Pyramid Dual Domain Injection Network for Pan-sharpening

Xuanhua He[1,2*] Keyu Yan[1,2*] Rui Li[2] Chengjun Xie[1,2] Jie Zhang[2†] Man Zhou[3†]
[1]University of Science and Technology of China, China
[2]Hefei Institute of Physical Science, Chinese Academy of Sciences, China
[3]Nanyang Technological University, Singapore
{hexuanhua,keyu}@mail.ustc.edu.cn, {lirui,cjxie,zhangjie}@iim.ac.cn,manzhountu@gmail.com

Abstract

Pan-sharpening, a panchromatic image guided low-spatial-resolution multi-spectral super-resolution task, aims to reconstruct the missing high-frequency information of high-resolution multi-spectral counterpart. Although the inborn connection with frequency domain, existing pan-sharpening research has almost investigated the potential solution upon frequency domain, thus limiting the model performance improvement. To this end, we first revisit the degradation process of pan-sharpening in Fourier space, and then devise a Pyramid Dual Domain Injection Pan-sharpening Network upon the above observation by fully exploring and exploiting the distinguished information in both the spatial and frequency domains. Specifically, the proposed network is organized with multi-scale U-shape manner and composed by two core parts: a spatial guidance pyramid sub-network for fusing local spatial information and a frequency guidance pyramid sub-network for fusing global frequency domain information, thus encouraging dual-domain complementary learning. In this way, the model can capture multi-scale dual-domain information to enable generating high-quality pan-sharpening results. Quantitative and qualitative experiments over multiple datasets demonstrate that our method performs the best against other state-of-the-art ones and comprises a strong generalization ability for real-world scenes.

1. Introduction

There is often a high demand for high-resolution multispectral (HRMS) images in the domain of agricultural monitoring and mapping services. However, due to the limitations of satellite sensors, the existing approaches must instead separately capture high-resolution panchromatic (PAN) and multispectral low-resolution (LRMS) images and fuse them using a technique known as pan-sharpening to get the desired HRMS images. Essentially, pan-sharpening is a super-resolution process that utilizes the richer details of PAN images to guide the reconstruction process of LRMS images. Pan-sharpening technology has expanded the use of remote sensing images in a wider range and gained considerable attention.

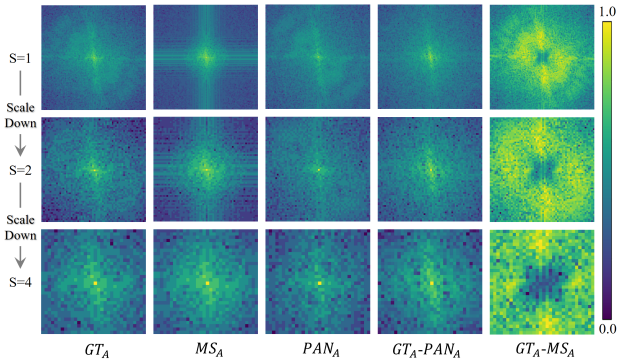

Figure 1. Frequency domain analysis of multi-scale PAN, MS, and Ground Truth images. The amplitude, denoted by A, is displayed for the original-scale image, the 2-times downsampled image, and the 4-times downsampled image, from top to bottom, respectively. The last two columns show the degradation of PAN and MS images relative to GT images.

Pan-sharpening has seen an increase in the use of deep learning-based techniques in recent years. The PNN [16], which started this trend, was motivated by the SRCNN [4] to introduce the convolutional neural network. The PNN presented a three-layer convolutional neural network that concatenated LRMS and PAN images as input and subsequently generated the output through the network. Since then, a flood of deep learning-based techniques has been developed to improve the model's representation capacity by employing deeper models or model-driven techniques [33, 29, 7, 32]. While significant progress has been made, there are still limitations need to be addressed. Prior approaches have primarily focused on spatial domain solutions, thereby neglecting the exploration of frequency domain solutions. Nevertheless, it has been established that CNNs in the spa-

*Co-first authors contributed equally, † Corresponding author.

tial domain exhibit a preference for low-frequency information and struggle to capture high-frequency information [1], resulting in insufficient texture in the generated images [14]. However, by incorporating frequency domain information, high-frequency features in low-dimensional manifolds can be efficiently captured [20]. Additionally, the LRMS images can be treated as degraded versions of HRMS images, which indicates that pan-sharpening is fundamentally an image restoration task. Image restoration are highly interrelated with frequency domain, as image degradation and content can be easily disentangled through the Fourier transform [9]. Thus, it is natural for us to investigate the feasibility of introducing frequency domain information into the pan-sharpening domain, as detailed in Figure 1 where the degradation effect of pan-sharpening is enlarged over multi-scale manner.

