Revisiting Spatial-Frequency Information Integration from a Hierarchical Perspective for Panchromatic and Multi-Spectral Image Fusion

Jiangtong Tan[1], Jie Huang[1], Naishan Zheng[1], Man Zhou[1], Keyu Yan[1], Danfeng Hong[2], Feng Zhao[1]*
[1]University of Science and Technology of China, [2]Chinese Academy of Sciences
{jttan, hj0117, nszheng, manman, keyu}@mail.ustc.edu.cn, hongdf@aircas.ac.cn, fzhao956@ustc.edu.cn

Abstract

*Pan-sharpening is a super-resolution problem that essentially relies on spectra fusion of panchromatic (PAN) images and low-resolution multi-spectral (LRMS) images. The previous methods have validated the effectiveness of information fusion in the Fourier space of the whole image. However, they haven't fully explored the Fourier relationships at different hierarchies between PAN and LRMS images. To this end, we propose a **H**ierarchical **F**requency **I**ntegration **N**etwork (**HFIN**) to facilitate hierarchical Fourier information integration for pan-sharpening. Specifically, our network consists of two designs: information stratification and information integration. For information stratification, we hierarchically decompose PAN and LRMS information into spatial, global Fourier and local Fourier information, and fuse them independently. For information integration, the above hierarchical fused information is processed to further enhance their relationships and undergo comprehensive integration. Our method extend a new space for exploring the relationships of PAN and LRMS images, enhancing the integration of spatial-frequency information. Extensive experiments robustly validate the effectiveness of the proposed network, showcasing its superior performance compared to other state-of-the-art methods and generalization in real-world scenes and other fusion tasks as a general image fusion framework. Code is available at https://github.com/JosephTiTan/HFIN.*

1. Introduction

In remote sensing imaging, due to the limitations of satellites, it's common to utilize sensors to acquire low-resolution multi-spectral (LRMS) image with high spectral resolution and panchromatic (PAN) image with high spatial resolution but low spectral resolution. Pan-sharpening technique aims to fuse the LRMS image with the PAN image to obtain high-resolution multi-spectral (HRMS) image with both high spectral and high spatial resolutions.

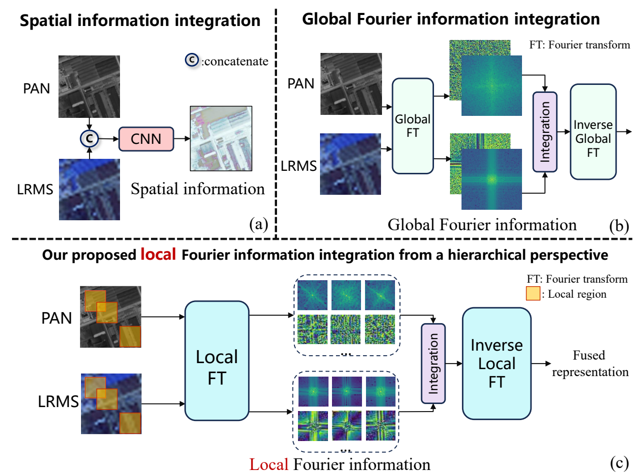


Figure 1. Illustration of different information integration process. (a): Spatial information integration; (b): Global Fourier information integration; (c): Our proposed local Fourier information integration. We explore the relationships of PAN and LRMS images from a hierarchical perspective, combining relationships in (a), (b) and (c) to achieve hierarchical information integration.

Over the past years, pan-sharpening technique has undergone rapid development and advancement. The traditional approaches employ mathematical models to fuse spatial and spectral information, typically assuming that the PAN image is a linear combination of different spectral bands of the HRMS image. However, excessive reliance on prior knowledge has constrained the applicability of these methods. With the rise of deep learning technology, convolutional neural networks have been employed in the field of pan-sharpening [1, 10, 28]. Subsequently, model structures have become increasingly complex, leading to impressive results in the field of pan-sharpening [8].

Despite the promising results achieved by these methods, most of them focus on learning in the spatial domain, neglecting the information in the frequency domain. Some studies have suggested that pan-sharpening is intri-

*Corresponding author.

cately linked to frequency domain information as a super-resolution task [17, 44, 45]. As mentioned in [45], the phase of the PAN image is more similar to the HRMS image comparing with LRMS image, while the disparity in amplitude between the PAN image and the HRMS image primarily resides in the low-frequency range, whereas the amplitude difference between the LRMS image and the HRMS image encompasses both low and high frequencies. Therefore, it is natural to utilize the Fourier Transform (FT) to obtain complementary information in frequency domain between PAN and LRMS images, further enhancing the representational capacity of the information and improving the performance of the model.

However, the previous methods only explored global Fourier fusion, neglecting the frequency relationships of PAN and LRMS images in local regions. On the other hand, spatial fusion cannot directly perform frequency fusion, as shown in Fig. 1. Due to the Fourier transform's capability to capture global frequency, we believe that capturing local Fourier information relationships of PAN and LRMS images is beneficial for modeling the local regions' global frequency integration of PAN and LRMS images, which can provide a compromise in the previous methods. Fig. 2 illustrates a simplified version of dividing the image into 16 regions, using local FT to analyze the disparities of local Fourier information between HRMS and PAN images, as well as HRMS and LRMS images in different local regions. We can clearly observe in the last column that the frequency differences in the red region are major, while in the yellow region are minor, meaning that the local Fourier information between PAN and LRMS images exhibits distinct complementarity, which further validates our argument and motivate us to combine it with spatial fusion and global Fourier fusion in previous methods to leverage hierarchical information for pan-sharpening.

