Mutual Information-driven Pan-sharpening

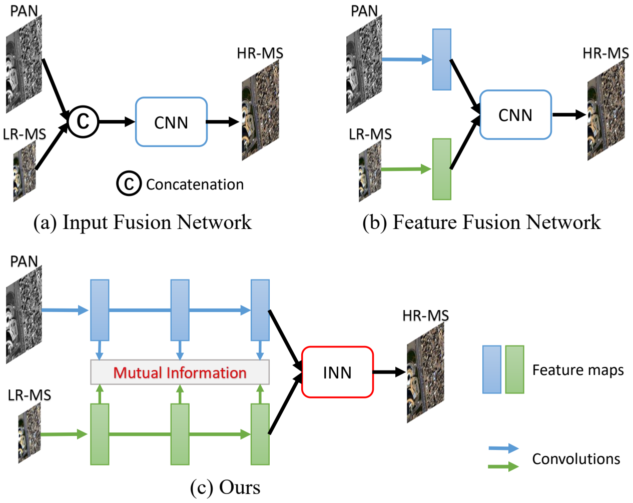
Man Zhou[1,2]*, Keyu Yan[1,2]*, Jie Huang[2], Zihe Yang[2], Xueyang Fu[2]†, Feng Zhao[2]
[1]Hefei Institute of Physical Science, Chinese Academy of Sciences, China
[2]University of Science and Technology of China, China
{manman,keyu,hj0117,yangzh10}@mail.ustc.edu.cn, {xyfu,fzhao956}@ustc.edu.cn

Abstract

*Pan-sharpening aims to integrate the complementary information of texture-rich PAN images and multi-spectral (MS) images to produce the texture-rich MS images. Despite the remarkable progress, existing state-of-the-art Pan-sharpening methods don't **explicitly** enforce the complementary information learning between two modalities of PAN and MS images. This leads to information redundancy not being handled well, which further limits the performance of these methods. To address the above issue, we propose a novel mutual information-driven Pan-sharpening framework in this paper. To be specific, we first project the PAN and MS image into modality-aware feature space independently, and then impose the mutual information minimization over them to **explicitly** encourage the complementary information learning. Such operation is capable of reducing the information redundancy and improving the model performance. Extensive experimental results over multiple satellite datasets demonstrate that the proposed algorithm outperforms other state-of-the-art methods qualitatively and quantitatively with great generalization ability to real-world scenes.*

1. Introduction

With the rapid development of remote sensors, explosive satellite images are available for a wide range of applications like military systems, environmental monitoring, and mapping services. Due to the physical limitations, satellites usually capture both multi-spectral (MS) and panchromatic (PAN) sensors to simultaneously obtain complementary information. To be specific, MS images possess high spectral but limited spatial resolution while PAN images have rich spatial information but low spectral resolution. To generate images with both high spectral and spatial resolutions, the Pan-sharpening technique by fusing the MS and PAN images has drawn increasing attention from both image processing and remote sensing communities.



(a) Input Fusion Network (b) Feature Fusion Network

(c) Ours

Figure 1. The categorization of existing Pan-sharpening methods.

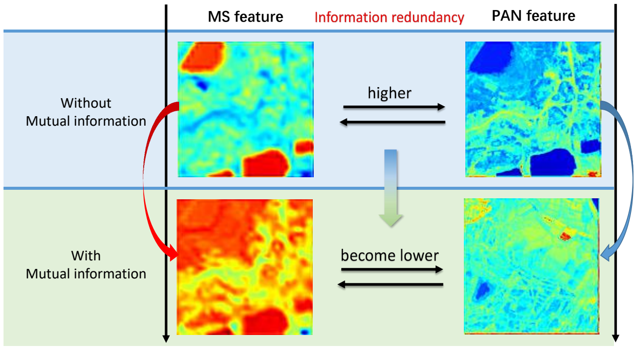


Figure 2. The information redundancy reduction after integrating mutual information minimization constraint, and the average features of P_2 and M_2 are visualized, defined in Figure 4.

Treated as a fusion task, considerable Pan-sharpening methods have been developed with two main fusion strategies: 1) image-level fusion and 2) feature-level fusion. As shown in Figure 1 (a), the first category directly concatenates the MS and PAN images along the channel dimension

*Co-first authors contributed equally, † corresponding author. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61901433; and in part by the USTC Research Funds of the Double First-Class Initiative under Grant YD2100002003.

before feeding them into the networks. Without conducting explicitly cross-modal fusion, the 'input fusion' strategy is therefore limited in studying the complementary information, leading to unsatisfactory performance. The second category attempts to extract the modality-aware features from PAN and MS images independently, and then performs the information fusion in feature space, as shown in Figure 1 (b). Although encouraging improvement has been achieved, it still suffers from the following issue. Since PAN and MS images capture the same scene in different modalities, they contain shared information as well as unique features, as demonstrated in Figure 3. However, existing state-of-the-art Pan-sharpening methods don't explicitly enforce the complementary information learning between two modalities of PAN and MS images, resulting in the redundancy of learned features and the so-called copy artifacts [5, 12, 46]. Considering the limitation of the current methods, in this paper, we make our efforts to enforce the complementary feature learning and reduce the information redundancy for improving the Pan-sharpening performance.

As shown in Figure 1 (c), we propose a novel mutual information-driven Pan-sharpening framework in a cascaded manner, and the detailed flowchart is illustrated in Figure 4. The MS and PAN images are firstly fed into two independent convolution branches for obtaining the modality-aware features, and then we impose the mutual information minimization over them to encourage the complementary information learning from the shallow to deep levels. To be specific, the obtained modality-aware features are further converted to low-dimensional feature vectors for calculating the mutual information where the latter-level feature vectors are obtained depending on two-folds: 1) the current-layer modality features and 2) the previous-layer immediate-process feature in feature vector calculating. Such operation is capable of reducing the information redundancy, visualized in Figure 2. After obtaining the refined features, a post-fusion module is devised for projecting them back to the expected MS images by equipping with effective invertible neural networks. Extensive experimental results over multiple satellite datasets demonstrate that the proposed algorithm outperforms other state-of-the-art methods qualitatively and quantitatively with great generalization ability to real-world scenes. Ablation studies also verify the effectiveness of the proposed components.

