Normalization-based Feature Selection and Restitution for Pan-sharpening

Man Zhou
University of Science and Technology of China
Hefei Institute of Physical Science, Chinese Academy of Sciences
manman@mail.ustc.edu.cn

Jie Huang*
University of Science and Technology of China
hj0117@mail.ustc.edu.cn

Keyu Yan
Hefei Institute of Physical Science, Chinese Academy of Sciences
University of Science and Technology of China
keyu@mail.ustc.edu.cn

Gang Yang
University of Science and Technology of China
yg1997@mail.ustc.edu.cn

Aiping Liu
University of Science and Technology of China
aipingl@ustc.edu.cn

Chongyi Li
Nanyang Technological University
lichongyi25@gmail.com

Feng Zhao†
University of Science and Technology of China
fzhao956@ustc.edu.cn

ABSTRACT

Pan-sharpening is essentially a panchromatic (PAN) image-guided low-spatial resolution MS image super-resolution problem. The commonly challenging issue of pan-sharpening is how to correctly select consistent features and propagate them, and properly handle inconsistent ones between PAN and MS modalities. To solve this issue, we propose a Normalization-based Feature Selection and Restitution mechanism, which is capable of filtering out the inconsistent features and promoting to learn the consistent ones. Specifically, we first modulate the PAN feature as the MS style in feature space by AdaIN operation [21]. However, such operation inevitably removes the favorable features. We thus propose to distill the effective information from the removed part and restitute it back to the modulated part. To better distillation, we enforce a contrastive learning constraint to close the distance between the restituted feature and the ground truth, and push the removed part away from the ground truth. In this way, the consistent features of PAN images are correctly selected and the inconsistent ones are filtered out, thus relieving the over-transferred artifacts in the process of PAN-guided MS super-resolution. Extensive experiments validate the effectiveness of the proposed network and demonstrate its favorable performance against other state-of-the-art methods. The source code will be released at https://github.com/manman1995/pansharpening.

*Both authors contributed equally to this research.
†Feng Zhao is the corresponding author.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547774

CCS CONCEPTS

• **Computing methodologies** → **Hyperspectral imaging**.

KEYWORDS

Normalization, contrastive learning, pan-sharpening

ACM Reference Format:
Man Zhou, Jie Huang, Keyu Yan, Gang Yang, Aiping Liu, Chongyi Li, and Feng Zhao. 2022. Normalization-based Feature Selection and Restitution for Pan-sharpening . In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3547774

1 INTRODUCTION

With the rapid development of satellite sensors, the satellite images have been used in a wide range of applications like military system, environmental monitoring, and mapping services. However, due to the technological and physical limitation of imaging devices, satellites are usually equipped with both multi-spectral (MS) and panchromatic (PAN) sensors to simultaneously measure the complementary images, MS images with low spatial resolution and high spectral resolution and PAN images with low spectral resolution and high spatial resolution. To obtain the images with both high spectral and high spatial resolutions, pan-sharpening technique that fuses the low resolution MS images and high spatial PAN images to break the technological limits for generating the expected high-resolution (HR) MS images, has drawn much attention from either image processing and remote sensing communities.

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 before feeding them into the networks. Without conducting explicitly cross-modal fusion, the "input fusion" strategy is therefore limited in studying the complementary

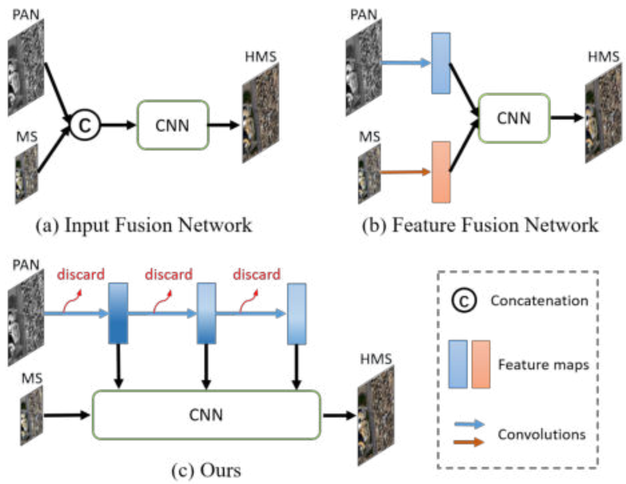

Figure 1: The categorization of existing Pan-sharpening methods.