Based on the above analysis, we propose Hierarchical Frequency Integration Network (HFIN) to leverage hierarchical information from both PAN and LRMS images, facilitating the integration of spatial-frequency information. Specifically, our network is composed of several fundamental modules called Spatial and Global-Local Fourier information Integration module (SGLI). The SGLI implements two functionalities: information stratification and information integration. For information stratification, we employ three blocks to extract hierarchical information from PAN and LRMS images: spatial block, global Fourier block and local Fourier block. The spatial block utilizes a conventional CNN to extract spatial information while the global Fourier branch employs discrete FT on the whole image to extract the global Fourier information. In local Fourier block, we employed a region partitioning way with 50% overlap to extract frequency information across different regions to get local Fourier information. Three blocks then

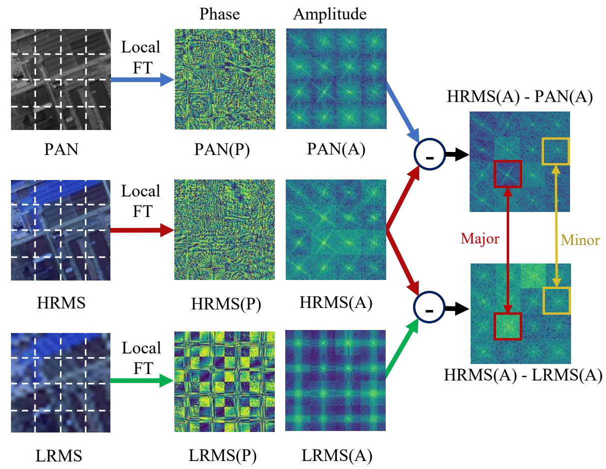


Figure 2. The observed disparities between the PAN image and the HRMS image, as well as between the LRMS image and the HRMS image, in terms of both magnitude and phase spectra in frequency domain of different regions. In the different regions, the local Fourier information between PAN and LRMS images exhibits distinct complementarity.

independently fuse PAN and LRMS information. For information integration, a crafted integration module is utilized to seamlessly integrate and complement the information from the three blocks. We extensively conduct experiments to analyze the effectiveness of the proposed network, showcasing its better performance qualitatively and quantitatively compared to other state-of-the-art methods, while also demonstrating its ability to generalize well in real-world scenes and other fusion tasks.

In summary, the contributions of our work are as follows:

- We propose a novel perspective for pan-sharpening, leveraging local Fourier information integration to explore the relationships between PAN and LRMS images. This approach complements the existing spatial and frequency fusion methods and enhances the overall performance of pan-sharpening.

- We introduce an innovative pan-sharpening framework that focuses on spatial fusion, global Fourier fusion, and local Fourier fusion at different hierarchies for information stratification, while further strengthening the performance of model by learning their interrelationships for information integration.

- Extensive experiments demonstrate that our proposed method is superior to existing state-of-the-art pan-sharpening algorithms qualitatively and quantitatively across multiple satellite datasets. Furthermore, this method can be extended to other fusion tasks and serve as a general image fusion framework.

2. Related work

2.1. Traditional pan-sharpening methods

Three commonly used traditional methods for pan-sharpening are Component Substitution (CS), Multi-resolution Analysis (MRA), and Variational Optimization (VO). The CS approaches [2, 12, 13, 23, 30], which is also called spectral methods, transform the original LRMS image into a domain suited for analysis to substitute the spatial components of PAN images. While CS methods may result in insufficient blending of the spectral and spatial information, leading to artifacts and inconsistencies in the fused image, the MRA approaches [22, 27, 29, 32] produce less spectral distortion than CS methods, which utilizes a multi-resolution decomposition of both the LRMS and PAN images to extract high-frequency spatial details. The VO approaches [4, 9, 31] assume that the PAN image is created through a linear combination of multiple HRMS image bands, which leverage various priors and constraints and performed well on pan-sharpening. However, excessive reliance on manual operations in these methods severely hampers the model's performance, resulting in degradation.

2.2. Deep learning based pan-sharpening methods

Due to the impressive representational capabilities of convolutional neural network (CNN), they have made substantial progress in the field of computer vision [11, 14, 18–21] and have found successful applications in remote sensing [36, 37, 43, 46]. [28] is the first to apply CNN in the domain of pan-sharpening, achieving superior results compared to traditional methods. In response to the challenges in pan-sharpening, researchers have explored various deep learning architectures [5, 16]. Moreover, alongside these advancements, there has been an emergence of model-driven CNN models that offer clear physical interpretations [6, 7, 35].

Recently, researchers have turned to the Discrete Fourier Transform (DFT) to tackle low-level problems [17, 40, 44], leveraging its robust capability in extracting and transforming global frequency information. [45] made pioneering attempts to address pan-sharpening in both spatial and frequency domains, introducing a global Fourier modeling approach to enhance its performance. However, the global FT completely the local Fourier information of PAN and LRMS images, which is not the optimal way for comprehensive information integration.

3. Proposed method

In this section, we will start by the properties of Fourier transform, then provide an overview of the proposed pan-sharpening network (See in Fig. 3), and finally explain the details of the key modules in our method (See in Figs. 3 and 4).

3.1. Fourier transform of images

The Discrete Fourier Transform (DFT) has long been utilized in the field of image processing because of its ability to decompose signals into frequency components. Given an image $x \in \mathbb{R}^{H\times W}$, the DFT can be expressed in the following form:

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

The DFT process is performed separately on each image channel. The amplitude and phase components can be represented by the following equation:

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

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

where $R(\cdot)$ and $I(\cdot)$ respectively represent the real and imaginary parts of the frequency representation $\mathcal{F}(\cdot)$.

It has been demonstrated that PAN and LRMS images exhibit complementary information in the frequency domain and global Fourier information integration can enhance the performance of pan-sharpening[45]. However, simply applying DFT on whole images cannot fully reflect the comprehensive relationships between PAN and LRMS images. Fig. 2 illustrates that the frequency information at different regions is also different, which is also crucial for the fusion of PAN and LRMS images. Therefore, incorporating local Fourier information with previous methods and hierarchically extracting information from PAN and LRMS images enable a more comprehensive restoration.