The contributions of this work are as follows:

- We design a novel Pan-sharpening framework by mutual information minimization in a cascaded manner. To the best of our knowledge, this is the first attempt to explicitly encourage the multi-modal learning between MS and PAN modalities. The proposed model reduces the information redundancy and alleviate the artifacts in Pan-sharpening.

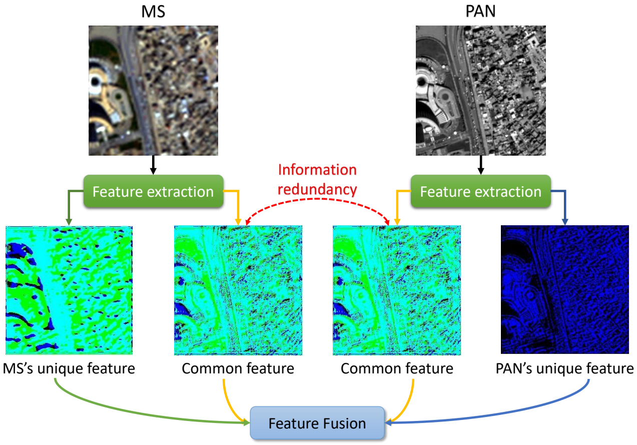


Figure 3. The decomposed components of the PAN and its corresponding MS image by the technique [5] .

- Extensive experimental results over multiple satellite datasets demonstrate the superiority of the proposed algorithm against other state-of-the-art methods. The great generalization ability is also verified over real-world full-resolution satellite scenes.

2. Related work

2.1. Classic Pan-sharpening methods

In the past few years, many classic Pan-sharpening methods have been proposed in an attempt to fuse the low-resolution multi-spectral (LR-MS) image and PAN image to obtain the high-resolution multi-spectral (HR-MS) image. A common way of dividing is to divide classic Pan-sharpening methods into the following categories: component substitution-based (CS) methods, multiresolution analysis-based (MRA) methods and variational optimization-based (VO) approaches [10, 14, 37, 38]. The core idea of the CS methods [9, 35] is to replace the spatial component of the LR-MS image with the component extracted from the PAN image. Generally, the CS methods consume less time than other classic methods but the sharpened image often has apparent spectral distortion. To reduce spectral distortion, the MRA methods [29, 34] construct the HR-MS image by injecting high spatial details extracted from the PAN image through multi-resolution decomposition technique into the LR-MS image. However, because high frequency details are injected in the transform domain, there are often frequency aliasing problems in actual use. In addition, there are also some hybrid methods [23, 56] that combine CS methods and MRA methods, trying to use the advantages of the two to complement each other. In the recent past, many VO approaches [2, 39, 40] have emerged as their good performance in the field of Pan-sharpening. The approaches are designed to find an optimized function

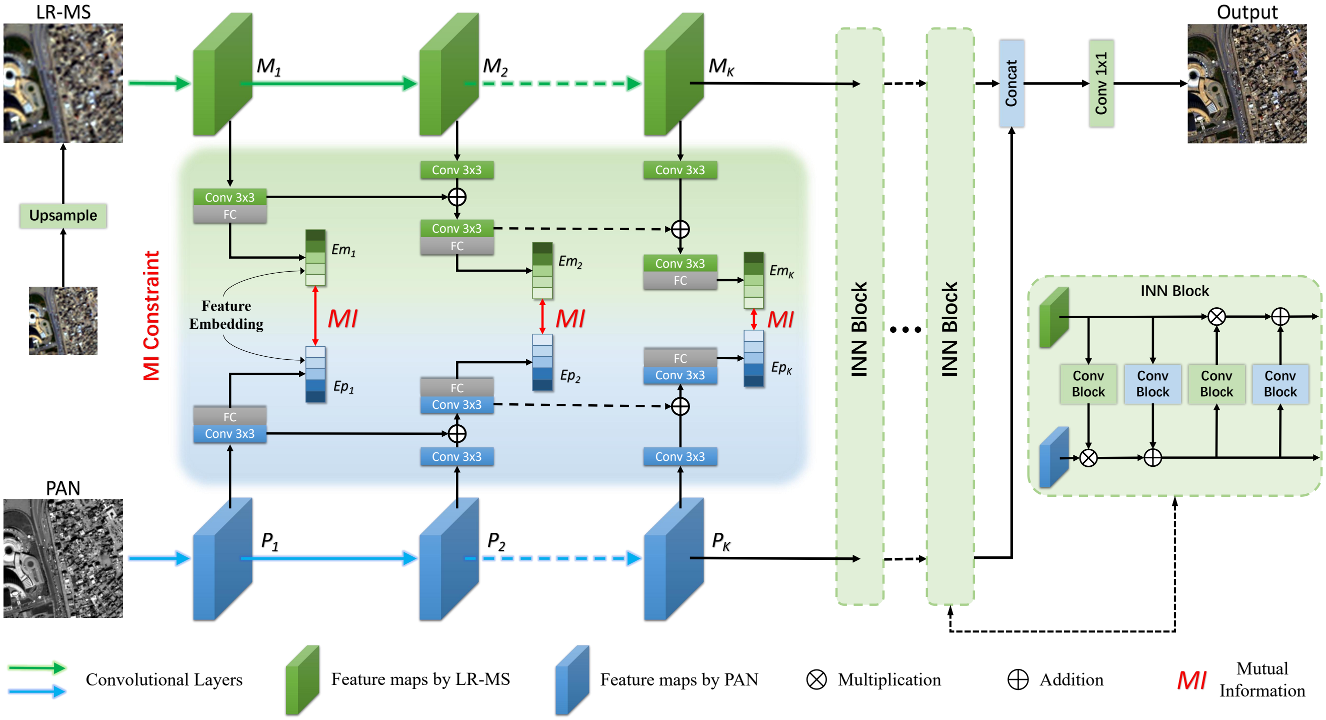


Figure 4. The detailed flowchart of our proposed method. PAN and its corresponding MS images are firstly projected into modality-aware feature maps as P_1, P_2, ..., P_K and M_1, M_2, ..., M_K from shallow to deep levels respectively. Next, we transform them to low-dimension feature embedding Ep_1, Ep_2,..., Ep_K and Em_1, Em_2,..., Em_K, and impose the mutual information minimization over them to explicitly encourage the complementary information learning. Above operation is capable of reducing the information redundancy and alleviating the artifacts. Finally, an effective post-fusion module is devised to project the features back to the expected MS images.

through certain prior constraints or assumptions. Unfortunately, it is a huge challenge for VO approaches to choose the appropriate prior constraint and reasonable hypothesis.

