Learned Image Reasoning Prior Penetrates Deep Unfolding Network for Panchromatic and Multi-Spectral Image Fusion

Man Zhou[1]*, Jie Huang[2]*, Naishan Zheng[2], Chongyi Li[3]†
[1]S-Lab, Nanyang Technological University, Singapore
[2]University of Science and Technology of China, China
[3]Nankai University, China

Abstract

The success of deep neural networks for pan-sharpening is commonly in a form of black box, lacking transparency and interpretability. To alleviate this issue, we propose a novel model-driven deep unfolding framework with image reasoning prior tailored for the pan-sharpening task. Different from existing unfolding solutions that deliver the proximal operator networks as the uncertain and vague priors, our framework is motivated by the content reasoning ability of masked autoencoders (MAE) with insightful designs. Specifically, the pre-trained MAE with spatial masking strategy, acting as intrinsic reasoning prior, is embedded into unfolding architecture. Meanwhile, the pre-trained MAE with spatial-spectral masking strategy is treated as the regularization term within loss function to constrain the spatial-spectral consistency. Such designs penetrate the image reasoning prior into deep unfolding networks while improving its interpretability and representation capability. The uniqueness of our framework is that the holistic learning process is explicitly integrated with the inherent physical mechanism underlying the pan-sharpening task. Extensive experiments on multiple satellite datasets demonstrate the superiority of our method over the existing state-of-the-art approaches. Code will be released at `https://manman1995.github.io/`.

1. Introduction

Pan-sharpening, a texture-rich panchromatic image-guided multi-spectral image super-resolution task, is to reason the unknown content at the pre-defined pixel positions according to the context of low-resolution (LR) multi-spectral (MS) image and high-resolution (HR) panchromatic (PAN) image. Owing to the physical constraints, satellites usually adopt both MS and PAN sensors to observe scenes, providing the MS images with high spectral but limited spatial resolution and the PAN images with high spatial but low spectral resolution. To obtain observation with both high spectral and high spatial resolutions, the pan-sharpening technique has drawn increasing attention in both image processing and remote sensing communities.

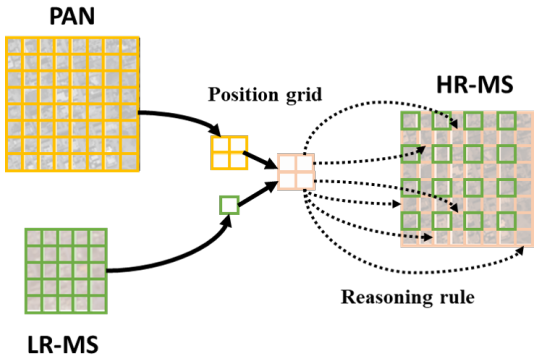

Figure 1: **Motivation.** Pan-sharpening is to reason the unknown content at the pre-defined pixel positions according to the context of low-resolution multi-spectral image and high-resolution panchromatic image.

Many research efforts have been devoted to solving the pan-sharpening problem, which can be categorized into two groups: traditional optimization methods and deep learning-based methods. Since an infinite number of HR-MS images can be downsampled to produce the same LR-MS image, reasoning the HR-MS images from the LR counterparts is highly ill-posed. To solve the ill-posedness, various natural images priors as regularization terms have been developed in traditional optimization methods, *e.g.*, low-rank prior [29] and sparse image priors [44]. However, these priors are not easy to be devised. Moreover, it is challenging to optimize these methods, hampering the practical applications. Besides, due to the hand-crafted designs, their limited representation ability results in unsatisfactory performance.

The powerful learning capability of deep neural net-

*Co-first authors contributed equally, † corresponding author. We also gratefully acknowledge the support of MindSpore, CANN, and Ascend AI Processor used for this research.

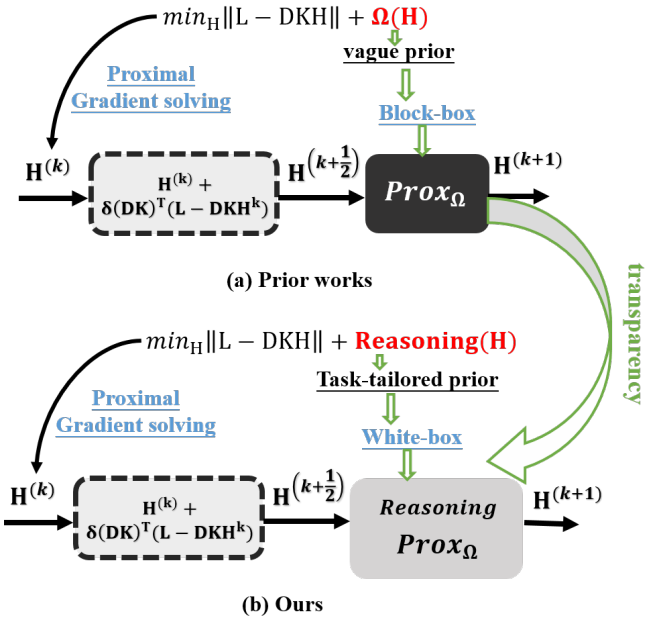

Figure 2: **Motivation.** The comparison between prior unfolding frameworks and ours.

works ignites renewed interest in this problem. As a pioneering work, PNN [18] employs three-layer convolution neural networks to account for pan-sharpening learning. Since then, more and more complicated and deeper architectures have been proposed to improve the performance of pan-sharpening [25, 28, 30]. Despite the remarkable progress, most of them focus on heuristically constructing network architectures in a black box fashion without considering the underlying rationality of pan-sharpening task, lacking transparency and interpretability.