3.2. Network framework

Based on the aforementioned analysis, we propose a novel Hierarchical Frequency Integration Network for pan-sharpening, as illustrated in Fig. 3. Given the PAN image $P \in \mathbb{R}^{H\times W\times 1}$ and upsampled LRMS image $L \in \mathbb{R}^{H\times W\times C}$ obtained by bicubic upsampling , convolution layers are employed to map PAN and HRMS to the same feature size. The PAN image goes through independent convolution network branches to extract effective information for HRMS restoration. Then, the obtained PAN and LRMS features are continuously processed by the key module SGLI for information stratification and information integration with exchanged branches. Finally, the concatenated global Fourier branch and local Fourier branch are combined with the residual to obtain the final output.

3.3. Information stratification

As shown in Fig. 3, information stratification includes spatial block, local Fourier block and global Fourier block. The

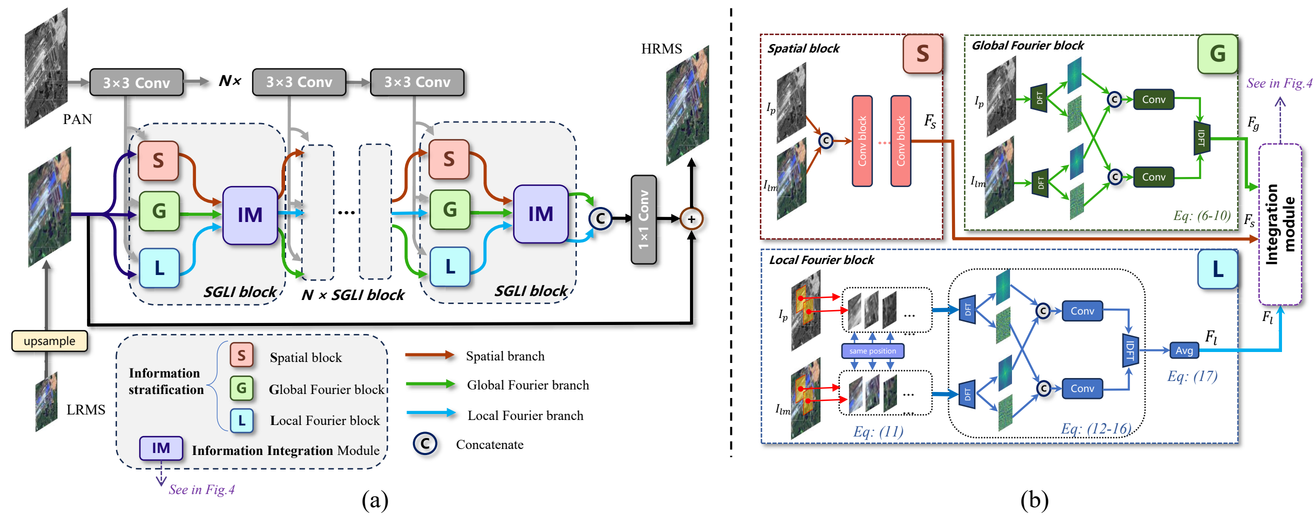


Figure 3. Overview of our method. (a): The framework of our proposed Hierarchical Frequency Integration Network. The network consists of the main module: Spatial and Global-Local Fourier information Integration block (SGLI). In detail, for information stratification, the SGLI first hierarchically decomponents different domains' information from PAN and LRMS images by three blocks: spatial block, global Fourier block and local Fourier block. Then a integration module (See in Fig. 4) is applied for information integration. After passing through several SGLI modules, the final HRMS image is obtained. (b): The three information stratification blocks in SGLI. The spatial block is entirely composed of CNNs, extracting spatial domain information; The global Fourier block extracts global Fourier information through DFT and combines the magnitude spectrum and phase spectrum separately. The local Fourier block first divides the image into regions and then performs DFT to extract local Fourier information. Then, the regions are concatenated, and the results in the overlapping areas are averaged. Finally, the information from three blocks is fed into the integration module.

three blocks respectively extract spatial information, local Fourier information and global Fourier information for hierarchical information fusion.

Spatial block. As shown in Fig. 3, the spatial block consists of convolution blocks composed of 3×3 convolution layers, which are used to extract local features F_s in the spatial domain. Convolution has a high spatial resolution, allowing it to extract information that complements frequency information, which can be observed in Fig. 5.

Global Fourier block. In the global Fourier block, as shown in Fig. 3, we first apply the DFT to obtain the magnitude spectrum and phase spectrum of both the PAN and LRMS images. Assuming the features of PAN and LRMS are I_p and I_m, respectively, the $\mathcal{F}(\cdot)$ refers to DFT and $\mathcal{F}^{-1}(\cdot)$ refers to Inverse DFT (IDFT), this process can be expressed as follows:

$$\mathcal{A}(I_p), \mathcal{P}(I_p) = \mathcal{F}(I_p), \tag{4}$$

$$\mathcal{A}(I_m), \mathcal{P}(I_m) = \mathcal{F}(I_m). \tag{5}$$

Then, we concatenate the magnitude spectra of the two images together and the phase spectra together, then respectively pass them through a three-layer 1×1 convolution neural network with ReLU activation. The resulting global frequency features are transformed back to the spatial domain using the IDFT, consistent with [45]. The entire process can be represented as follows:

$$\mathcal{A}(I_g) = conv_{1\times1}\Big(Cat_c\big(\mathcal{A}(I_p), \mathcal{A}(I_m)\big)\Big), \tag{6}$$

$$\mathcal{P}(I_g) = conv_{1\times1}\Big(Cat_c\big(\mathcal{P}(I_p), \mathcal{P}(I_m)\big)\Big), \tag{7}$$

$$F_g = \mathcal{F}^{-1}(\mathcal{A}(I_g), \mathcal{P}(I_g)), \tag{8}$$

where $Cat_c(\cdot)$ refers to concatenation operation by channel dimension. Although completely loses spatial domain information, DFT extracts global Fourier information F_g in Fig. 5, which enables modeling of contextual information with a big receptive field of the image.