2.2. Deep learning based methods

Nowadays, with the success of deep learning-based (DL) methods in the field of hyperspectral image [7, 8, 15, 21, 43] and remote sensing image [6, 19, 20, 26, 45], the DL methods [5, 47, 54, 58] have also begun to be used in Pan-sharpening and make a great improvement. A famous DL method called PNN [31] is based on a three-layer convolutional neural network. Subsequently, PANNet [50] introduces the high-pass filtering domain in the training process to preserve spatial information. MSDCNN [52] takes into account the problem of multi-scale, adding multi-scale modules to the network to promote performance. Furthermore, based on SRCNN [13], Cai *et al.* [3] applied super-resolution method to Pan-sharpening. The networks mentioned above are designed based on the residual block which limit the reuse of shallow network features. Wang *et al.* [44] adopted U-shaped network to solve this problem. Moreover, WSDFNet [22] propagate shallow features by adaptive skip weighter to deep layers. Additionally, there are also some networks based on generative adversarial models such as Pan-GAN [28]. Model-driven deep networks for Pan-sharpening like GPPNN [48] increase the model interpretability but performance has decreased.

2.3. Mutual information

InfoMax principle motivates explosive representation learning researcher works where they maximize the mutual information to achieve the effective representation. The work [41] provides an empirical evidence about the connection and application direction over three folds: 1) global features and local features of the same image, 2) multiple views of different image modality over the same scene, and 3) a sequential component of the data. Since then, Zhang *et.al* [55] introduce mutual information minimization to explicitly encourage the multi-modal information learning between RGB image and depth data. Sanghi *et.al* [33] maximize the mutual information between 3D objects and their geometric transformed versions to improve the representations. However, the information redundancy naturally exists in Pan-sharpening task and leads to the so-called copy artifacts [5, 12, 46]. To this end, we introduce the mutual information minimization between two modalities PAN and MS image to encourage the multi-modal learning.

3. Methods

We denote PAN image as $P \in R^{H \times W \times 1}$ and its corresponding MS image is firstly up-sampled with the same spatial resolution of P as $M \in R^{H \times W \times C}$.

3.1. Model architecture

As detailed in Figure 4, our proposed method consists of three modules: modality-aware feature extraction of PAN and MS images, mutual information constraint and the post-fusion module based on Invertible neural networks (INN).

Modality-aware feature extraction. We firstly employ two independent feature extraction branches with the basic component of convolution layer with kernel size 3×3 to project the PAN and MS images to modality-aware feature maps from shallow to deep levels. Specifically, these features are denoted as $P_1, P_2, \dots, P_K$ and $M_1, M_2, \dots, M_K$ respectively. Both of them are equipped with the size of $H \times W \times C$. Suppose that two branches denote as f_p and f_m, the process can be written as

$$P_1, P_2, \dots, P_K = f_p(P), \tag{1}$$

$$M_1, M_2, \dots, M_K = f_m(M). \tag{2}$$

Mutual information. Referring to the above PAN features $P_1, P_2, \dots, P_K$ and MS features $M_1, M_2, \dots, M_K$, we firstly transform them into low-dimensional feature vectors for preparing the mutual information. Particularly, the first-layer features P_1 and M_1 are fed into an additional convolution layer with kernel size of 3×3 for channel dimension as P_T^1 and M_T^1, and then followed by two fully-connected layers that receives above reshaped features to obtain the low-dimensional feature vectors Ep_1 and Em_1

$$P_T^1, M_T^1 = C_3(P_1), C_3(M_1), \tag{3}$$

$$Ep_1, Em_1 = FCs(P_T^1), FCs(M_T^1), \tag{4}$$

where C_3, FCs represent the convolution layers of kernel size 3×3 and full-connected layers respectively. In terms of the latter-layer features, taking the i-layer for example, the feature vectors are obtained by combining the previous intermediate feature transformation P_T^{i-1}, M_T^{i-1} and the current modality features P_i, M_i. Particularly, it follows three steps: 1) P_i and M_i are fed into two different convolution layers for channel reduction; 2) the reduction features are added with previous intermediate features P_T^{i-1} and M_T^{i-1}; 3) the obtained features are further passed through the convolution layers and two fully-connected layers to generate the low-dimensional feature vectors Ep_i and Em_i as

$$P_T^i = C_3(C_3(P_i) + P_T^{i-1}), \tag{5}$$

$$M_T^i = C_3(C_3(M_i) + M_T^{i-1}), \tag{6}$$

$$Ep_i, Em_i = FCs(P_T^i), FCs(M_T^i), \tag{7}$$

where C_3 denotes the convolution with kernel size 3×3 in no-sharing weights manner. Finally, given the modality-aware feature vectors Ep_1, Ep_2, ..., Ep_K of PAN image and Em_1, Em_2, ..., Em_K of MS image, we introduce the mutual information minimization to enforce the complementary information learning of two modalities, thus reducing the information redundancy.