To improve interpretability, model-driven deep unfolding methods have been proposed. Xu *et al.* [26] propose the first deep unfolding network for pan-sharpening. The basic idea behind it is to formulate pan-sharpening as an optimization problem and employ the proximal gradient descent algorithm to solve it. The optimization process can be reformulated as:

$$\hat{\mathbf{H}} = \min_{\mathbf{H}} \frac{1}{2} ||\mathbf{L} - \mathbf{DKH}||_2^2 + \lambda\Omega(\mathbf{H}|\mathbf{P}), \qquad (1)$$

where λ is a hyper-parameter to weight the first term (data fidelity term accords with the degradation) and the second term (regularization term $\Omega(\cdot)$). $\mathbf{D}$, and $\mathbf{K}$ denote the downsampling and blur operators, respectively. L, P, and H respectively represent the LR-MS image, PAN image, and HR-MS image. Technically, the proximal gradient descent algorithm can be approximatively expressed as an iterative convergence problem by solving the following iterative function:

$$\mathbf{H}^{(\mathrm{k})} = \mathrm{prox}_{\lambda,\Omega,\mathrm{P}}(\mathbf{H}^{(\mathrm{k}-1)} + \delta(\mathbf{DK})^{\mathrm{T}}(\mathbf{L} - \mathbf{DKH}^{(\mathrm{k}-1)})), \qquad (2)$$

where k and δ denote the iterative step and learning factor.

Motivation. In terms of the function $\mathrm{prox}_{\lambda,\Omega,\mathrm{P}}$ that involves the embedded prior term $\Omega(\cdot)$, existing works usually heuristically employ diverse network architectures in a black box fashion and deliver the proximal operator networks as the uncertain and vague priors, vislized in Figure 2. These methods thus lack clear physical meanings of the prior terms. In addition, most deep unfolding-based methods [26, 27] cannot extract the pan-sharpening customized prior with clear physical patterns well, which is caused by weak interpretable prior operations. Hence, we argue that the deep unfolding framework with sufficient interpretability has the potential of improving performance.

Solution. The first function of pan-sharpening is to reason the unknown information at the pre-defined pixel positions using the context, detailed in Figure 1. In this work, we propose a novel model-driven deep unfolding framework for pan-sharpening with interpretability. Inspired by the content reasoning ability of masked autoencoders (MAE) [9], our framework endows the holistic learning process of deep unfolding with the explicitly integrated with inherent physical mechanism underlying the pan-sharpening task. *The key insights of our framework are (1) we embed the pre-trained MAE with spatial masking strategy into the unfolding architecture, acting as intrinsic reasoning prior and (2) we treat the pre-trained MAE with spatial-spectral masking as the regularization term within loss function to constrain the spatial-spectral consistency.* Such new designs penetrate the image reasoning prior into deep unfolding network while improving the interpretability and representation ability. Besides, our framework also outperforms existing state-of-the-art approaches on multiple satellite datasets.

Our main contributions are summarized as follows:

- We propose to embed the pre-trained MAE into deep unfolding architecture, resorting to its image reasoning ability for pan-sharpening. Such reasoning prior makes the framework transparent and interpretable.
- We creatively treat the pre-trained MAE with spatial-spectral masking strategy as regularization term within loss function to constrain the spatial-spectral consistency. The tailored regularization term with intrinsic knowledge of spatial-spectral reasoning empowers the unfolding framework.
- In contrast to previous works, our framework as the first attempt pushes the frontiers of pan-sharpening towards the designs of the deep prior term. It shows outstanding performance on three satellite datasets, outperforming the state-of-the-art algorithms.

2. Related Work

Traditional Methods. Traditional pan-sharpening methods can be roughly divided 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 [3], the principal component analysis (PCA) methods [13, 22], Brovey transforms [6], and Gram-Schmidt (GS) orthogonalization method [14] are CS-based approaches. These CS-based approaches are fast since LR-MS images simply need spectral treatment to remove and replace spatial components, but the resultant HR-MS images 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) [17], high-pass filter fusion (HPF) [21], Laplacian pyramid (LP) [23], smoothing filter-based intensity modulation (SFIM) [16], and atrous wavelet transform (ATWT) [19] are typical MRA-based methods that reduce spectral distortion and improve resolution, but they heavily rely on multi-resolution techniques, which may cause local spatial artifacts. In recent years, VO-based methods are used because of the fine fusion effect on ill-posed problems. These various constraints can only inadequately reflect the limited structural relations of the images.

Deep Learning-based Methods. Deep learning-based methods have been widely used for pan-sharpening [41, 27, 34, 35, 11, 32, 40, 43, 39, 36, 37, 33, 42]. PNN [18] uses three convolutional units to directly map the relationship between PAN, LR-MS, and HR-MS images. Inspired by PNN, a large number of pan-sharpening studies based on deep learning emerge. For example, PANNet [28] adopts the residual learning module in Resnet [10]. MSDCNN [30] adds multi-scale modules on the basis of residual connection. SRPPNN [2] refers to the design idea of SRCNN [4].

Recently, some model-driven deep models with physical meaning emerge. The basic idea is to use prior knowledge to formulate optimization problems, then unfold the optimization algorithms and replace the steps in the algorithm with deep neural networks. For example, Xu *et al.* [26] propose the model-based deep learning network MH-Net and GPPNN for pan-sharpening, respectively. In terms of the function design $\mathrm{prox}_{\lambda,\Omega}$ that takes for the embedded prior term, existing works heuristically employ diverse network architectures in the black-box fashion, thus resulting in weak physical meanings. It motivates us to explore the task-driven customized prior with clear physical patterns.