Motivation. We conducted multi-scale Fourier analysis on LRMS, PAN, and HRMS images since the degradation mostly affects the amplitude of the image and the diversity scale of remote sensing landforms has an effect on the pan-sharpening process. The analysis results of image spectrum is shown in the Figure 1. We can draw two inspirations from the figure: 1) The difference in amplitudes between the PAN and HRMS images is primarily in the low frequency component, whereas the difference between the amplitudes of the HRMS and LRMS images is observed in both the high and low frequency components. 2) As the scale decreases, we can observe the missing information in the PAN image becomes more concentrated in the low-frequency component, while the degradation of the LRMS image becomes more concentrated in the high frequency component, and the degradation of images at various scales is not exactly the same. We may deduce from observation 1 that the frequency information in the PAN image can be used to fill in the missing information of the LRMS image. For observation 2, it is crucial to restore the information at several scales since the degradation is varied under multi-scale conditions. To encourage the learning of context information and multi-scale reconstruction, frequency domain information should be handled at various scales. Additionally, we can encourage the interaction of global and local, low-frequency and high-frequency features through multi-scale processing in the spatial and frequency domains in accordance with CNN's low-frequency preference in spatial domain and the global receptive field described by the spectral convolution theorem [6] in frequency domain to enhance the model's performance.

In light of the aforementioned observation, we propose a brand new Pyramid Dual Domain Injection Network (PDDIN) to inject the multi-scale details of the PAN image into the LRMS image in both the spatial and frequency domains to aid in the restoration process. A Feature Extraction Pyramid, a Frequency Injection Pyramid, and a Spatial Injection Pyramid compensate the three major key components of the network. The Feature Extraction Pyramid is responsible for extracting multi-scale feature information from the PAN image. By using these features, the Frequency Injection Pyramid may inject global multi-scale features into the LRMS image, facilitating the recovery of global and high-frequency information in the spatial domain. To promote the reconstruction of local and spatial information, the spatial domain feature of PAN images is injected into the LRMS image in the Spatial Injection Pyramid. This approach effectively compensates for degradation at various scales, leading to a more thorough and precise reconstruction of the image. We conducted comprehensive experiments on multiple datasets to assess the effectiveness of our suggested network. The results show the efficacy of our method, by outperforming state-of-the-art techniques in both qualitative and quantitative evaluations.

Our contributions are summarized below:

- In this work, we revisit the degradation process of pan-sharpening in Fourier space, and then devise a Pyramid Dual Domain Injection Pan-sharpening Network, inspired by the above observation in both the spatial and frequency domains.
- To fully explore and exploit the distinguished information, we devised a frequency guidance pyramid sub-network and a spatial guidance pyramid sub-network. This design allows for the fusion of multi-scale information from both the frequency and spatial domains, lifting the learning capability of the model.
- Extensive experiments on multiple satellite datasets demonstrate our approach's quantitative and qualitative superiority to existing state-of-the-art pan-sharpening algorithms.

2. Related work

2.1. Traditional pan-sharpening method

Three sorts of traditional techniques are widely adapted in the realm of pan-sharpening: component substitution (CS) method, MRA method, and VO-based method. The component substitution approach generates high-resolution images by substituting the LRMS image components with the spatial detail components from the PAN image. IHS algorithm [10], Brovey [8], PCA [12], and GS algorithm [11] are examples of CS methods. Due to the fact that spectral information is not fully considered in CS methods, the pan-sharpened images suffer from serious spectral distortion. In light of this, the MRA method reduces spectral distortion by fusing PAN and LRMS through multi-resolution decomposition. DWT [15] ATWT [17] HPF [19] and LP [22] algorithms are typical MRA algorithms. The VO-based methods, in contrast to the above two, treat pan-sharpening as

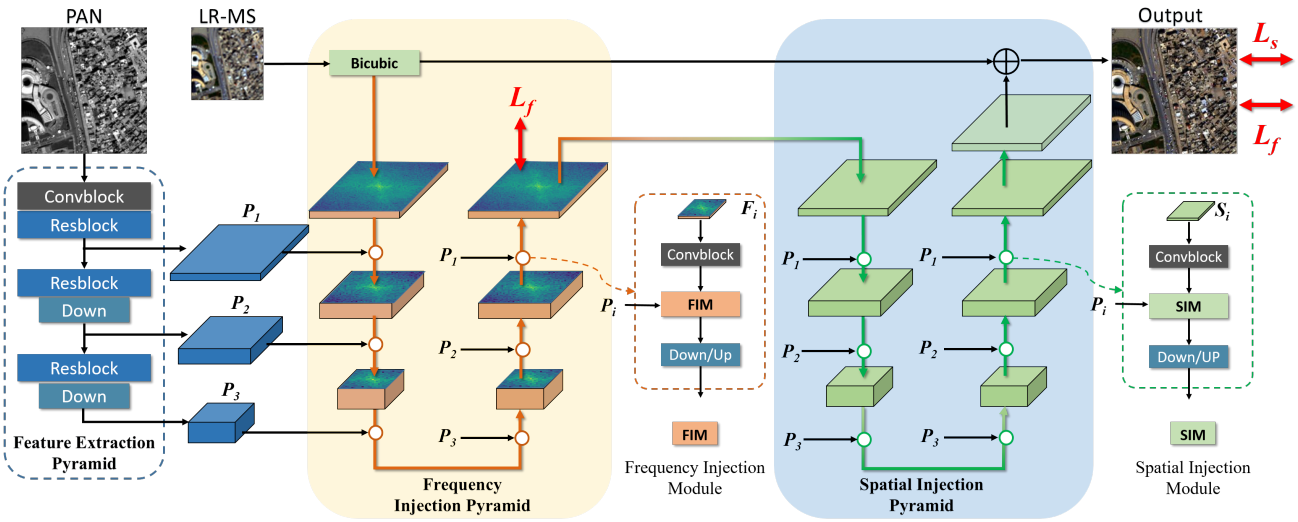

Figure 2. The framework of the Pyramid Dual Domain Injection Network. It consists of three key components: Feature Extraction Pyramid, Frequency Injection Pyramid and Spatial Injection Pyramid.

an optimization problem, design the loss function under a prior constraints, and then generate results by iterative solution and minimization of the loss function. The Bayesian technique [5] and the total variation algorithm [18] are common approaches in VO methods. The VO method can deliver decent results, but designing the optimization model and choosing the loss function is a significant challenge.