Local Fourier block. For the local Fourier block shown in Fig. 3, we partition regions with 50% overlap to extract information from different positions (See comparison in Sec. 4.4). To reduce computational complexity, we only dividing it into four regions with different weights. Assuming the PAN and LRMS images are divided by the i-th region partition function $\sigma_i(\cdot)$, the process could be expressed as follows:

$$I_{p,\sigma_i}, I_{m,\sigma_i} = \sigma_i(I_p), \sigma_i(I_m), \tag{9}$$

where $\sigma_i(I)$ denotes the i-th selected rectangular region of a image. The information within each region undergoes the DFT operation, followed by the concatenation of magnitude

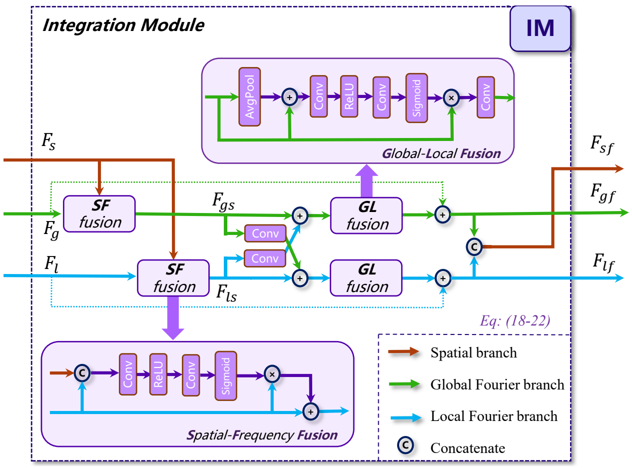


Figure 4. Architecture of integration module. In integration module, the SF fusion is employed to integrate spatial domain information into frequency information. The cross-add convolution and GL fusion enable the interaction between global Fourier information and local Fourier information.

spectra and as well as phase spectra, similar to the Global Fourier block, which could be expressed as follows:

$$\mathcal{A}(I_{p,\sigma_i}), \mathcal{P}(I_{p,\sigma_i}) = \mathcal{F}(I_{p,\sigma_i}), \tag{10}$$

$$\mathcal{A}(I_{m,\sigma_i}), \mathcal{P}(I_{m,\sigma_i}) = \mathcal{F}(I_{m,\sigma_i}). \tag{11}$$

After passing through convolution layers, we obtain local Fourier features in different regions, which are then transformed back to the spatial domain using the IDFT. We retain the non-overlapping information while averaging the overlapped parts. Finally, we obtain features with the same size as other blocks:

$$\mathcal{A}(I_{l,\sigma_i}) = conv_{1\times1}\Big(Cat_c\big(\mathcal{A}(I_{p,\sigma_i}), \mathcal{A}(I_{m,\sigma_i})\big)\Big), \tag{12}$$

$$\mathcal{P}(I_{l,\sigma_i}) = conv_{1\times1}\Big(Cat_c\big(\mathcal{P}(I_{p,\sigma_i}), \mathcal{P}(I_{m,\sigma_i})\big)\Big), \tag{13}$$

$$F_{l,\sigma_i} = \mathcal{F}^{-1}(\mathcal{A}(I_{l,\sigma_i}), \mathcal{P}(I_{l,\sigma_i})), \tag{14}$$

$$F_l = Cat_h\big(Cat_w(F_{l,\sigma_{1:M}}, \frac{1}{D}\sum_{d=0}^{d=D} F_{l,\sigma_{1:N,d}})\big), \tag{15}$$

where Cat_h and Cat_w refer to concatenation operation by height and width dimensions, respectively. M represents M non-overlapping regions, while N represents N overlapping regions. Within the overlapping regions, each pixel has D overlapping values, then we take the average of them to get the final feature, while the non-overlapping regions retain original value. The hierarchical information extracted in the three blocks complements each other, enabling comprehensive restoration, as shown in Fig. 5.

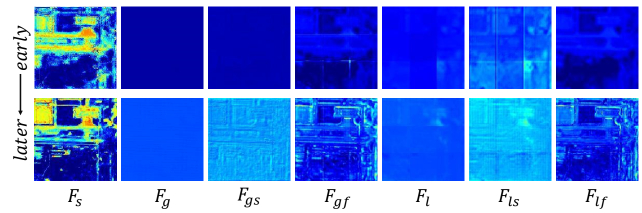


Figure 5. The Visualization of feature maps in the process of SGLI. From the top to the bottom, it displays the different stages of SGLI. The spatial feature F_s, global Fourier feature F_g and local Fourier feature F_l complements each other. In SF fusion, the spatial details in F_s complements F_g and F_l to generate F_{gs} and F_{ls}. In GL fusion, F_{gs} and F_{ls} further complements each other to generate F_{gf} and F_{lf}.