3. Methodology

3.1. Model Formulation

In general, pan-sharpening aims to obtain the HR-MS 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 blur kernel and downsampling operation, and $\mathbf{n_s}$ is usually assumed to be additive white Gaussian noise (AWGN). The degradation process by using the maximum a posterior (MAP) principle can be reformulated as:

$$\min_{\mathbf{H}} \frac{1}{2} ||\mathbf{L} - \mathbf{DKH}||_2^2 + \lambda \Omega(\mathbf{H}|\mathbf{P}), \tag{3}$$

where λ is a hyper-parameter to weight the first term (data fidelity term accords with degradation) and the regularization term $\Omega(.)$. We solve the optimization problem using the half-quadratic splitting (HQS) algorithm. By introducing one auxiliary variables $\mathbf{U}$, Eq. (3) can be reformulated as a non-constrained optimization problem:

$$\min_{\mathbf{H},\mathbf{U}} \frac{1}{2} ||\mathbf{L} - \mathbf{DKH}||_2^2 + \frac{\eta}{2} ||\mathbf{U} - \mathbf{H}||_2^2 + \lambda \Omega(\mathbf{U}|\mathbf{P}), \tag{4}$$

where η is the penalty parameter. When η approaches infinity, Eq. (4) converges to Eq. (3). Minimizing Eq. (4) involves updating $\mathbf{U}$ and $\mathbf{H}$ alternately.

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

$$\mathbf{U}^{(k)} = \arg\min_{\mathbf{U}} \frac{\eta}{2} \left|\left|\mathbf{U} - \mathbf{H}^{(k)}\right|\right|_2^2 + \lambda \Omega(\mathbf{U}|\mathbf{P}). \tag{5}$$

We can derive the solution of Eq. (5) as

$$\mathbf{U}^{(\mathrm{k})} = \mathrm{prox}_{\Omega(.),\lambda,\eta}(\mathbf{H}^{(\mathrm{k})}, \mathrm{P}), \tag{6}$$

where $\mathrm{prox}_{\Omega}(\cdot)$ is the proximal operator corresponding to the implicit local prior $\Omega(\cdot)$.

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

$$\mathbf{H}^{(k+1)} = \arg\min_{\mathbf{H}} \frac{1}{2} ||\mathbf{L} - \mathbf{DKH}||_2^2 + \frac{\eta}{2} \left|\left|\mathbf{U}^{(k)} - \mathbf{H}\right|\right|_2^2. \tag{7}$$

By applying the proximal gradient method [20] to Eq. (7), we update $\mathbf{H}$ using the gradient decent method. Consequently, the updated equation for $\mathbf{H}$ is

$$\mathbf{H}^{(k+1)} = \mathbf{H}^{(k)} - \delta_2 \nabla f_2(\mathbf{H}^{(k)}), \tag{8}$$

where δ_2 is the step size, and the gradient $\nabla f_2(\mathbf{H}^{(k)})$ is

$$\nabla f_2(\mathbf{H}^{(k)}) = (\mathbf{DK})^T(\mathbf{DKH}^{(k)} - \mathbf{L}) + \eta(\mathbf{H}^{(k)} - \mathbf{U}^{(k)}), \tag{9}$$

where T is the matrix transpose operation.

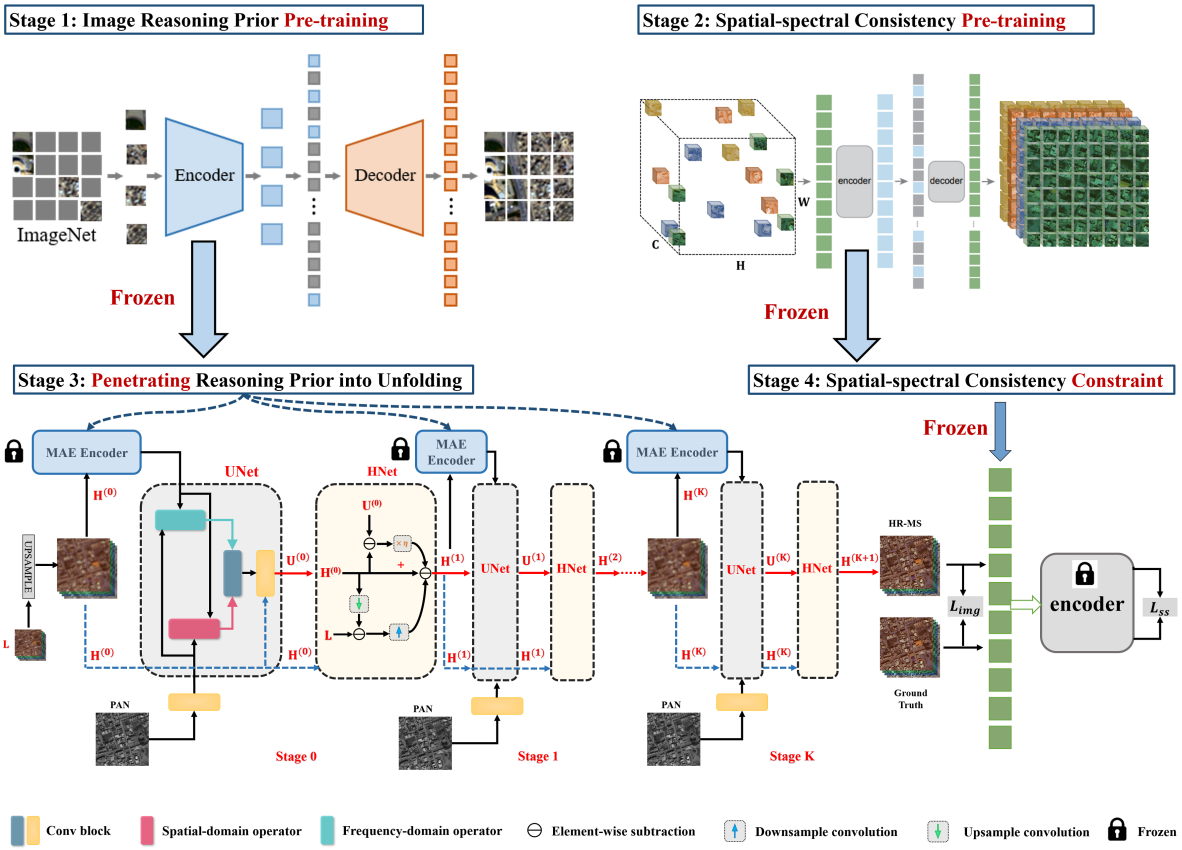

Figure 3: The overall architecture of our proposed method. In detail, LR-MS image is firstly up-sampled and then performs the stage-wise iteration, updating U and H in the overall K stages where the pre-trained MAE acting as reasoning prior penetrates Deep Unfolding process. To promote the spatial-spectral consistency, the pre-trained extension version is employed as the loss constraint. (Best viewed in color.)

