for pan-sharpening. In: CVPR. pp. 1366–1375 (June 2021)
38. Yang, J., Fu, X., Hu, Y., Huang, Y., Ding, X., Paisley, J.: Pannet: A deep network architecture for pan-sharpening. In: IEEE International Conference on Computer Vision. pp. 5449–5457 (2017)
39. Yuan, Q., Wei, Y., Meng, X., Shen, H., Zhang, L.: 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)
40. Yuhas, R.H., Goetz, A.F.H., Boardman, J.W.: Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. Proc. Summaries Annu. JPL Airborne Geosci. Workshop pp. 147–149 (1992)
41. Zhang, J., Ghanem, B.: Ista-net: Interpretable optimization-inspired deep network for image compressive sensing. In: CVPR. pp. 1828–1837 (2018)