2.2. CNN-based pan-sharpening method

In recent years, the dominant approach in this area has been the CNN-based pan-sharpening method. PNN [16], which is inspired by SRCNN [4], is the pioneering work that introduced CNN to this field, giving rise to an unprecedented number of novel CNN-based techniques [27, 29, 7, 32, 33]. Despite having a clear and simple network structure, the results of PNN outperformed traditional techniques, demonstrating the promising potential of CNN techniques. Following that, Pannet [30] employs residual connection and high-frequency filtering to learn high-frequency features of images, while multi-scale convolution is taken into account in MSDCNN [31] to handle the varied remote sensing landforms. The model-driven models such as GPPNN [25], MMNet [28] and ARFNet [26] achieve interpretable models by utilizing deep unfolding techniques to construct the network structure. Additionally, SRPPNN [2] uses the progress super resolution method to enhance the performance of model and construct a profoundly deep network. Recently, MutNet [33] utilizes a mutual information mechanism to learn a more effective feature representation. The study of these techniques has also been expanded to include generative networks [13] and unsupervised pansharpening [21] technologies.

3. Methods

In this part, we first discuss the pertinent Fourier transform properties before delving into further details about our network shown in Figure 2, 3 and 4. It is organized into three parts: the Feature Extraction Pyramid, which extracts the multi-scale features from PAN images; the Frequency Injection Pyramid, which incorporates multi-scale frequency features from PAN into LRMS images for global and high-frequency information restoration; and the Spatial Injection Pyramid, which introduces the multi-scale spatial information from PAN into LRMS images to reconstruct local details. Finally, we will describe our loss function.

3.1. Fourier transform of images

Due to its capacity to decompose signals into their frequency components, the Fourier transform is widely adapted in the image processing community. This transformation makes it possible to analyze the image from a different perspective and offers valuable insights into its features. The Fourier transform of an input image $x \in R^{h \times w}$ can be defined as follows:

$$\mathcal{F}(x)(u,v) = \frac{1}{\sqrt{HW}} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} x(h,w) e^{-j2\pi(\frac{h}{H}u + \frac{w}{W}v)} \tag{1}$$

In our work, the Fourier transform is individually applied to each image channel. The amplitude and phase components can be described as follows:

$$\mathcal{A}(x)(u,v) = [R^2(x)(u,v) + I^2(x)(u,v)]^{\frac{1}{2}}, \tag{2}$$

$$\mathcal{P}(x)(u,v) = \arctan[\frac{I(x)(u,v)}{R(x)(u,v)}] \tag{3}$$

where $I(x)$ and $R(x)$ indicate imaginary and real parts of the image's frequency representation $\mathcal{F}(x)$, correspondingly.

The Fourier transform can be beneficial in the task of pan-sharpening by disentangling and analyzing image degradation in the amplitude component. We may obtain important insights into the level of degradation existing in the image by carefully evaluating the amplitude information, ultimately resulting in more effective restoration attempts. Figure 1 illustrates that whereas LRMS images degrade in both the high-frequency and low-frequency regions, while PAN images degrade mostly in the low-frequency region. Moreover, it is important to note that the frequency of image degradation can vary across different scales. Following frequency analysis, it is obvious that the frequency domain information from a PAN image can be effectively integrated into a LRMS image to assist with its restoration. This process can be further optimized by injecting the frequency domain information at multiple scales, allowing for a more comprehensive reconstruction of the image's contextual information.

3.2. Network framework

As shown in Figure 2, given an input of a PAN and LRMS pair, we start by upsampling the LRMS image to the size of the PAN image. After that, a feature extraction pyramid is employed to process the PAN image, obtaining multi-scale features. Once the LRMS image is upsampled, the frequency injection pyramid injects multi-scale features into the LRMS image. The frequency injection pyramid is used to integrate the frequency domain features of various scales to further explore the frequency information. Then, we adopt spatial injection pyramid to reconstruct local details. Finally, the reconstruction image is obtained through a skip connection between network output and the upsampled LRMS image.

3.3. The key building components

Three essential elements make up the architecture of our network, and they are shown in Figures 2, 3, and 4. We go into great depth on each of these parts in the part that follow.

Feature Extraction Pyramid. The Feature Extraction Pyramid is in charge of extracting multi-scale features from the PAN images, which can be utilized by other modules for further feature injection. To obtain the original size feature, the PAN image is first projected into the feature space using a convolution layer. After that, two sets of convolution and downsampling operations are carried out to generate features that are half as large and a quarter as large as the original. These features, namely P_1, P_2, and P_3, are then injected as supplemental information to the LRMS images to aid in the reconstruction process. The convolution operation with a stride of 2 is employed as a down-sampling operation in the pyramid. In the end, this module produces multi-scale feature maps for further information injection utilizing PAN image as input.