3.2. Model Flowchart

As detailed in Figure 3, the proposed image reasoning prior-embedded framework consists of four training stages: **Stage 1**: Image Reasoning Prior Pre-training; **Stage 2**: Spatial-Spectral Consistency Pre-training; **Stage 3**: Penetrating Reasoning Prior into Unfolding Network; **Stage 4**: Spatial-Spectral Consistency Constraint as loss function.

3.3. Structure Flow

Based on the iterative algorithm, we construct deep unfolding network for pan-sharpening as shown in Figure 3. This network is an implementation of the algorithm for solving Eq. (3). In terms of the function design $\text{prox}_{\lambda,\Omega}$ that takes for the embedded prior term $\Omega(.)$, we stand on the shoulder of MAE proposed by He *et al*, where the MAE is trained in a self-supervised manner and empowered with the reasoning ability, acting as the image prior. Based on the above analysis, the MAE is naturally treated as the prior term $\Omega(.)$ and meets the function of pan-sharpening.

Stage 1: Based on the original MAE, we employ the pure convolution neural network as the encoder and decoder to implement its network architecture with masking patch strategies: (1) evenly divide the input image, randomly sample the small subset of regions and mask the remaining ones while keeping the whole image architecture; and (2) both the small subset of visible patches and mask tokens are processed by the encoder and decoder that reconstructs the original image in pixels. Note that the input into the encoder is the whole image, not the image patch.

Stage 3: **UNet.** Based on the pre-trained MAE encoder $\text{f}_{\text{CMAE}}(.)$, we implement the whole architecture of **UNet** that is presented in Figure 5. To be specific, the k-th iteration $\mathbf{H}^{(k-1)}$ is fed into the encoder part $\text{E}_{\text{CMAE}}(.)$ of the pre-trained MAE to generate the reasoning feature representation as

$$\mathbf{H}_{\text{rs}} = \text{E}_{\text{CMAE}}(\mathbf{H}^{(k-1)}), \tag{10}$$

Then, PAN image P is projected into the shallow feature space by the convolution units as

$$\mathbf{F}_{\text{p}} = \text{Conv}(\mathbf{P}). \tag{11}$$

Referring to the representation $\mathbf{H}_{\text{rs}}$ and the texture-rich PAN information $\mathbf{F}_{\text{p}}$, we further incorporate them to re-

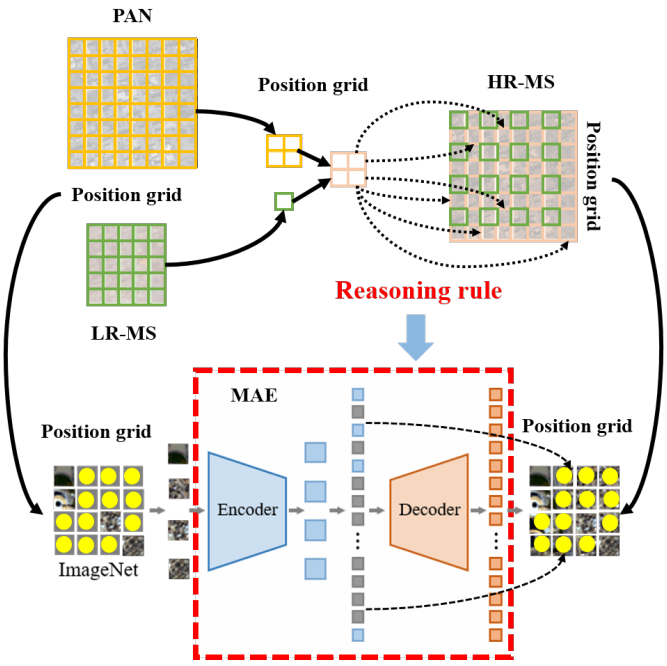

Figure 4: **Motivation.** Masked Autoencoders as the image reasoning learners. During pre-training, a large random subset of image patches is masked out, remarked as yellow ones. The encoder is applied to the small subset of visible patches. Mask tokens are introduced after the encoder, and the full set of encoded patches and mask tokens is processed by a decoder that reconstructs the original image in pixels.

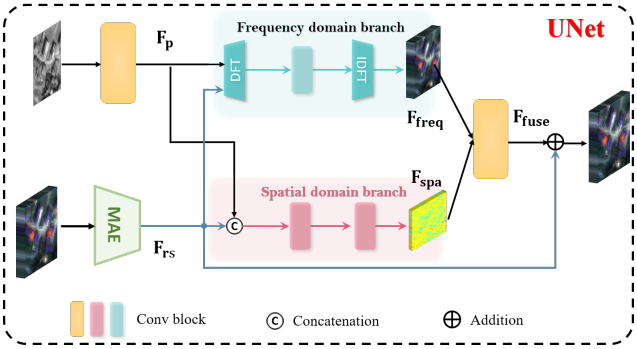

Figure 5: The detailed structure of UNet.