In Information theory, mutual information aims to measure the amount of information obtained about a random variable Ep_i by observing some other random variable Em_i or vice versa as

$$MI(Ep_i, Em_i) = H(Ep_i) - H(Ep_i|Em_i), \tag{8}$$

$$MI(Em_i, Ep_i) = H(Em_i) - H(Em_i|Ep_i), \tag{9}$$

where

$$H(Em_i, Ep_i) = H(Ep_i) + H(Em_i|Ep_i), \tag{10}$$

$$H(Em_i, Ep_i) = H(Em_i) + H(Ep_i|Em_i). \tag{11}$$

Combing above two equations, we can obtain

$$H(Em_i) - H(Em_i|Ep_i) = H(Ep_i) - H(Ep_i|Em_i), \tag{12}$$

where $H(.)$ represents the entropy, $i \in [1, K]$ with K being the stage number of feature extraction, $H(Em_i)$, $H(Ep_i)$ indicate the marginal entropies, $H(Ep_i, Em_i)$ and and $H(Em_i, Ep_i)$ are the joint entropy, $H(Ep_i|Em_i)$ and $H(Em_i|Ep_i)$ are the conditional entropy. Afterward, integrating above equations, we can infer

$$MI(Ep_i, Em_i) = H(Ep_i) + H(Em_i) - H(Ep_i, Em_i), \tag{13}$$

where we also introduce Kullback-Leibler divergence (KL) to calculate the entropy following previous works [33,41]

$$H(Ep_i) = -\sum Ep_i \log(Em_i) - KL(Ep_i||Em_i), \tag{14}$$

$$H(Em_i) = -\sum Em_i \log(Ep_i) - KL(Em_i||Ep_i). \tag{15}$$

Note that we enforce the $MI(Ep_i, Em_i)$ calculation over $i = 1, \dots K$ levels, shown in Figure 4.

INN block. With the mutual information minimization, the redundancy of modality features is reduced. Followed, we design an effective post-fusion module based on invertible neural networks. The basic component is the coupling layer proposed in [36] and stacked for effectively fusing above the refined modality features [30,57], thus projecting back to the expected MS images. Deepening into the coupling layer, the convolution block is implemented by Half-Instance normalization module [11].

3.2. Optimization

As shown in Figure 4, the overall loss function consists of two parts: one for the reconstructing the ground-truth MS image by $L1$ loss and the other for reducing the information redundancy between two modalities, written as:

$$L = \|f(M,P) - H\|_1 + \lambda \sum_{i=1}^{K} MI(Ep_i, Em_i), \tag{16}$$

where $f(.)$ denotes the mapping function of our method, $MI(.)$ indicates mutual information calculating. H is the ground truth MS image and λ is the parameters to balance the two terms in loss function. In our setting, λ is set as 0.1.

4. Experiments

In this section, we conduct extensive experiments over three satellite image datasets of the WorldView II, WorldView III, and GaoFen2 to evaluate the model performance.

4.1. Datasets and benchmark

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 [42]. Specifically, given the MS image $H \in R^{M \times N \times C}$ and the PAN image $P \in R^{rM \times rN \times b}$, both of them are downsampled with ratio r, and then are denoted by $L \in R^{M/r \times N/r \times C}$ and $p \in R^{M \times N \times b}$ 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 the WorldView II, GaoFen2 and WorldView III are adopted to construct image datasets. For each database, PAN images are cropped into patches with the size of 128 × 128 pixels while the corresponding MS patches are with the size of 32 × 32 pixels.

To evaluate the results of our proposed method, several commonly-recognized state-of-the-art Pan-sharpening methods are selected, which are classified into two-folds: 1) five representative deep-learning based methods, PNN [32], PANNET [51], MSDCNN [53], SRPPNN [4], and GPPNN [49]; 2) five promising traditional methods, SFIM [27], Brovey [16], GS [24], IHS [17], and GFPCA [25].

4.2. Implementation details and metrics

We implement our networks in PyTorch framework on the PC with a single NVIDIA GeForce GTX 2080Ti GPU. In the training phase, they are optimized by Adam optimizer over 1000 epochs with a batch size of 4. The learning rate is initialized with 8×10^{-4} and decayed by multiplying 0.5 when reaching 200 epochs. Several widely-used image quality assessment (IQA) metrics are adapted for performance measurement, including the PSNR, SSIM, SAM [18], ERGAS [1], the three non-reference metrics of D_λ, D_S, QNR for real-world full-resolution scenes.

Table 1. The quantitative results on WorldView-II datasets. The best values are highlighted by the red bold. The up or down arrow indicates higher or lower metric corresponds to better images.

Method	WorldView II			
	PSNR↑	SSIM↑	SAM↓	ERGAS↓
SFIM	34.1297	0.8975	0.0439	2.3449
Brovey	35.8646	0.9216	0.0403	1.8238
GS	35.6376	0.9176	0.0423	1.8774
IHS	35.2926	0.9027	0.0461	2.0278
GFPCA	34.5581	0.9038	0.0488	2.1411
PNN	40.7550	0.9624	0.0259	1.0646
PANNET	40.8176	0.9626	0.0257	1.0557
MSDCNN	41.3355	0.9664	0.0242	0.9940
SRPPNN	41.4538	0.9679	0.0233	0.9899
GPPNN	41.1622	0.9684	0.0244	1.0315
Ours	**41.6773**	**0.9705**	**0.0224**	**0.9519**

Table 2. The quantitative results on GaoFen2 test datasets. The best values are highlighted by the red bold.

Method	GaoFen2			
	PSNR↑	SSIM↑	SAM↓	ERGAS↓
SFIM	36.9060	0.8882	0.0318	1.7398
Brovey	37.7974	0.9026	0.0218	1.3720
GS	37.2260	0.9034	0.0309	1.6736
IHS	38.1754	0.9100	0.0243	1.5336
GFPCA	37.9443	0.9204	0.0314	1.5604
PNN	43.1208	0.9704	0.0172	0.8528
PANNET	43.0659	0.9685	0.0178	0.8577
MSDCNN	45.6874	0.9827	0.0135	0.6389
SRPPNN	47.1998	0.9877	0.0106	0.5586
GPPNN	44.2145	0.9815	0.0137	0.7361
Ours	**47.3042**	**0.9892**	**0.0102**	**0.5481**

Table 3. Comparisons on flops and parameter numbers.