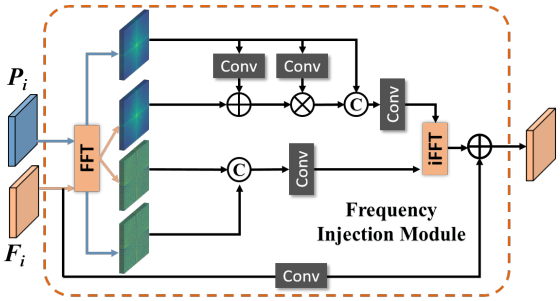

Figure 3. Architecture of Frequency Injection Module (FIM). The phase and amplitude components of P_i and F_i are obtained using the Fourier transform. Feature fusion is performed in the frequency domain, and the resulting frequency domain feature map is combined with the spatial feature to facilitate frequency domain information recovery and dual-domain complementary learning.

Spatial Injection Pyramid. The Spatial Injection Pyramid is designed to inject the local, multi-scale spatial information from PAN into the LRMS image. This module may effectively concentrate on small objects by leveraging the locality of the convolution operation in the pixel domain. In this module, the spatial features of P_1, P_2, and P_3 are injected into LRMS images at various scales to aid in reconstruction. For instance, given the output of the frequency pyramid presented by Ms_i and the feature map P_i at the original scale, further feature extraction is performed using a convolution layer to get S_i, and then the features P_i are injected into S_i using the spatial injection module (SIM). Following injection, a down-sampling or up-sampleing operation is applied to generate the feature map of a different size. Taking the down-sampling branch for example, the process can be described as follows:

$$S_i = conv_{3\times3}(Ms_i) \tag{4}$$

$$F_{spainj} = SIM(S_i, P_i) \tag{5}$$

$$Ms_{i+1} = down(F_{spainj}) \tag{6}$$

The suggested spatial injection module, shown in the figure 4, uses the AFF [3] as a basis block and is built on attention mechanism to enable effective and consistent injection of spatial details. The module works by enabling interaction between the PAN and LRMS branches, which enhances detail injection and facilitates the acquisition of LRMS features with injected spatial details.

Frequency Injection Pyramid. The purpose of the frequency injection pyramid is to provide LRMS features with high-frequency and global information from PAN image. By utilizing the multi-scale frequency injection, the restoration of LRMS image's frequency domain information is enhanced by leveraging the frequency information of PAN im-

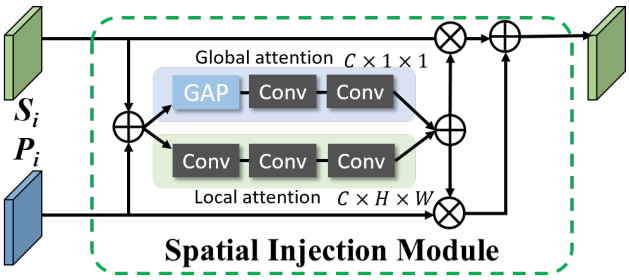

Figure 4. The proposed module for injecting spatial domain information, dubbed SIM, is based on attention mechanism and uses a basic module built on attentional fusion block.

age. Processing information in frequency domain enables the network to quickly capture high-frequency features that are challenging to acquire using pixel domain CNN, encourages dual-domain complementary learning, and generates clear texture features. Specifically, given the feature map of the LRMS image presented by Mf_i and the feature map P_i, first M_i are fed to a convolution block to extract a further feature F_i. Next, the frequency injection module (FIM) is used to inject the frequency domain features of P_i into F_i. The resulting feature map is then down-sampled or up-sampled for the next scale of frequency domain information injection. Taking the up-sampling branch as an example, this process is summarized as follows:

$$F_i = conv_{3\times3}(Mf_i) \tag{7}$$
$$F_{freinj} = FIM(F_i, P_i) \tag{8}$$
$$Mf_{i+1} = up(F_{freinj}) \tag{9}$$

where up is implemented by pixel-shuffle.

With the inputs F_i and P_i, the FIM shown in Figure 3 employs the Fourier transform to produce amplitude and phase components. Following that, the phase and amplitude information of P_i is individually injected into F_i to aid in the restoration of the LRMS image in the frequency domain.

In particular, the following procedure is used to acquire the relevant amplitude and phase components:

$$\mathcal{A}(F_i), \mathcal{P}(F_i) = \mathcal{F}(F_i), \tag{10}$$
$$\mathcal{A}(P_i), \mathcal{P}(P_i) = \mathcal{F}(P_i), \tag{11}$$

For the $\mathcal{A}(F_i)$, we use SFT [24] to inject frequency information from $\mathcal{A}(P_i)$, and use $Conv_{1\times1}$ to fuse phase components, for the difference between two modalities mainly lay in amplitude:

$$\alpha, \beta = Conv_{3\times3}(\mathcal{A}(P_i)), Conv_{3\times3}(\mathcal{A}(P_i)), \tag{12}$$
$$\mathcal{A}(F_i) = \mathcal{A}(F_i) \cdot \alpha + \beta, \tag{13}$$
$$\mathcal{A}(F) = Conv_{1\times1}(Cat([\mathcal{A}(P_i), \mathcal{A}(F_i)])), \tag{14}$$
$$\mathcal{P}(F) = Conv_{1\times1}(Cat([\mathcal{P}(P_i), \mathcal{P}(F_i)])) \tag{15}$$