construct the HR-MS image by the spatial-frequency information transformation module SFT that derived from [38], which is shown in Figure 5. The information transformation process is detailed as

$$\mathrm{F}_{\mathrm{fuse}} = \mathrm{SFT}(\mathrm{H}_{\mathrm{rsl}}, \mathrm{F}_{\mathrm{p}}). \tag{12}$$

HNet. To transform the update process of $\mathbf{H}^{(k+1)}$ in Eq. (8) into a network. Firstly, we need to implement two operations, i.e., $Down\downarrow_s$ and $Up\uparrow_s$, using the network. Specifically, $Down\downarrow_s$ is implemented by a spatial identify transformation convolution operator, and an additional s-strides followed convolution module with spatial resolution reduction:

$$\mathbf{DKH}^{(k)} = \mathrm{Conv}^{(s)} \downarrow (\mathbf{KH}^{(k)}), \tag{13}$$

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

$$\mathbf{UH}^{(k)} = \mathrm{Conv}^{(s)} \uparrow (\mathbf{L} - \mathbf{DKH}^{(k)}), \tag{14}$$

where $\mathrm{Conv}\uparrow_{(s)}$ aims to perform the s times up-sampling.

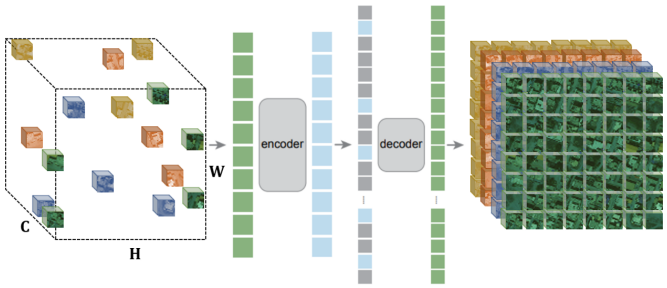

Figure 6: **Motivation.** Masked Autoencoders as spatiotemporal learners. It masks a large subset of random patches in spacetime. An encoder operates on a set of visible patches. A decoder then processes the full set of encoded patches and mask tokens to reconstruct the input. Except for patch and positional embeddings, the encoder, the decoder, and the masking strategy have no spatiotemporal inductive bias.

3.4. Optimization Flow

Stage 2: To highlight, the uniqueness of our proposed method is that the entire learning process is fully and explicitly integrated with the inherent physical mechanism underlying the pan-sharpening task. Specifically, based on the MAE with spatial-spectral masking strategy that is tailored with the spatial-frequency representation learner, we redevelop the MAE as the regularization term within the loss function to constrain the spatial-spectral consistency of the model output and its corresponding ground truth. As shown in Figure 6, standing on the shoulders of the video-version extension of masked Autoencoders [5] proposed by He *et al*, we redevelop the masked image modeling as "learned loss function" to constraint the spatial-spectral representation consistency by the following implementation details:

- randomly mask out spatial-spectral patches in the ground truth image and then learn an autoencoder to reconstruct them;
- the only spatial-spectral specific inductive bias is on embedding the patches and their positions; all other

Table 1: Quantitative comparison with the state-of-the-art methods. The best results are highlighted in **bold**.

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 (RS'16)	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 (ICCV'17)	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 (TGRS'21)	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 (TGRS'20)	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 (CVPR'21)	41.1622	0.9684	0.0244	1.0315	44.2145	0.9815	0.0137	0.7361	30.1785	0.9175	0.0776	3.2596
MutNet (CVPR'22)	41.6773	0.9705	0.0224	0.9519	47.3042	0.9892	0.0102	0.5481	30.4907	0.9223	0.0749	3.1125
MANet (ECCV'22)	41.8577	0.9697	0.0229	0.9420	47.2668	0.9890	0.0102	0.5472	30.5451	0.9214	0.0769	3.1032
Ours	**41.8735**	**0.9731**	**0.0220**	**0.9413**	**47.3931**	**0.9892**	**0.0089**	**0.5435**	**30.5560**	**0.9225**	**0.0733**	**3.0072**

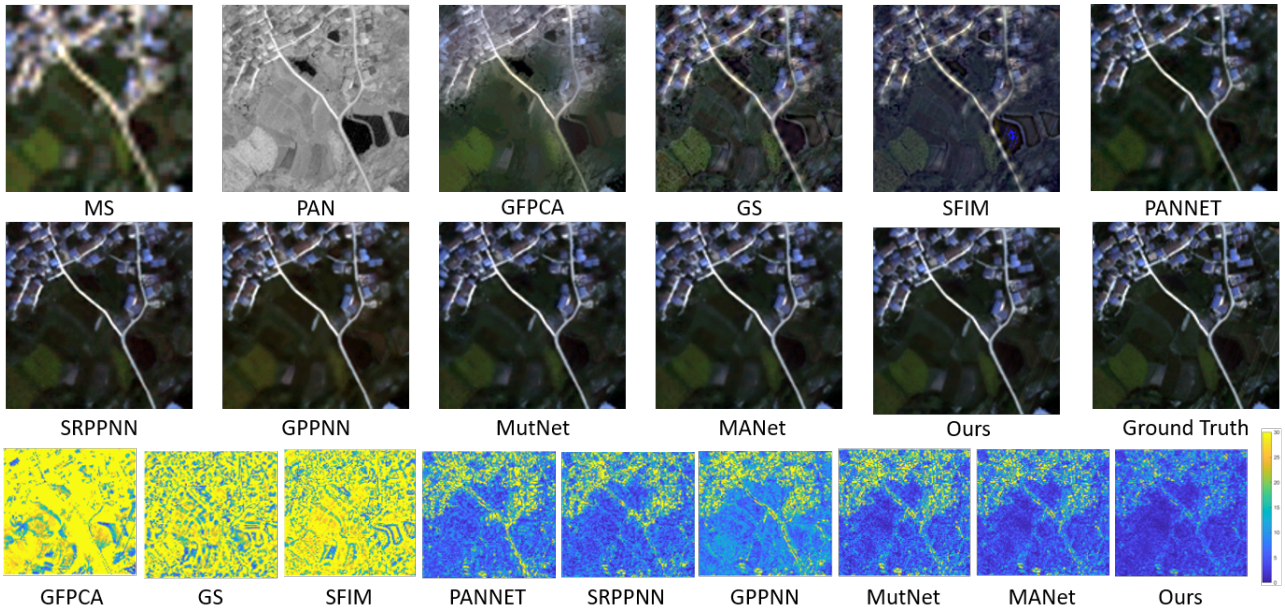

Figure 7: Visual comparison of the HR-MS images produced by different methods for processing the LR-MS image from the WorldView-II dataset. Images in the last row visualize the mean squared error image between the output and ground truth.