3.4. Information integration

Information integration refers to effectively combining hierarchical information from the three blocks by a integration module in Fig. 4. For the integration module, due to the substantial disparity between spatial information and frequency information, we first fuse the spatial information F_s with frequency information F_g and F_l, denoted as Spatial-Frequency (SF) Fusion. In SF Fusion, we concatenate the two branches and pass them through two convolution layers with ReLU activation. We then use the sigmoid function to obtain the importance weight for each pixel in the spatial feature. The fusion feature is obtained by multiplying the weight with the frequency branch and adding them together. The process can be expressed as follows:

$$F_{gs} = SF_{fusion}(F_s, F_g), \tag{16}$$

$$F_{ls} = SF_{fusion}(F_s, F_l). \tag{17}$$

After integrating with spatial information, both the global Fourier branch and the local Fourier branch then go through a 3×3 convolution layer and are added each other, then undergo a Global-Local (GL) Fusion process to fuse global and local Fourier information, where the resulting features F_{gf} and F_{lf} serve as the LRMS features for the next module. Furthermore, we concatenate the results from the global Fourier branch and the local Fourier branch as the LRMS feature F_{sf} for the spatial branch of the next module:

$$F_{gf} = GL_{fusion}(F_{gs} + conv_{3\times3}(F_{ls})) + F_g, \tag{18}$$

$$F_{lf} = GL_{fusion}(F_{ls} + conv_{3\times3}(F_{gs})) + F_l, \tag{19}$$

$$F_{sf} = Cat_c(F_{gf}, F_{lf}). \tag{20}$$

Lastly, we employed the L_1 loss in our study. The designed module enhances the network's ability to extracting hierarchical Fourier information and facilitating the integration of comprehensive fusion information, which promotes the fusion of PAN and LRMS images.

Table 1. Quantitative comparison on three datasets. The best and the second best values are highlighted in **bold** and underline. ↑ indicates that the larger the value, the better the performance, and ↓ indicates that the smaller the value, the better the performance.

Method	Params (M)	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	22.5060	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
PNN	0.0689	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	0.0688	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	0.2390	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	1.7114	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	0.1198	41.1622	0.9684	0.0244	1.0315	44.2145	0.9815	0.0137	0.7361	30.1785	0.9175	0.0776	3.2593
SFIINET	0.0871	41.6144	0.9689	0.0229	0.9460	47.8541	0.9877	0.0104	0.5191	30.4184	0.9182	0.0775	3.1285
PanFlowNet	0.0873	41.8584	0.9712	0.0224	0.9335	47.2533	0.9884	0.0103	0.5512	30.4873	**0.9221**	0.0751	3.1142
Ours	0.0772	**42.2319**	**0.9714**	**0.0215**	**0.8807**	**48.8783**	**0.9898**	**0.0093**	**0.4591**	**30.6147**	0.9203	**0.0742**	**3.0786**

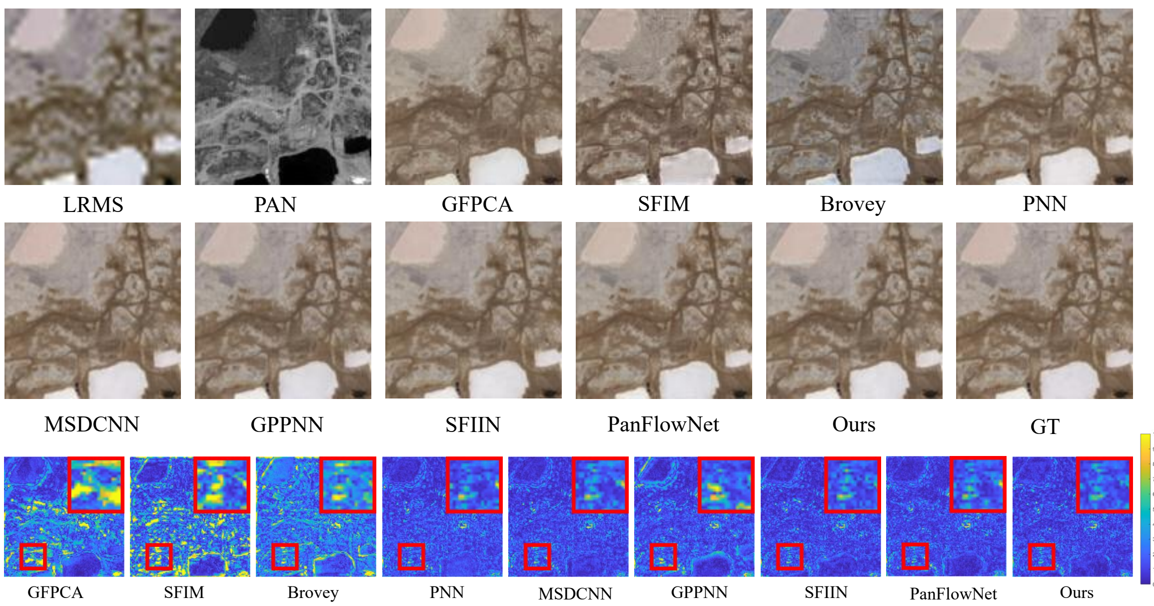


Figure 6. The result of our method compared with other methods on WorldView-II dataset.

We perform more visualizations to help readers better understand the effectiveness of hierarchical information. As depicted in Fig. 5, after undergoing SF fusion the frequency information of F_g and F_l is complemented by spatial information F_s with spatial details. In the process of GL fusion, F_{gs} has more contextual information with global receptive field and F_{ls} has more details in local regions, meaning that they can complement each other to generate F_{gf} and F_{lf}. We can also observe that as the SGLI stages increase, the restoration performance improves progressively. These visualizations further demonstrate the effectiveness of the integration of hierarchical Fourier information.