The injected components are then transformed back to the pixel domain using the inverse transform, and are subsequently fused with spatial features to inject frequency features into spatial information and promote dual-domain complementary learning:

$$F_{fre} = Conv_{1\times1}(\mathcal{F}^{-1}(\mathcal{A}(F), \mathcal{P}(F))) \tag{16}$$
$$F_{spa} = Conv_{3\times3}(F_i), \tag{17}$$
$$F_{freinj} = F_{fre} + F_{spa} \tag{18}$$

3.4. Loss Function

To improve the perception of high-frequency information, we propose a loss function that combines spatial and frequency domain losses in this study. Provided that Y represents the model output, Y_f represents the output of frequency injection pyramid and H represents the ground truth, the L_1 distance between Y and H is employed to define the spatial domain loss:

$$\mathcal{L}_s = ||Y - H||_1. \tag{19}$$

The frequency domain loss, which aims to improve the perception of high-frequency information, is determined using the following formula:

$$\mathcal{L}_f = ||log(\mathcal{A}(Y_f)) - log(\mathcal{A}(H))||_1 + ||\mathcal{P}(Y) - \mathcal{P}(H)||_1, \tag{20}$$

where the amplitude and phase components of the Fourier transform are denoted by $\mathcal{A}$ and $\mathcal{P}$ respectively. The complete loss function is a combination of the dual domain losses, and involves a hyper-parameter λ which is set to 0.1 based on empirical evidence.

$$\mathcal{L} = \mathcal{L}_s + \lambda\mathcal{L}_f. \tag{21}$$

4. Experiment

4.1. Datasets and benchmark

Three datasets—WorldView-II (WV2), WorldView-III (WV3), and Gaofen2 (GF2) were used in our experiments. Due to the unavailability of real HRMS images, we utilized the Wald protocol [23] to generate the training and test data. In further detail, we extracted patches from the same location from the PAN and LRMS images, with the PAN patch being r times larger than the LRMS patch. The original LRMS patch was utilized as the ground truth, and both the LRMS and PAN patches were downsampled by a factor of r and used as input data. The selected patch sizes were 128x128 and 32x32, respectively.

To evaluate the efficacy of our approach, we compared it with five traditional and six advanced CNN-based methods. Specifically, we included SFIM, Brovey [8], GS [11], IHS [10], and GFPCA [12] as traditional methods, and PANNET [30], MSDCNN [31], SRPPNN [2], GPPNN [25], MutNet [33] and INNformer [32] as advanced methods.

Table 1. Quantitative comparison on three datasets. Best results are highlighted by red. ↑ indicates that the larger the value, the better the performance, and ↓ indicates that the smaller the value, the better the performance.

Method	WorldView-II				GaoFen2				Worldview-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	22.5060	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
PANNET	40.8176	0.9626	0.0257	1.0557	43.0659	0.9685	0.0178	0.8577	29.6840	0.9072	0.0851	3.4263
MSDCNN	41.3355	0.9664	0.0242	0.9940	45.6847	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	3.2593
MutNet	41.6773	0.9705	0.0224	0.9519	47.3042	0.9892	0.0102	0.5481	30.4907	0.9223	0.0749	3.1125
INNformer	41.6903	0.9704	0.0227	0.9514	47.3528	0.9893	0.0102	0.5479	30.5365	0.9225	0.0747	3.0997
Ours	41.8473	0.9707	0.0222	0.9358	47.4939	0.9895	0.0100	0.5360	30.8502	0.9255	0.0726	2.9945

Table 2. Evaluation of the proposed method on real-world full-resolution scenes from the GaoFen2 dataset.

Metric	SFIM	Brovey	GS	IHS	GFPCA	PANNET	MSDCNN	SRPPNN	GPPNN	MutNet	INNformer	Ours
$D_\lambda \downarrow$	0.0822	0.1378	0.0696	0.0770	0.0914	0.0737	0.0734	0.0767	0.0782	0.0694	0.0693	0.0682
$D_S \downarrow$	0.1087	0.2605	0.2456	0.2985	0.1635	0.1224	0.1151	0.1162	0.1253	0.1118	0.1138	0.1101
QNR ↑	0.8214	0.6390	0.0725	0.6485	0.7615	0.8143	0.8215	0.8173	0.8073	0.8247	0.8259	0.8307

4.2. Implementation details

The experiment was conducted on an RTX 3060 GPU, using the PyTorch framework. The Adam optimizer was employed with an initial learning rate of 5×10^{-4}. To ensure stable training, the learning rate was decayed to 5×10^{-8} using the cosine annealing method after 500 epochs. Several commonly used evaluation metrics were chosen, including PSNR, SSIM, SAM, and ERGAS, as well as non-reference indices such as D_S, D_λ, and QNR.

4.3. Comparison with state-of-the-art methods

Evaluation on Reduced-Resolution Scene. The evaluation results of our approach on three datasets are shown in Table 1. The results demonstrate that our model performs better than the state-of-the-art approaches in the majority of assessment measures, and the improvements are significant. To be specific, our method outperforms the cutting-edge method INNformer on most indicators, on the WV2 dataset, our model achieves a 0.16 dB improvement in PSNR compared to MIDN. Similarly, on the GF2 and WV3 datasets, our model achieves a 0.14 dB and 0.31 dB improvement in PSNR, respectively, demonstrating its excellent performance. Comparable improvements are also observed in other evaluation metrics.