	PNN	PANNET	MSDCNN	SRPPNN	GPPNN	Ours
params	0.0689	0.0688	0.2390	1.7114	0.1198	0.0714
flops	1.1289	1.1275	3.9158	21.1059	1.3967	1.1815

Table 4. The quantitative results on WorldView-III test datasets. The best values are highlighted by the red bold.

Method	WorldView III			
	PSNR↑	SSIM↑	SAM↓	ERGAS↓
SFIM	21.8212	0.5457	0.1208	8.9730
Brovey	22.5060	0.5466	0.1159	8.2331
GS	22.5608	0.5470	0.1217	8.2433
IHS	22.5579	0.5354	0.1266	8.3616
GFPCA	22.3344	0.4826	0.1294	8.3964
PNN	29.9418	0.9121	0.0824	3.3206
PANNET	29.6840	0.9072	0.0851	3.4263
MSDCNN	30.3038	0.9184	0.0782	3.1884
SRPPNN	30.4346	0.9202	0.0770	3.1553
GPPNN	30.1785	0.9175	0.0776	3.2593
Ours	**30.4907**	**0.9223**	**0.0749**	**3.1125**

Table 5. The non-reference metrics on full-resolution dataset. The best values are highlighted by the red bold.

Method	Full-resolution Dataset		
	$D_\lambda \downarrow$	$D_s \downarrow$	QNR↑
PNN	0.0746	0.1164	0.8191
PANNET	0.0737	0.1224	0.8143
MSDCNN	0.0734	0.1151	0.8215
SRPPNN	0.0767	0.1162	0.8173
GPPNN	0.0782	0.1253	0.8073
Ours	**0.0694**	**0.1118**	**0.8259**

Table 6. The effect of weighting parameter λ in loss function.

λ	0.01	0.05	0.1	0.5	1
PSNR	41.5859	41.6079	41.6773	41.5581	41.4867

4.3. Parameter Numbers vs model performance

In this section, we investigate the comparisons on parameter numbers and model performance (representation by PSNR) are shown in Table 3. It can be seen that our network is able to achieve a good trade-off and achieves the best performance with comparably fewer parameters compared to other deep learning-based methods. We use the tensor with $1 \times 4 \times 32 \times 32$ and $1 \times 1 \times 128 \times 128$ to represent the MS and PAN roles for evaluation.

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

4.4.1 Evaluation on reduced-resolution scene

Quantitative comparison. The comparison results over three satellite datasets are reported in Table 1, Table 2 and Table 4 respectively where the best values are highlighted by red bold. As can be seen clearly, our proposed method achieves the best overall results than other promising Pan-sharpening methods over all the satellite datasets. Specifically, the average gains of our method over the second-best SRPPNN are 0.24dB, 0.16dB, 0.10dB in reference metric PSNR on WorldView-II, GaoFen2 and WorldView-III datasets, respectively. In addition to PSNR, consistent improvements can be observed in the other metrics, indicating the lower spectral distortion and spatial texture preservation. When compared with the other approaches, our method is far ahead.

Qualitative comparison. We also show the qualitative comparison of the visual results to testify the effectiveness of our method in Figure 5 and Figure 6 over the representative samples from WorldView-II and WorldView-III dataset. Images in the last row are the MSE residuals between the output pan-sharpened results and the ground truth. Compared with other competing methods, our model has minor spatial and spectral distortions. As for the MSE residuals, it's noticed that our proposed method is closest to the ground truth than other comparison methods. The state-of-the-art performance of our method demonstrate the effectiveness of the proposed mutual information minimization mechanism, which is capable of reducing the information redundancy and improving the Pan-sharpening results.

4.4.2 Evaluation on full-resolution scene

In order to demonstrate the real-world application value, we further perform experiments on 200 sets of full-resolution data obtained by Gaofen2. Due to the unavailability of ground-truth MS images in the real-world full-resolution scenes, the commonly-used three non-reference metrics of D_λ, D_s and QNR are adapted for evaluation. The quantitative comparison between representative CNN-based methods and our method are shown in Table 5. The lower D_λ, D_s and the higher QNR correspond to the better image quality where the best results are remarked by red bold. As can be seen clearly, our methods surpass other competitive Pan-sharpening methods in all the indexes.

4.5. Ablation experiments

We conduct the ablation studies to further verify the components of our model, including the mutual information minimization (MI), the invertible neural network-base fusion module (INN) and the weighting parameter λ of the overall loss. The visual comparison is shown in Figure 8,