4. Experiment

4.1. Dataset and benchmarks

We conduct experiments on three datasets in our research: WorldView-II (WV2), Gaofen2 (GF2) and WorldView-III (WV3). Due to the unavailability of HRMS images, we follow the same approach as previous methods and use the Wald protocol[34] to generate training and testing data. Given the LRMS image $M_h \in R^{C \times H \times W}$ and the PAN image $P_h \in R^{C \times rH \times rW}$, both are downsampled by a ratio r to obtain $M_l \in R^{C \times \frac{H}{r} \times \frac{W}{r}}$ and $P_l \in R^{C \times H \times W}$, respectively. During training, M_l and P_l are used as inputs, while M_h serves as the ground truth. For each dataset, The size of the PAN image is cropped to 128×128, while the LRMS

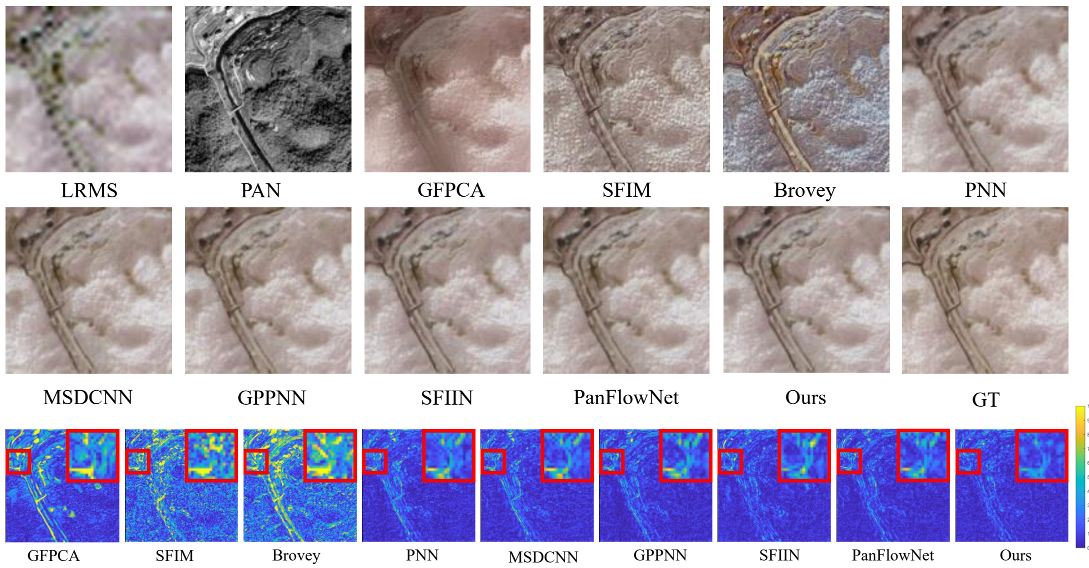


Figure 7. The result of our method compared with other methods on GanFen2 dataset.

image is cropped to 32×32. To validate the effectiveness of our method, we compare it with several state-of-the-art pan-sharpening methods, including PNN[28], PANNET[39], MSDCNN[41], SRPPNN[5], GPPNN[35], SFIINET[45], and PanFlowNet[38], as well as several traditional methods, including SFIM[26], Brovey[13], GS[24], IHS[15], and GFPCA[25].

4.2. Implementation details

In our experiments, all networks are implemented using the PyTorch framework and trained on an NVIDIA GeForce GTX 3090 GPU. During the training phase, these networks are optimized using Adam optimizer with a learning rate 1×10^{-3}. After reaching 200 epochs, the learning rate is halved. We employ commonly used evaluation metrics, including PSNR, SSIM, SAM[42], and ERGAS[3], as well as unsupervised metrics such as D_s, D_λ, and QNR[33] for real-world full-resolution scenes.

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

Evaluation on reduced-resolution scene. The evaluation results of our proposed method are presented in Table 1. The results demonstrate that our method outperforms state-of-the-art approaches in almost all metrics. Specifically, compared to the second-best method, our method achieves improvements of 0.4dB, 1.0dB, and 0.1dB in terms of PSNR on the WV2, GF2, and WV3 datasets, respectively. Similar improvements can be observed in other metrics as well. Our method almost outperforms other deep learning algorithms, validating the effectiveness of hierarchical information for the fusion process.

Furthermore, in terms of qualitative comparison, we

Table 2. Evaluation of the proposed method on real-world full-resolution scenes from the GaoFen2 dataset. The best and the second best values are highlighted in **bold** and underline.

Method	MSDCNN	SRPPNN	GPPNN	SFIINET	PanFlowNet	Ours
$D_s \downarrow$	0.0734	0.0767	0.0782	0.0724	**0.0665**	<u>0.0710</u>
$D_\lambda \downarrow$	0.1151	0.1162	0.1253	0.1230	<u>0.1113</u>	**0.1098**
QNR$\uparrow$	0.8215	0.8173	0.8073	0.8146	<u>0.8257</u>	**0.8261**

compare the results obtained by our method with other approaches on the WV2 and GF2 datasets, as shown in Figs. 6 and 7. To assess the differences between the results and the ground truth (GT), we generate residual maps to visualize the magnitude of the differences. Brighter regions in the maps indicate larger differences. It can be observed that our method exhibits the smallest differences in both spatial and spectral aspects compared to the GT, with fewer bright spots. This further demonstrates the superiority of our method over other approaches.

Evaluation on full-resolution scene and other fusion tasks. To further validate the generalization capability of our method, we conduct testing on the real-world full-resolution GF2 dataset. We first train the model on the GF2 dataset and then evaluate its performance on the real-world full-resolution GF2 dataset. Since no reference image is available, we utilize only no-reference evaluation metrics. As shown in Table 2, our method almost achieve the best performance across almost all metrics.

Additionally, we evaluate our method in other image fusion tasks including visible and infrared image fusion on RoadScene dataset and depth image SR on NYU v2 dataset using the corresponding evaluation metrics. As shown in

Table 3. Quantitative comparison on other fusion tasks. (a): Visible and infrared image fusion on RoadScene dataset with metrics MI, VIF and FMI; (b): Depth image SR on NYU v2 dataset at different ratios ($\times 4$, $\times 8$ and $\times 16$) with metric RMSE that lower values indicate higher performance. The best values are highlighted in **bold**.