Additionally, for qualitative evaluation, we compared the results obtained from our method with those of the other nine methods on the WV2 and GF2 datasets. To evaluate the similarity between the generated results and the ground truth, a residual map was generated by computing their mean squared error. The brighter areas of residual map indicate the larger differences with the ground truth. Our approach produces images with the finest texture and most realistic spectrum, as shown in Figures 5 and 6. The residual maps exhibit the least amount of bright spots, indicating high congruence with the ground truth. These findings are further supported by the evaluation indicators presented in the Table 1.

Evaluation on Full-Resolution Scene. To test the generalization ability of our model, we trained it on the GF2 dataset and evaluated full-resolution data from GF2 that was not downsampled, using the reserved portion of the dataset. Since there was no reference image available, we evaluated the dataset using no-reference indices. As shown in Table 2, our method achieved the best QNR index on this dataset, indicating its strong generalization ability compared to traditional and deep learning methods. This demonstrates that our approach can be applied to real-world scenes.

Table 3. The outcomes of the ablation experiments conducted on the WordView-II dataset are presented.

Config	Multi-scale	FIM	WorldView-II				GaoFen2				WorldView-III			
			PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
(I)	✗	✓	41.6361	0.9701	0.0225	0.9581	47.3788	0.9892	0.0102	0.5466	30.7849	0.9252	0.0717	3.0240
(II)	✓	✗	41.6399	0.9695	0.0227	0.9600	47.4588	0.9890	0.0101	0.5403	30.8140	0.9256	0.0718	3.0106
Ours	✓	✓	41.8435	0.9711	0.0222	0.9478	47.4939	0.9895	0.0100	0.5360	30.8645	0.9258	0.0757	2.9851

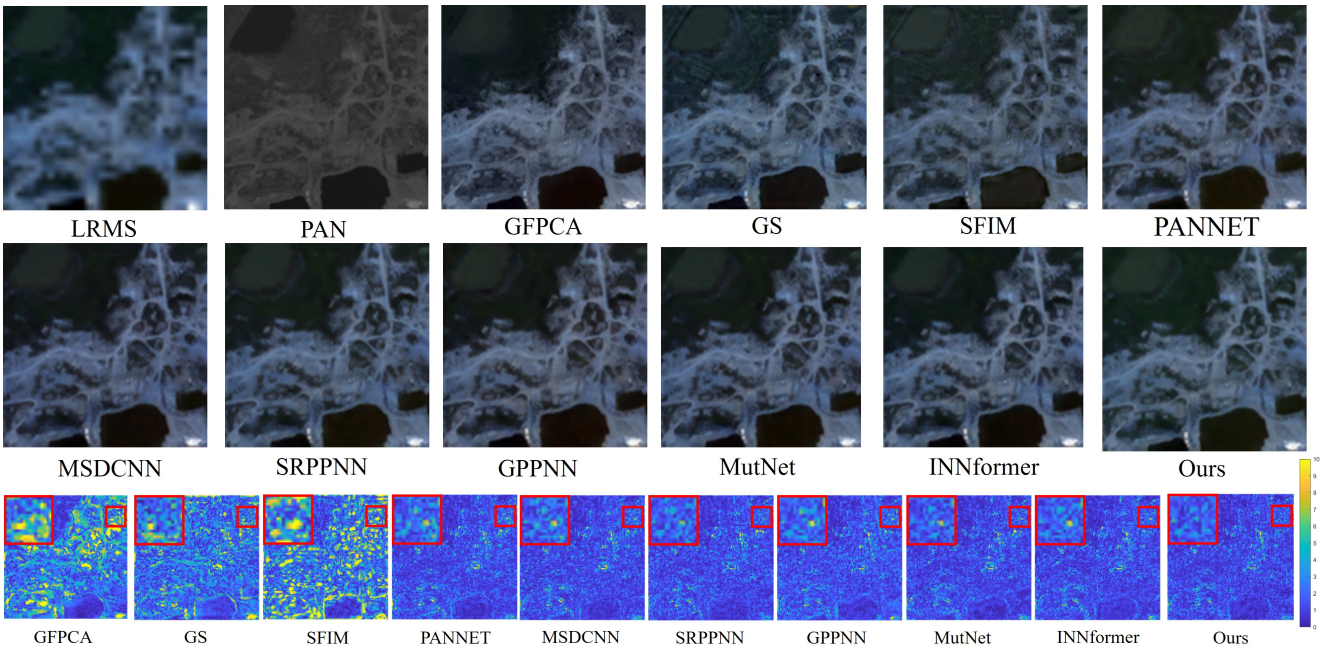

Figure 5. The result of our approach was compared against nine other methods on WorldView-II dataset.