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 captured in the same scene share the consistent information, they also have the modality-aware unique information. It is natural to transfer the effective part of the PAN modality-aware unique information to guide the MS modality super-resolution and reduce the wrong influence of the PAN guidance to predict the expected MS reconstruction properly. The key is how to correctly select the effective part of PAN modality and transfer it into MS modality. However, existing state-of-the-art Pan-sharpening methods don't explicitly enforce the consistent information learning and filtering out the inconsistent information between two modalities of PAN and MS images, resulting in the modality discrepancy and further the over-transformed artifacts. Considering the limitation of the current methods, in this paper, we make our efforts to enforce the consistent feature learning and reduce the modality discrepancy for improving the Pan-sharpening performance, as shown in Figure 1 (c).

To solve this issue, we propose a Normalization-based Feature Selection and Restitution mechanism, which is capable of filtering out the inconsistent features and promoting to learn the consistent ones. Specifically, we first modulate the PAN feature as the MS style in feature space by AdaIN operation [21]. However, such operation inevitably removes the favorable features. We thus propose distill the effective information from the removed part and restitute it back to the modulated part. To better distillation, we enforce a contrastive learning constraint to close the distance of the restituted feature and the ground truth, and push the removed part away from the ground truth. In this way, the consistent features of PAN images are correctly selected and the inconsistent ones are filtered out, thus relieving the over-transferred artifacts in the process of PAN-guided MS super-resolution. We conduct extensive experiments to analyze the effectiveness of the proposed network and demonstrate the favorable performance against state-of-the-art methods qualitatively and quantitatively while generalizing well to real-world scenes.

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

- To the best of our knowledge, this is the first attempt to introduce the normalization mechanism into pan-sharpening to explicitly address the modality discrepancy.
- The Normalization-based Feature Selection and Restitution mechanism is proposed to explicitly filter out the inconsistent features and promote to learn the consistent ones in PAN and MS modality.
- Extensive experiments over different satellite datasets demonstrate that our proposed method performs the best qualitative and quantitative while generalizing well to real-world full-resolution scenes.

2 RELATED WORK

2.1 Traditional pan-sharpening methods

Traditional pan-sharpening methods are classified into three types: Component Substitution (CS), Multi-resolution Analysis(MRA), and Variational Optimization (VO) [44, 45]. The most common methods of CS are intensity hue-saturation (IHS) fusion [11], the principal component analysis (PCA) methods [32, 43], Brovey transforms [16], and Gram-Schmidt (GS) orthogonalization method [34]. There are also some improvements based on the above methods proposed by researchers, such as the nonlinear IHS (NIHS) method [15] to reduce the spectrum distortion of IHS and the GSA method [1] with adaptive capability for the GS method. These CS methods are very fast to calculate, but the generated images are easy to contain artifacts. Compared with the CS methods, MRA methods bring less spectral distortion while sharpening MS images. Typical MRA methods include decimated wavelet transform (DWT) [39], high-pass filter fusion (HPF) [42], indusion method [31], Laplacian pyramid (LP) [47] and atrous wavelet transform (ATWT) [41]. P+XS pan-sharpening approach [3], the first variational method, assumes that PAN image is derived from the linear combination of various bands of HRMS, whereas the upsampled low resolution multi-spectral (LRMS) image is from the blurred HRMS image. Subsequently, various constraints are introduced into pan-sharpening task, such as dynamic gradient sparsity property (SIRF) [12], local gradient constraint (LGC) [13], group low-rank constraint for texture similarity (ADMM) [45] and so on. These various priors and constraints requiring the manual setting of parameters can only inadequately reflect the limited structural relations of the images, which can also result in degradation.

2.2 CNN-based pan-sharpening methods

Owing to the rapid development of convolutional neural networks (CNN) in computer vision, CNN that has powerful learning capabilities has been widely used in hyperspectral images [10, 14, 18, 24–28, 49] and remote sensing images [5, 7–9, 23, 29, 30, 37, 54, 55, 60–64]. Recently, Various CNN-based methods [38, 52, 59] have been put forward to promote the fusion quality of pan-sharpening. For example, Masi *et al.* [40] are the first to use CNN to deal with the issue of pan-sharpening. Although the structure is simple, the effect