components are agnostic to the spatial-spectral nature of the problem. In particular, the encoder and decoder are both vanilla Vision Transformers with no factorization or hierarchy, and our random mask sampling is agnostic to the spatial-spectral structures.

To generate pleasing pan-sharpening results, we construct our training objective function using the mean absolute error loss over image-level measurement, which is defined as

$$\mathbf{L_{img}} = \sum_{i=1}^{N} \left\| \mathbf{H}_i^{(K+1)} - \mathbf{H}_{gt,i} \right\|_1, \tag{15}$$

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

Stage 4: Suppose that the pre-trained MAE model is $f_{mae}(.)$ and its encoder part is $E_{mae}(.)$, it is employed as the complementary loss function to the original image-level

Table 2: The average quantitative results on the GaoFen2 dataset in the full resolution case.

Metrics	SFIM	GS	Brovey	IHS	GFPCA	PNN	PANNET	MSDCNN	SRPPNN	GPPNN	MutNet	MANet	**Ours**
$D_\lambda \downarrow$	0.0822	0.0696	0.1378	0.0770	0.0914	0.0746	0.0737	0.0734	0.0767	0.0782	0.0694	0.0681	**0.0676**
$D_s \downarrow$	**0.1087**	0.2456	0.2605	0.2985	0.1635	0.1164	0.1224	0.1151	0.1162	0.1253	0.1118	0.1119	0.1112
$QNR \uparrow$	0.8214	0.7025	0.6390	0.6485	0.7615	0.8191	0.8143	0.8251	0.8173	0.8073	0.8259	0.8266	**0.8287**

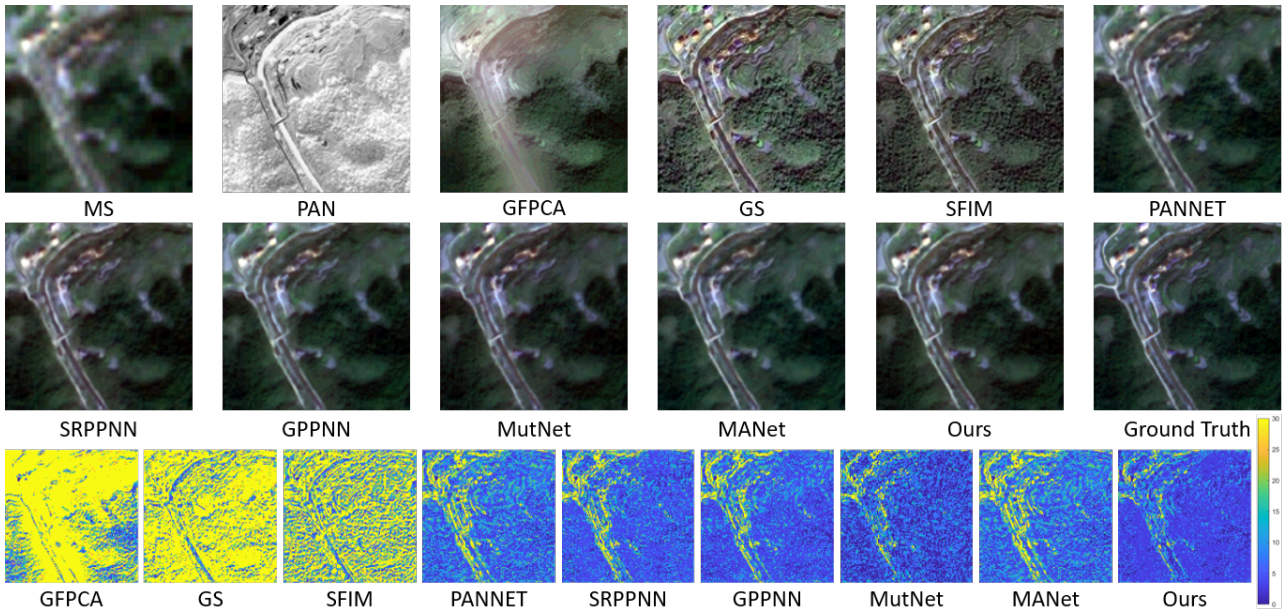

Figure 8: Qualitative visualization comparison of our method with other representative counterparts on a typical satellite image pair from the GaoFen2 dataset. Images in the last row visualizes the MSE between the output and ground truth.

loss function $\mathbf{L_{img}}$ as

$$\mathbf{L_{ss}} = \sum_{i=1}^{N} ||E_{mae}(\mathbf{H}_{gt,i}) - E_{mae}(\mathbf{H}_i^{(K+1)})||_1. \quad (16)$$

The total loss function is remarked as

$$\mathbf{L} = L_{img} + \lambda \times L_{ss}. \quad (17)$$

where λ is weighted factor and set as 1 in our work.