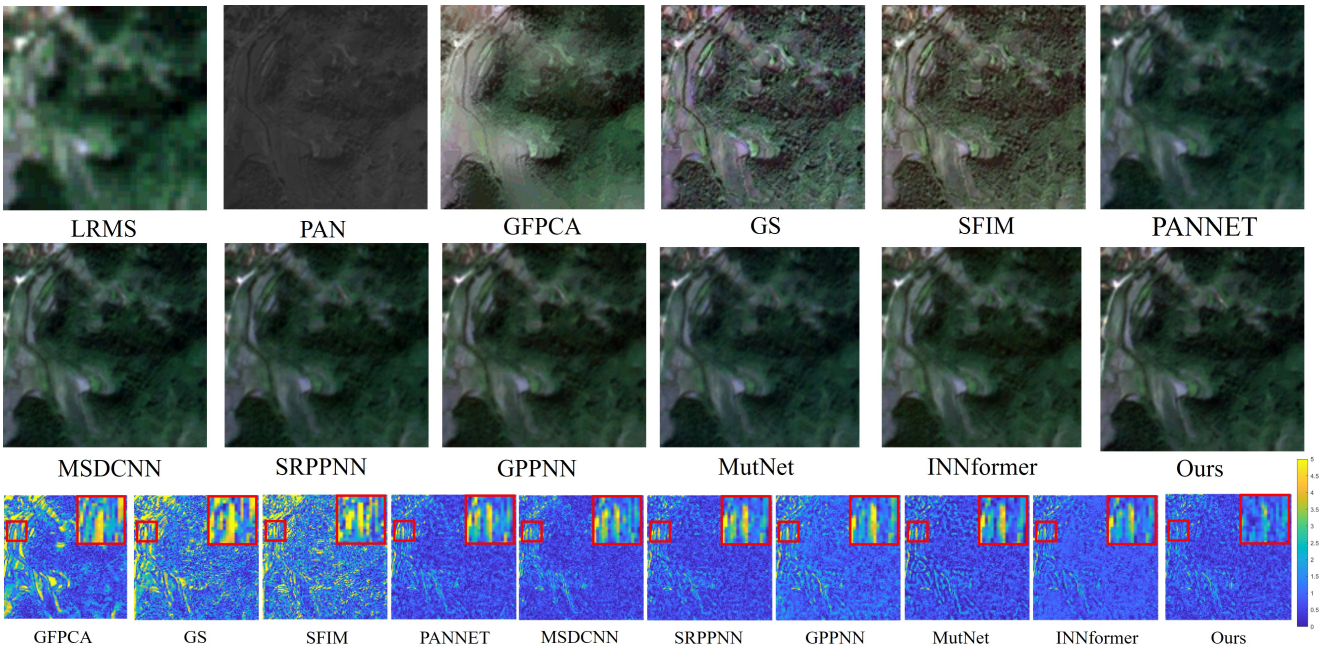

Figure 6. The result of our approach was compared against nine other methods on Gaofen2 dataset.

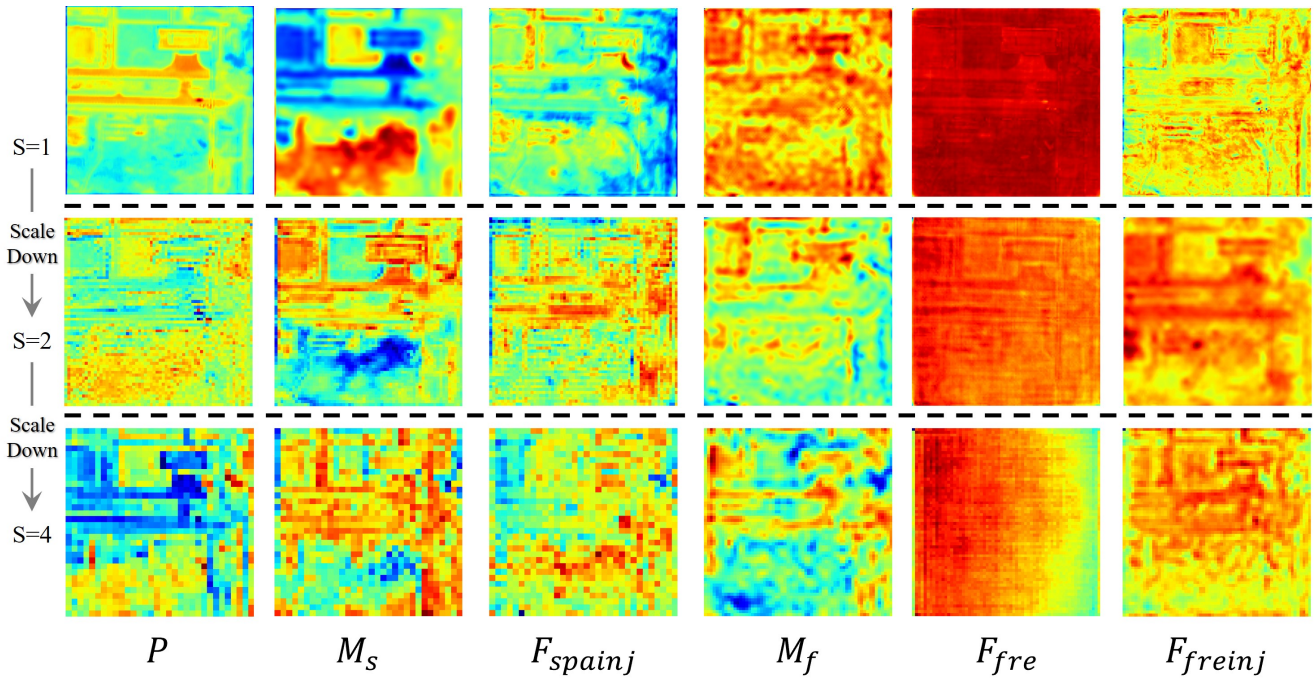

Figure 7. The feature map display result reduces in size from the top to the bottom, corresponding to the output of the pyramid. It should be noted that the last two rows have been magnified for better visualization.