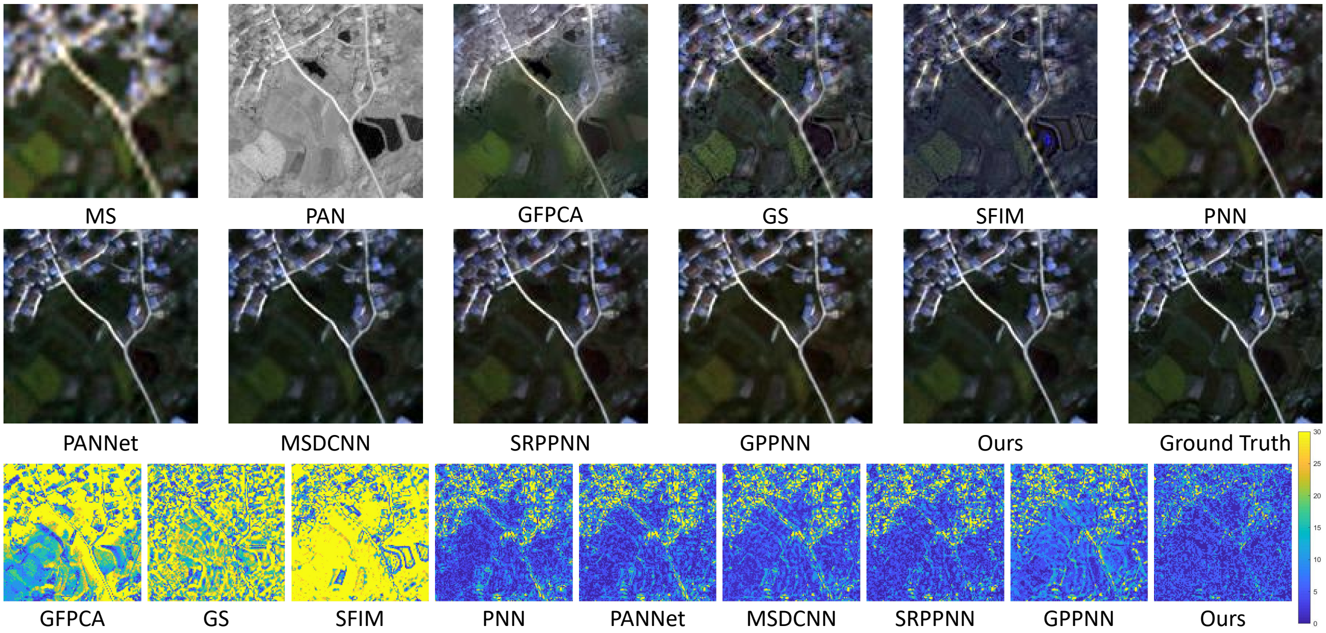


Figure 5. Visual comparisons over WorldView-II.

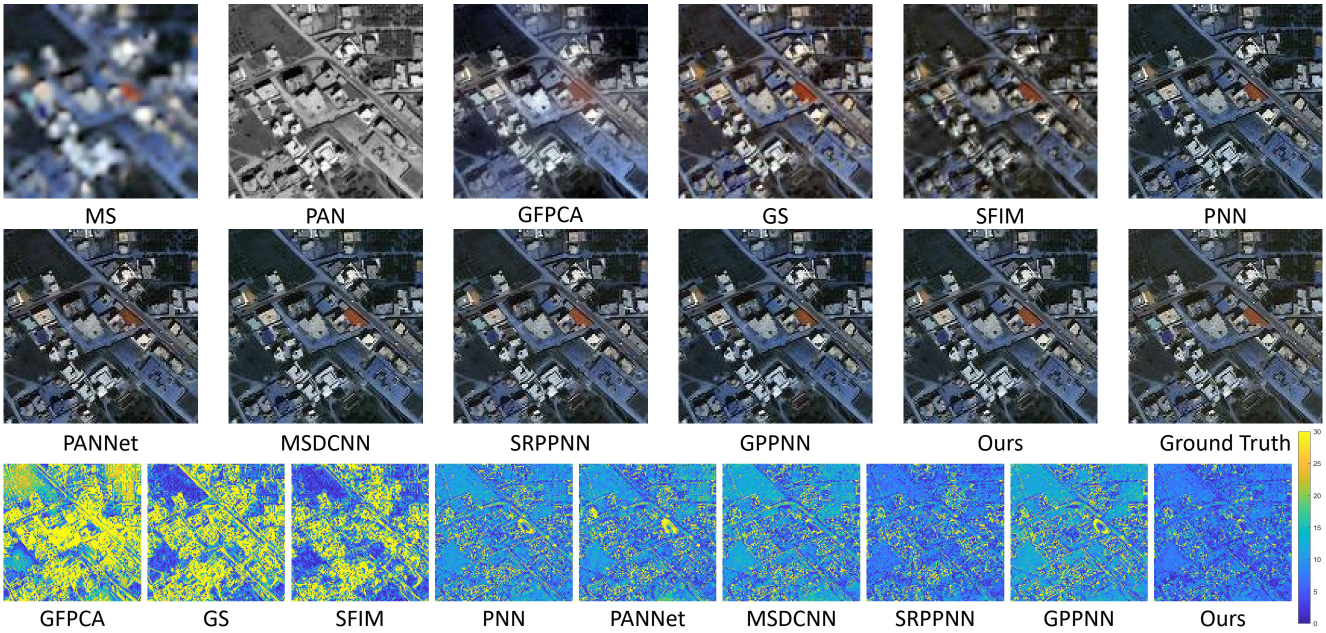


Figure 6. Visual comparisons over WorldView-III.

where "ours" indicate the whole network, (I) and (II) are the variants of "ours" by deleting the MI and replacing INN.

The mutual information minimization. In the first row of Table 7, we delete the mutual information minimization constraint to verify its necessity. Table 7 shows that deleting it will degrade all the metrics dramatically, thus verifying its positive effect to reduce the information redundancy. The visual comparisons in Figure 8 between (I) and "ours" also testify the vital importance of this component.

The invertible neural network-based fusion module. In the second row of Table 7, we replace INN with pure ResNet block under the constraint of the same numbers of parameters. The results demonstrate that replacing it will weaken our network's performance, indicating its effectiveness. The visual comparisons in Figure 8 between (II) and "ours" also verify its necessity.

Table 7. The results of ablation experiments over three datasets. The best values are highlighted by the red bold. “MI” and “INN” represent the components of mutual information and the post-fusion invertible neural module respectively.

Config	MI	INN	WorldView II				GaoFen2				WorldView III			
			PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	EGAS↓	PSNR↑	SSIM↑	SAM↓	EGAS↓
(I)	✗	✓	41.4940	0.9685	0.0231	0.9711	47.2316	0.9878	0.0117	0.5492	30.2893	0.9192	0.0789	3.1584
(II)	✓	✗	41.5863	0.9699	0.0228	0.9620	47.2775	0.9885	0.0104	0.5488	30.3511	0.9214	0.0785	3.1311
Ours	✓	✓	**41.6773**	**0.9705**	**0.0224**	**0.9519**	**47.3042**	**0.9892**	**0.0102**	**0.5481**	**30.4907**	**0.9223**	**0.0749**	**3.1125**