4. Experiments

4.1. Settings

Datasets. Due to the unavailability of ground-truth MS images, we follow the previous works to generate the training set by employing the Wald protocol tool [24]. Specifically, given the MS image $\mathrm{H} \in \mathrm{R}^{\mathrm{M}\times\mathrm{N}\times\mathrm{C}}$ and the PAN image $\mathrm{P_H} \in \mathrm{R}^{\mathrm{rM}\times\mathrm{rN}\times\mathrm{b}}$, both of them are downsampled by a ratio r, and then are denoted as $\mathrm{L} \in \mathrm{R}^{\mathrm{M/r}\times\mathrm{N/r}\times\mathrm{C}}$ and $\mathrm{P} \in \mathrm{R}^{\mathrm{M}\times\mathrm{N}\times\mathrm{b}}$, respectively. In the training set, L and p are regarded as the inputs, while H is the ground truth. In our work, three satellite images of WorldView II, GaoFen2, and WorldView III are adopted to construct image datasets. For each database, PAN images are cropped into patches with a size of 128×128 while the corresponding MS patches are with a size of 32×32.

Baselines. Several state-of-the-art Pan-sharpening methods are compared, including seven representative deep learning-based methods: PNN [18], PANNET [28], MSDCNN [30], SRPPNN [2], GPPNN [26], MANet [27], and MutNet [41] and five promising traditional methods: SFIM [16], Brovey [7], GS [14], IHS [8], and GFPCA [15].

Metrics. Several widely-used image quality assessment (IQA) metrics are employed for performance measurement, including PSNR, SSIM, SAM [31], ERGAS [1], and three non-reference metrics D_λ, D_S, and QNR for real-world full-resolution scenes.

Implementations. In our experiments, all the designed networks are implemented with PyTorch framework and trained on the PC with a single NVIDIA GeForce GTX 3090 GPU. In the training phase, these networks are optimized by the Adam optimizer [12] over 1000 epochs with a mini-batch size of 4. The learning rate is initialized with

5×10^{-4}. When reaching 200 epochs, the learning rate is decayed by multiplying 0.5.

4.2. Comparisons

Evaluation on reduced-resolution scenes. The comparison results on three satellite datasets are reported in Table 1. As can be seen, our proposed method achieves the best overall results than other pan-sharpening methods across all the satellite datasets. Specifically, the average gains of our method over the second-best MANet are 0.12dB, 0.32dB, and 0.10dB in terms of PSNR on WorldView-II, GaoFen2, and WorldView-III datasets, respectively. In addition to PSNR, consistent improvements can be observed in the other metrics, indicating lower spectral distortion and spatial texture preservation. Our method outperforms other compared methods by a large margin. The corresponding visual comparisons shown in Figure 7 and Figure 8 also support the above claim. More visual results can be found in the supplementary material.

Evaluation on full-resolution scenes. In order to demonstrate the real-world application, we further perform experiments on 200 sets of full-resolution data obtained by the additional Gaofen2. Due to the unavailability of ground-truth MS images in real-world full-resolution scenes, the commonly-used three non-reference metrics of D_λ, D_s, and QNR are adopted for evaluation. The quantitative comparisons between representative deep learning-based methods and our method are shown in Table 2. Our methods surpass other pan-sharpening methods in all metrics.

Table 3: Quantitative results of the model with different number stages.

Stage Number (K)	PSNR↑	SSIM↑	SAM↓	ERGAS↓
1	41.2459	0.9655	0.0250	1.0123
2	41.4962	0.9679	0.0240	0.9838
3	41.7152	0.9722	0.0223	0.9506
4	**41.8735**	**0.9731**	**0.0220**	**0.9413**
5	41.8461	0.9697	0.0226	0.9421
6	41.7429	0.9733	0.0221	0.9506

Table 4: Quantitative results with the ablation of key components.

$\mathrm{U_{MAE}}$	$\mathrm{L_{MAE}}$	PSNR↑	SSIM↑	SAM↓	ERGAS↓
		41.4655	0.9669	0.0253	0.9724
✓		41.6576	0.9681	0.0241	0.9679
	✓	41.8382	0.9695	0.0231	0.9423
✓	✓	**41.8735**	**0.9731**	**0.0220**	**0.9413**

4.3. Ablation Study

To explore the contribution of different hyper-parameters and the key components, we conduct ablation studies on the WorldView-II dataset.

Impact of the Stage Numbers. To investigate the impact of the number of unfolded stages, we experiment proposed method with varying numbers of stages K. Observing the results from Table 3, we found that the model's performance has obtained considerable improvement as the number of stages increases until reaching 4. When further increasing the K, the results show a decreasing trend, which may be caused by the difficulty of gradient propagation. We set $\mathrm{K} = 4$ as default stage number to balance the performance and computational complexity.

Effect of Key Components. To investigate the contribution of the devised modules in our network, we take the model with $\mathrm{K} = 4$ as the baseline and then conduct the comparison by observing the difference before and after removing the components. The corresponding quantitative comparisons are reported in Table 4, where $\mathrm{U_{MAE}}$ represents the MAE within the UNet network and $\mathrm{L_{MAE}}$ represents the MAE within loss function. As can be observed, equipping with both the MAE priors significantly improves the model performance.

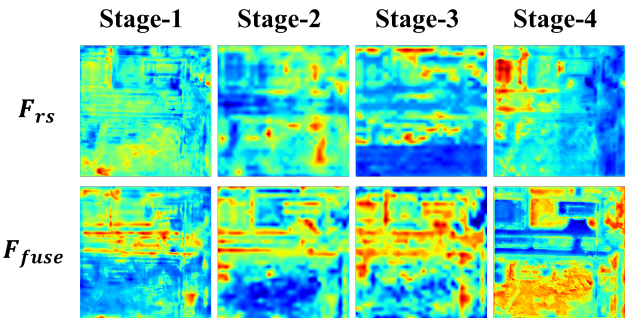

Figure 9: The feature visualization in Unet upon the iterative stage increasing of deep unfolding network.