Method	RoadScene		
	MI↑	VIF↑	FMI↑
DDcGAN	2.6177	0.5945	0.859
DenseFuse	3.1275	0.8025	0.868
AUIF	3.1109	0.8466	0.856
DIDFuse	3.1840	0.8274	0.853
ReCoNet	3.1594	0.7955	0.858
SDNet	3.4225	0.8207	0.863
TarDAL	3.4639	0.7871	0.852
U2Fusion	2.8109	0.7401	0.861
UMFusion	3.2018	0.7912	0.866
Ours	**4.8114**	**0.8670**	**0.878**

(a)

Method	NYU v2		
	$\times 4$	$\times 8$	$\times 16$
Bicubic	4.71	8.29	13.17
GF	5.84	7.86	12.41
TGV	3.64	10.97	39.74
DGF	3.21	5.92	10.45
DJF	2.80	5.33	9.46
DMSG	3.02	5.38	9.17
DJFR	2.38	4.94	9.18
DSRNet	3.00	5.16	8.41
PacNet	1.89	3.33	6.78
Ours	**1.53**	**3.19**	**6.44**

(b)

Table 3, our method outperforms other methods (See more details and tests in *Supplementary material*).

These experiments further demonstrate the strong generalization capability of our method, which has the ability to transfer to other fusion tasks and can serve as a general image fusion framework.

4.4. Ablation experiments

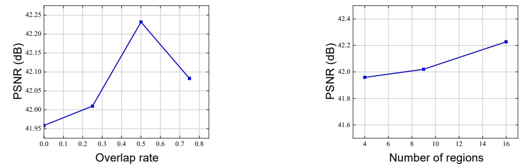
We conduct ablation experiments on the WV2 dataset to further demonstrate the validity of our approach, as shown in Table 4. The local Fourier block and integration module are the core aspects of our method. We independently conduct ablation experiments. Additionally, we also test the degree of overlap for the regions to prove that dividing images into 50% overlap four regions is a more reasonable choice.

Local Fourier block. To validate the effectiveness of the local Fourier block, we eliminate the operation of partitioning regions while keeping the parameter count unchanged. The results in Table 4 clearly demonstrate that removing the local Fourier information leads to a performance decline, thus confirming the indispensability of it. Moreover, since the parameter count remains unchanged, this also proves that the performance improvement is attributed to local Fourier information relationships of PAN and LRMS images rather than the increase in parameter count.

Integration module. The integration module consists of the SF fusion module and GL fusion module in Fig. 4. We independently remove each module to validate the rationality of these two fusion processes. As evident from Table 4, removing the SF fusion module can result in a performance decline due to the loss of spatial guidance from frequency information. Similarly, eliminating the GL fusion module led to a performance drop as the interaction between local

Table 4. Ablation studies comparison on the WorldView-II datasets. The best values are highlighted in **bold**.

Config	Local Fourier	SF fusion	GL fusion	PSNR↑	SSIM↑	SAM↓	ERGAS↓
(I)	✗	✓	✓	42.0354	0.9705	0.0219	0.9021
(II)	✓	✗	✓	40.8078	0.9627	0.0257	1.0486
(III)	✓	✓	✗	42.0655	0.9707	0.0219	0.8986
Ours	✓	✓	✓	**42.2319**	**0.9714**	**0.0215**	**0.8807**



(a) (b)

Figure 8. The region configuration results: (a) Effect of region overlap rate; (b) Effect of number of regions.

and global Fourier information is lost. This demonstrates the effectiveness of our designed integration module in promoting hierarchical information integration among the three branches, thereby enhancing the model's performance.

Region configuration. Regarding the degree of region overlap, we conduct tests with overlap sizes of 0%, 25%, 50%, and 75%, with four partitioning regions. As shown in Fig. 8a, the model achieved the highest PSNR at 50% overlap. We also test the impact of increasing the number of regions, as shown in Fig. 8b. It can be observed that as the number of regions increases, there is a slight improvement in performance. However, it is obvious that this slight improvement comes with a significant increase in the number of parameters, so it is more efficient to divide images into four regions with 50% overlap.

5. Conclusion

In this paper, we revisit spatial-frequency information integration from a hierarchical perspective for pan-sharpening, for which we propose a Hierarchical Frequency Integration Network to facilitate hierarchical Fourier information integration of PAN and LRMS images, which consists of the main module SGLI for information stratification and information integration. Extensive experiments demonstrate that our method outperforms SOTA methods and exhibits excellent generalization capabilities.

6. Acknowledgment

This work was supported by the JKW Research Funds under Grant 20-163-14-LZ-001-004-01, and the Anhui Provincial Natural Science Foundation under Grant 2108085UD12. We acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

References

[1] Paolo Addesso, Gemine Vivone, Rocco Restaino, and Jocelyn Chanussot. A data-driven model-based regression applied to panchromatic sharpening. *IEEE Transactions on Image Processing*, 29:7779–7794, 2020. 1

[2] Bruno Aiazzi, Stefano Baronti, and Massimo Selva. Improving component substitution pansharpening through multivariate regression of ms + pan data. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10):3230–3239, 2007. 3

[3] Luciano Alparone, Lucien Wald, Jocelyn Chanussot, Claire Thomas, Paolo Gamba, and Lori Mann 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. 7

[4] Coloma Ballester, Vicent Caselles, Laura Igual, Joan Verdera, and Bernard Rougé. A variational model for p+ xs image fusion. *International Journal of Computer Vision*, 69:43–58, 2006. 3

[5] 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, 2020. 3, 7

[6] Xiangyong Cao, Xueyang Fu, Danfeng Hong, Zongben Xu, and Deyu Meng. Pancsc-net: A model-driven deep unfolding method for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2021. 3

[7] Xiangyong Cao, Yang Chen, and Wenfei Cao. Proximal pannet: A model-based deep network for pansharpening. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 176–184, 2022. 3