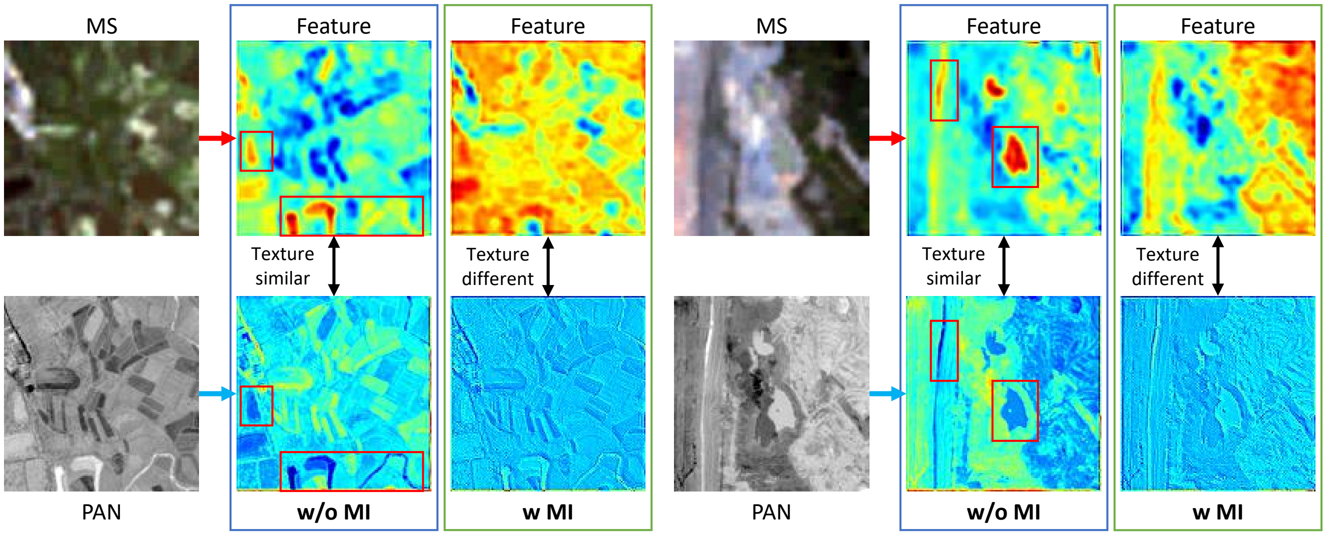


Figure 7. The features difference of PAN and MS in our model without (w/o) and with (w) integrating the mutual information constraint.

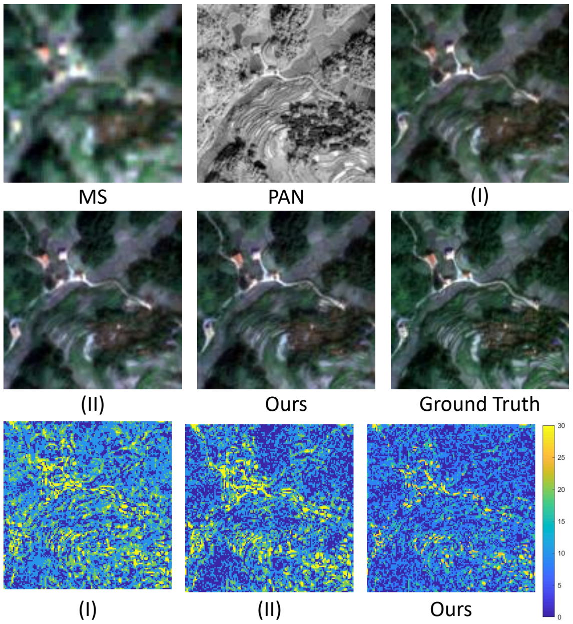


Figure 8. The visual comparison over GaoFen2 satellite.

The weighting parameter λ. In Table 6, we show the different values of weighting parameters λ to verify its effect over WorldView-II. With its increasing, the effect of mutual information becomes larger. Moreover, being smaller is inadequate of reducing information redundancy while being larger will destroy the connection between two-modalities. It is obvious that the setting as 0.1 is the optimal solution.

4.6. Visualization of feature maps

To verify the effect of mutual information (MI) constraint, we show the change of modality features before and after integrating with mutual information constraint in Figure 7 and Figure 2. It is clearly seen that integrating MI enforces the complementary features learning and reduces the information redundancy, especially in red box. To be specific, the PAN features focus more on texture details while MS features focus more on spectral characteristics.

5. Conclusion

In this paper, we propose a novel Pan-sharpening framework. Specifically, we introduce the mutual information minimization regularization to reduce the information redundancy between two modalities of PAN and MS images. To the best of our knowledge, this is the first attempt to explicitly encourage the complementary information learning. Extensive experimental results over multiple satellites demonstrate the effectiveness of the proposed algorithm.

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. 5

[2] 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(1):43–58, 2006. 2

[3] 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

[4] 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. 5

[5] 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*, pages 1–13, 2021. 2, 3

[6] Xiangyong Cao, Xueyang Fu, Chen Xu, and Deyu Meng. Deep spatial-spectral global reasoning network for hyperspectral image denoising. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–14, 2021. 3

[7] Xiangyong Cao, Jing Yao, Zongben Xu, and Deyu Meng. Hyperspectral image classification with convolutional neural network and active learning. *IEEE Transactions on Geoscience and Remote Sensing*, 58(7):4604–4616, 2020. 3

[8] Xiangyong Cao, Feng Zhou, Lin Xu, Deyu Meng, Zongben Xu, and John Paisley. Hyperspectral image classification with markov random fields and a convolutional neural network. *IEEE Transactions on Image Processing*, 27(5):2354–2367, 2018. 3

[9] Wjoseph Carper, Thomasm Lillesand, and Ralphw Kiefer. The use of intensity-hue-saturation transformations for merging spot panchromatic and multispectral image data. *Photogrammetric Engineering and remote sensing*, 56(4):459–467, 1990. 2

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

[11] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration, 2021. 4

[12] Xin Deng and Pier Luigi Dragotti. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3333–3348, 2021. 2, 3

[13] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE image prior migration on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 3

[14] 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. 2

[15] 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

[16] 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. 5

[17] 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. 5

[18] A. F. Goetz J. R. H. Yuhas and J. M. 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. 5