4.4. Effect of MAE prior

To verify the effect of the designed MAE prior, we deepen into the feature maps of F_{rs}, F_{fuse}. As illustrated in section 3.3 that the MAE prior takes account for predicting the missing information of F_{rs} and then enhances the representation F_{fuse}, with the stage increasing, the MAE-prior reasoned feature F_{rs} is gradually enhanced and the resulted F_{fuse} becomes more informative, thus supporting the powerful capability of the MAE prior, detailed in Figure 9.

5. Conclusion

In this paper, we proposed the first work to focus on the designs of the deep prior term. We employ the learned

MAE in a self-supervised manner acting as an image prior and then embed the pre-trained MAE with reasoning ability to penetrate deep unfolding architecture, thus making it more transparent. We also redevelop the pre-trained MAE with a spatial-spectral masking strategy and employ it as the regularization term within loss function to constrain the spatial-spectral consistency. The contained intrinsic knowledge over MAE loss term empowers the main unfolding network learning ability. Extensive experiments on three satellite datasets demonstrate its superiority.

References

[1] L. Alparone, L. Wald, J. Chanussot, C. Thomas, P. Gamba, and L. M. Bruce. 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] Jiajun Cai and Bo Huang. Super-resolution-guided progressive pansharpening based on a deep convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[3] 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.

[4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2015.

[5] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022.

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

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

[8] R. Haydn, G. W. Dalke, J. Henkel, and J. E. Bare. 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.

[9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[11] Xuanhua He, Keyu Yan, Jie Zhang, Rui Li, Chengjun Xie, Man Zhou, and Danfeng Hong. Multiscale dual-domain guidance network for pan-sharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023.

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[13] P Kwarteng and A Chavez. Extracting spectral contrast in landsat thematic mapper image data using selective principal component analysis. *Photogrammetric Engineering and remote sensing*, 55(339-348):1, 1989.

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

[15] W. Liao, H. Xin, F. V. Coillie, G. Thoonen, and W. Philips. 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.

[16] J. G. 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.

[17] SG Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

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

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

[20] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *Siam J Control Optim*, 14(5):877–898, 1976.

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

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

[23] 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 Transactions on Geoscience and Remote Sensing*, 53(5):2565–2586, 2014.

[24] 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*, 63:691–699, 11 1997.

[25] 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.

[26] 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.

[27] Keyu Yan, Man Zhou, Li Zhang, and Chengjun Xie. Memory-augmented model-driven network for pansharpening. 2022.

[28] Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, and John Paisley. Pannet: A deep network architecture for pan-sharpening. In *IEEE International Conference on Computer Vision*, pages 5449–5457, 2017.

[29] Fei Ye, Yecai Guo, and Peixian Zhuang. Pan-sharpening via a gradient-based deep network prior. *Signal Processing: Image Communication*, 74:322–331, 2019.

[30] 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.

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

[32] Kaiwen Zheng, Jie Huang, Man Zhou, Danfeng Hong, and Feng Zhao. Deep adaptive pansharpening via uncertainty-aware image fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

[33] Man Zhou, Jie Huang, Xueyang Fu, Feng Zhao, and Danfeng Hong. Effective pan-sharpening by multiscale invertible neural network and heterogeneous task distilling. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

[34] Man Zhou, Jie Huang, Chun-Le Guo, and Chongyi Li. Fourmer: An efficient global modeling paradigm for image restoration. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 42589–42601. PMLR, 23–29 Jul 2023.

[35] Man Zhou, Jie Huang, Danfeng Hong, Feng Zhao, Chongyi Li, and Jocelyn Chanussot. Rethinking pan-sharpening in closed-loop regularization. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2023.

[36] Man Zhou, Jie Huang, Chongyi Li, Hu Yu, Keyu Yan, Naishan Zheng, and Feng Zhao. Adaptively learning low-high frequency information integration for pan-sharpening. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 3375–3384, New York, NY, USA, 2022. Association for Computing Machinery.

[37] Man Zhou, Jie Huang, Keyu Yan, Gang Yang, Aiping Liu, Chongyi Li, and Feng Zhao. Normalization-based feature selection and restitution for pan-sharpening. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 3365–3374, New York, NY, USA, 2022. Association for Computing Machinery.

[38] Man Zhou, Jie Huang, Keyu Yan, Hu Yu, Xueyang Fu, Aiping Liu, Xian Wei, and Feng Zhao. Spatial-frequency domain information integration for pan-sharpening. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*, pages 274–291. Springer, 2022.

[39] Man Zhou, Jie Huang, Keyu Yan, Hu Yu, Xueyang Fu, Aiping Liu, Xian Wei, and Feng Zhao. Spatial-frequency domain information integration for pan-sharpening. In *Computer Vision – ECCV 2022*, pages 274–291, Cham, 2022. Springer Nature Switzerland.

[40] Man Zhou, Keyu Yan, Xueyang Fu, Aiping Liu, and Chengjun Xie. Pan-guided band-aware multi-spectral feature enhancement for pan-sharpening. *IEEE Transactions on Computational Imaging*, 9:238–249, 2023.

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

[42] Man Zhou, Keyu Yan, Jinshan Pan, Wenqi Ren, Qi Xie, and Xiangyong Cao. Memory-augmented deep unfolding network for guided image super-resolution. *International Journal of Computer Vision*, 131(1):215–242, 2023.

[43] man zhou, Hu Yu, Jie Huang, Feng Zhao, Jinwei Gu, Chen Change Loy, Deyu Meng, and Chongyi Li. Deep fourier up-sampling. In *Advances in Neural Information Processing Systems*, volume 35, pages 22995–23008. Curran Associates, Inc., 2022.

[44] Xiao Xiang Zhu and Richard Bamler. A sparse image fusion algorithm with application to pan-sharpening. *IEEE transactions on geoscience and remote sensing*, 51(5):2827–2836, 2012.