[8] 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. 1

[9] Xueyang Fu, Zihuang Lin, Yue Huang, and Xinghao Ding. A variational pan-sharpening with local gradient constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10265–10274, 2019. 3

[10] Xueyang Fu, Wu Wang, Yue Huang, Xinghao Ding, and John Paisley. Deep multiscale detail networks for multiband spectral image sharpening. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):2090–2104, 2020. 1

[11] Ying Fu, Zhiyuan Liang, and Shaodi You. Bidirectional 3d quasi-recurrent neural network for hyperspectral image super-resolution. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:2674–2688, 2021. 3

[12] Morteza Ghahremani and Hassan Ghassemian. Nonlinear ihs: A promising method for pan-sharpening. *IEEE Geoscience and Remote Sensing Letters*, 13(11):1606–1610, 2016. 3

[13] 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. 3, 7

[14] Juan Mario Haut, Mercedes E Paoletti, Javier Plaza, Jun Li, and Antonio Plaza. Active learning with convolutional neural networks for hyperspectral image classification using a new bayesian approach. *IEEE Transactions on Geoscience and Remote Sensing*, 56(11):6440–6461, 2018. 3

[15] R Haydn. Application of the ihs color transform to the processing of multisensor data and image enhancement. In *Proc. of the International Symposium on Remote Sensing of Arid and Semi-Arid Lands, Cairo, Egypt*, 1982. 7

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[17] Xuanhua He, Keyu Yan, Rui Li, Chengjun Xie, Jie Zhang, and Man Zhou. Pyramid dual domain injection network for pan-sharpening. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12908–12917, 2023. 2, 3

[18] Junjun Jiang, Jiayi Ma, Zheng Wang, Chen Chen, and Xianming Liu. Hyperspectral image classification in the presence of noisy labels. *IEEE Transactions on Geoscience and Remote Sensing*, 57(2):851–865, 2018. 3

[19] Junjun Jiang, Jiayi Ma, and Xianming Liu. Multilayer spectral–spatial graphs for label noisy robust hyperspectral image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):839–852, 2020.

[20] Kui Jiang, Zhongyuan Wang, Peng Yi, and Junjun Jiang. A progressively enhanced network for video satellite imagery superresolution. *IEEE Signal Processing Letters*, 25(11): 1630–1634, 2018.

[21] Kui Jiang, Zhongyuan Wang, Peng Yi, Junjun Jiang, Guangcheng Wang, Zhen Han, and Tao Lu. Gan-based multi-level mapping network for satellite imagery superresolution. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 526–531. IEEE, 2019. 3

[22] Muhammad Murtaza Khan, Jocelyn Chanussot, Laurent Condat, and Annick Montanvert. Indusion: Fusion of multispectral and panchromatic images using the induction scaling technique. *IEEE Geoscience and Remote Sensing Letters*, 5(1):98–102, 2008. 3

[23] P Kwarteng and A Chavez. Extracting spectral contrast in landsat thematic mapper image data using selective principal component analysis. *Photogramm. Eng. Remote Sens*, 55(1): 339–348, 1989. 3

[24] Craig A Laben and Bernard V Brower. Process for enhancing the spatial resolution of multispectral imagery using pansharpening, 2000. US Patent 6,011,875. 7

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

[26] JG Liu. Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details. *International Journal of Remote Sensing*, 21(18): 3461–3472, 2000. 7

[27] Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989. 3

[28] Giuseppe Masi, Davide Cozzolino, Luisa Verdoliva, and Giuseppe Scarpa. Pansharpening by convolutional neural networks. *Remote Sensing*, 8(7):594, 2016. 1, 3, 7

[29] 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. 3

[30] 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. 3

[31] Xin Tian, Yuerong Chen, Changcai Yang, and Jiayi Ma. Variational pansharpening by exploiting cartoon-texture similarities. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021. 3

[32] 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. 3

[33] Gemine Vivone, Mauro Dalla Mura, Andrea Garzelli, Rocco Restaino, Giuseppe Scarpa, Magnus O Ulfarsson, Luciano Alparone, and Jocelyn Chanussot. A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods. *IEEE Geoscience and Remote Sensing Magazine*, 9(1):53–81, 2020. 7

[34] Lucien Wald, Thierry Ranchin, and Marc Mangolini. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogrammetric engineering and remote sensing*, 63(6):691–699, 1997. 6

[35] Shuang Xu, Jiangshe Zhang, Zixiang Zhao, Kai Sun, Junmin Liu, and Chunxia Zhang. Deep gradient projection networks for pan-sharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1366–1375, 2021. 3, 7

[36] 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. 3

[37] 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. 3

[38] Gang Yang, Xiangyong Cao, Wenzhe Xiao, Man Zhou, Aiping Liu, Xun Chen, and Deyu Meng. Panflownet: A flow-based deep network for pan-sharpening. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16857–16867, 2023. 7

[39] Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, and John Paisley. Pannet: A deep network architecture for pan-sharpening. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5449–5457, 2017. 7

[40] Hu Yu, Jie Huang, Feng Zhao, Jinwei Gu, Chen Change Loy, Deyu Meng, Chongyi Li, et al. Deep fourier up-sampling. *Advances in Neural Information Processing Systems*, 35:22995–23008, 2022. 3

[41] 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. 7

[42] Roberta H Yuhas, Alexander FH Goetz, and Joe W Boardman. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. In *JPL, Summaries of the Third Annual JPL Airborne Geoscience Workshop. Volume 1: AVIRIS Workshop*, 1992. 7

[43] 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*, pages 3553–3561, 2022. 3

[44] 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*, pages 3375–3384, 2022. 2, 3

[45] 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 *European Conference on Computer Vision*, pages 274–291. Springer, 2022. 2, 3, 4, 7

[46] 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. 3