4.4. Ablation experiments

We conducted two sets of ablation experiments on three datasets to examine the effects of our suggested approach on model performance to further demonstrate the effectiveness of our method. Multi-scale information modeling and dual-domain information injection are the fundamental parts of our model, and we conducted ablation experiments by removing each part independently.

Multi-scale feature extraction. The results of our model are shown in the first row of table 3 after the multi-scale feature extraction and reconstruction capabilities have been removed. We kept feature map size constant, eliminated all upsampling operations, and substituted the downsampling operation with standard convolution of stride 1 to conduct these trials. The results underscore the significance of multi-scale modeling for remote sensing images by showing that the model's performance declined after losing the capability of multi-scale modeling.

Frequency Injection Pyramid. The impact of eliminating frequency domain information from the model is shown in the second row of table 3. In order to remove frequency domain information, we used SIM to replace FIM in the frequency injection pyramid and remove the frequency loss to conduct these studies. The importance of introducing frequency information into the pan-sharpening process is revealed by the observation that the different indicators deteriorated when the global frequency domain information was removed.

4.5. Visualization of feature maps in different pyramids

To further demonstrate the multi-scale and dual-domain modeling capabilities of our model, we present P, M_s, F_{spainj}, M_f, F_{fre}, and F_{freinj} at different scales. These feature maps correspond to the output of the feature extraction pyramid, the features described in formula 6, and the features expressed in formulas 5, 9, 16, and 8, respectively. The Figure 7 illustrate that the features of different scales capture different regions of remote sensing images. The spatial pyramid features focus well on local information, while the frequency domain features effectively capture global and high-frequency features. The features after frequency domain injection combine the two types of information effectively.

5. Conclusion

In this paper, we investigate the amplitude degradation of LRMS and PAN images across various scales. Based on the results of our investigation, we suggest a pyramid dual-domain injection network that injects multi-scale frequency and spatial domain information from the PAN image into the LRMS image to aid in its restoration. Our rigorous experiments on a variety of datasets demonstrate that our suggested model surpasses SOTA methods, exhibits good generalization, and effectively handles real-world scenarios.

Acknowledgment

This work was Supported by the HFIPS Director's Fund, Grant No.2023YZGH04.

References

[1] Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017.

[2] Jiajun Cai and Bo Huang. Super-resolution-guided progressive pansharpening based on a deep convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(6):5206–5220, 2021.

[3] Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard. Attentional feature fusion. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3560–3569, 2021.

[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, 2016.

[5] Dominique Fasbender, Julien Radoux, and Patrick Bogaert. Bayesian data fusion for adaptable image pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 46(6):1847–1857, 2008.

[6] Matteo Frigo and Steven G Johnson. Fftw: An adaptive software architecture for the fft. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 3, pages 1381–1384. IEEE, 1998.

[7] Xueyang Fu, Zeyu Xiao, Gang Yang, Aiping Liu, Zhiwei Xiong, et al. Unfolding taylor's approximations for image restoration. *Advances in Neural Information Processing Systems*, 34:18997–19009, 2021.

[8] 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.

[9] Xin Guo, Xueyang Fu, Man Zhou, Zhen Huang, Jialun Peng, and Zheng-Jun Zha. Exploring fourier prior for single image rain removal. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 935–941. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

[10] 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.

[11] C.A. Laben and B.V. Brower. Process for enhancing the spatial resolution of multispectral imagery using pansharpening. *US Patent 6011875A*, 2000.

[12] 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.

[13] 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.

[14] Salma Abdel Magid, Yulun Zhang, Donglai Wei, Won-Dong Jang, Zudi Lin, Yun Fu, and Hanspeter Pfister. Dynamic high-pass filtering and multi-spectral attention for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4288–4297, 2021.

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

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

[17] 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.

[18] Frosti Palsson, Johannes R Sveinsson, and Magnus O Ulfarsson. A new pansharpening algorithm based on total variation. *IEEE Geoscience and Remote Sensing Letters*, 11(1):318–322, 2013.

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

[20] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020.

[21] Tatsumi Uezato, Danfeng Hong, Naoto Yokoya, and Wei He. Guided deep decoder: Unsupervised image pair fusion. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI*, pages 87–102. Springer, 2020.

[22] 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.

[23] 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.

[24] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by

deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 606–615, 2018.

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

[26] Keyu Yan, Man Zhou, Jie Huang, Feng Zhao, Chengjun Xie, Chongyi Li, and Danfeng Hong. Panchromatic and multispectral image fusion via alternating reverse filtering network. *Advances in Neural Information Processing Systems*, 35:21988–22002, 2022.

[27] Keyu Yan, Man Zhou, Liu Liu, Chengjun Xie, and Danfeng Hong. When pansharpening meets graph convolution network and knowledge distillation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022.

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

[29] Gang Yang, Man Zhou, Keyu Yan, Aiping Liu, Xueyang Fu, and Fan Wang. Memory-augmented deep conditional unfolding network for pan-sharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1788–1797, 2022.

[30] 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.

[31] Q. Yuan, Y. Wei, X. Meng, H. Shen, and L. 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.

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

[33] 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.