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

[20] Kui Jiang, Zhongyuan Wang, Peng Yi, Junjun Jiang, Emily Xiao, and Yuan Yao. Deep distillation recursive network for remote sensing imagery super-resolution. *Remote Sensing*, 10:1700, 10 2018. 3

[21] Kui Jiang, Zhongyuan Wang, Peng Yi, Guangcheng Wang, Tao Lu, and Junjun Jiang. Edge-enhanced gan for remote sensing image superresolution. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8):5799–5812, 2019. 3

[22] Zi-Rong Jin, Tian-Jing Zhang, Cheng Jin, and Liang-Jian Deng. Weighted shallow-deep feature fusion network for pansharpening. *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 2632–2635, 2021. 3

[23] Chiman Kwan, Joon Hee Choi, Stanley Chan, Jin Zhou, and Bence Budavari. A super-resolution and fusion approach to enhancing hyperspectral images. *Remote Sensing*, 10, 09 2018. 2

[24] C.A. Laben and B.V. Brower. Process for enhancing the spatial resolution of multispectral imagery using pansharpening. *US Patent 6011875A*, 2000. 5

[25] 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. 5

[26] Junmin Liu, Shijie Li, Changsheng Zhou, Xiangyong Cao, Yong Gao, and Bo Wang. Sraf-net: A scene-relevant anchor-free object detection network in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–1, 2021. 3

[27] 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. 5

[28] 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. 3

[29] SG Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989. 2

[30] Zhou Man, Huang Jie, Fang Yanchi, Fu Xueyang, and Liu Aiping. Pan-sharpening with customized transformer and invertible neural network. *AAAI Conference on Artificial Intelligence (AAAI)*, 2022. 4

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

[32] Giuseppe Masi, Davide Cozzolino, Luisa Verdoliva, and Giuseppe Scarpa. Pansharpening by convolutional neural networks. *Remote Sensing*, 8(7), 2016. 5

[33] Aditya Sanghi. Info3d: Representation learning on 3d objects using mutual information maximization and contrastive learning. *CoRR*, abs/2006.02598, 2020. 3, 4

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

[35] Vijay P Shah, Nicolas H Younan, and Roger L King. An efficient pan-sharpening method via a combined adaptive pca approach and contourlets. *IEEE image prior migration on geoscience and remote sensing*, 46(5):1323–1335, 2008. 2

[36] Takeshi Teshima, Isao Ishikawa, Koichi Tojo, Kenta Oono, Masahiro Ikeda, and Masashi Sugiyama. Coupling-based invertible neural networks are universal diffeomorphism approximators. *Neural Information Processing Systems (NeurIPS 2020)*, 2020. 4

[37] Xin Tian, Yuerong Chen, Changcai Yang, Xun Gao, and Jiayi Ma. A variational pansharpening method based on gradient sparse representation. *IEEE Signal Processing Letters*, 27:1180–1184, 2020. 2

[38] Xin Tian, Yuerong Chen, Changcai Yang, and Jiayi Ma. Variational pansharpening by exploiting cartoon-texture similarities. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–16, 2021. 2

[39] Xin Tian, Yuerong Chen, Changcai Yang, and Jiayi Ma. Variational pansharpening by exploiting cartoon-texture similarities. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–16, 2021. 2

[40] Xin Tian, Kun Li, Zhongyuan Wang, and Jiayi Ma. Vp-net: An interpretable deep network for variational pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–16, 2021. 2

[41] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic. On mutual information maximization for representation learning. *International Conference on Learning Representations*, 2019. 3, 4

[42] 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. 5

[43] Xinya Wang, Jiayi Ma, and Junjun Jiang. Hyperspectral image super-resolution via recurrent feedback embedding and spatial-spectral consistency regularization. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–13, 2021. 3

[44] Yudong Wang, Liang-Jian Deng, Tian-Jing Zhang, and Xiao Wu. Ssconv: Explicit spectral-to-spatial convolution for pansharpening. In *Proceedings of the 29th ACM International Conference on Multimedia (ACM MM)*, page DOI: 10.1145/3474085.3475600., 2021. 3

[45] Zhongyuan Wang, Kui Jiang, Peng Yi, Zhen Han, and Zheng He. Ultra-dense gan for satellite imagery super-resolution. *Neurocomputing*, 398:328–337, 2020. 3

[46] Han Xu, Jiayi Ma, Zhenfeng Shao, Hao Zhang, Junjun Jiang, and Xiaojie Guo. Sdpnet: A deep network for pan-sharpening with enhanced information representation. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5):4120–4134, 2021. 2, 3

[47] Han Xu, Jiayi Ma, Zhenfeng Shao, Hao Zhang, Junjun Jiang, and Xiaojie Guo. Sdpnet: A deep network for pan-sharpening with enhanced information representation. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5):4120–4134, 2021. 3

[48] 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. 3

[49] 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. 5

[50] 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. 3

[51] 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. 5

[52] 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. 3

[53] 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. 5

[54] Hao Zhang and Jiayi Ma. Gtp-pnet: A residual learning network based on gradient transformation prior for pansharpening. *ISPRS Journal of Photogrammetry and Remote Sensing*, 172:223–239, 2021. 3

[55] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Xin Yu, Yiran Zhong, Nick Barnes, and Ling Shao. Rgb-d saliency detection via cascaded mutual information minimization. In *IEEE ICCV*, 2021. 3

[56] Zi-Yao Zhang, Ting-Zhu Huang, Liang-Jian Deng, Jie Huang, and Hong-Xia Dou. Pan-sharpening via rog-based filtering. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 2790–2793, 2019. 2

[57] Man Zhou, Xueyang Fu, Jie Huang, Feng Zhao, Aiping Liu, and Rujing Wang. Effective pan-sharpening with transformer and invertible neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022. 4

[58] Man Zhou, Keyu Yan, Jinshan Pan, Wenqi Ren, Qi Xie, and Xiangyong Cao. Memory-augmented deep unfolding network for guided image super-resolution, 2021. 3