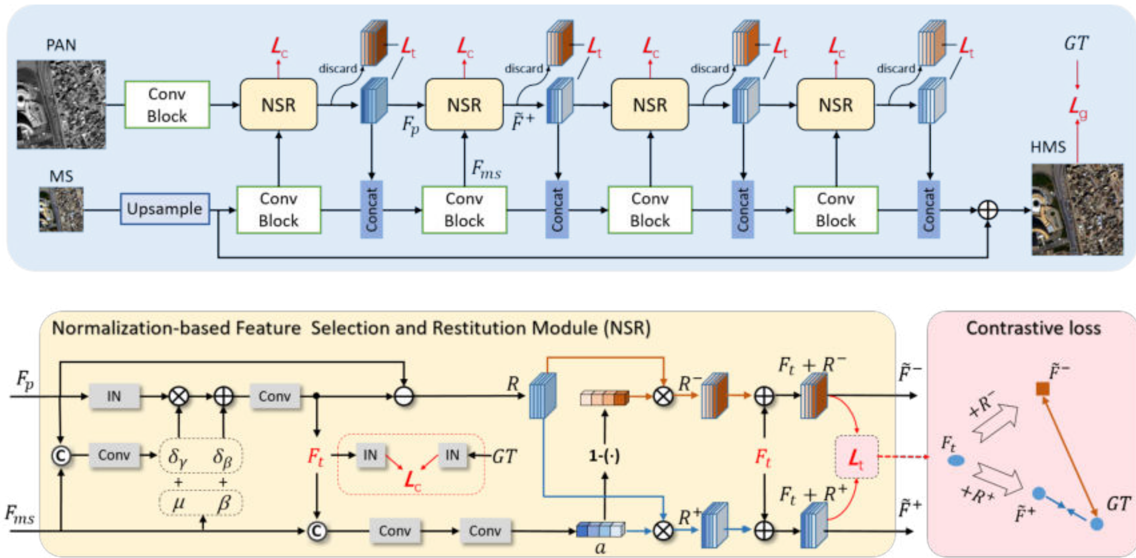

Figure 2: The pipeline of our proposed pan-sharpening framework and the core Normalization-based feature selection and restitution module. It is capable of correctly selecting consistent features and propagating them and properly filtering out inconsistent ones between PAN and MS modalities.

is much better than the traditional methods. Then, Yang *et al.* [56] designed a deeper convolutional network by relying on resblock in [20]. Meanwhile, Yuan *et al.* [57] introduced multi-scale module into the basic CNN architecture. Later, Cai *et al.* [4] and Wu *et al.* [50] have the similar idea, that is, continuously introduce images of different scales into the backbone network. The difference between the two approaches is that one uses PAN images and the other uses MS images. Recently, some model-driven CNN models with clear physical meaning emerged. The basic idea is to use prior knowledge to formulate optimization problems for computer vision tasks, then unfold the optimization algorithms into deep neural networks. For example, Xu *et al.* [53] developed two separate priors of PAN and MS to design the unfolding structure for pan-sharpening. The model-driven methods have interpretability and clear physical meaning. Cao *et al.* [6] unfolded an alternate optimization algorithm into CNN. Tian *et al.* [46] and Wu *et al.* [51] combined variational optimization and deep residual CNN.

3 METHODS

In this section, we will first present the overall flowchart of the proposed pan-sharpening framework, illustrated in Figure 2. We further provide the detail of our devised Normalization-based feature selection and restitution module. Finally, we deepen into the newly-designed loss function.

3.1 Framework

Targeting at pan-sharpening, it aims to super-resolve the low-resolution MS images, conditioning on the paired high-resolution PAN images. Since PAN and MS images captured in the same scene share the consistent information, they also have the modality-aware unique information. It is natural to transfer the effective part of the PAN modality-aware unique information to guide the MS modality super-resolution. The key is how to correctly select the effective part of PAN modality and transfer it into MS modality.

To this end, we first attempt to address this issue from the normalization perspective and devise a Normalization-based feature selection and restitution module, which is capable of filtering out the inconsistent features and promoting to learn the consistent ones. Equipped with the above module, our proposed method is constructed, thus relieving the over-transferred artifacts in the process of PAN-inserted MS super-resolution.

Figure 2 shows the overall flowchart of our framework. Remarkably, given PAN image $P \in R^{H \times W \times 1}$ and MS image $L \in R^{H/r \times W/r \times C}$, the network first applies the convolution layer to project the r-times L by Bibubic upsampling into shallow feature representations while P is fed into the convolution block to extract the informative features. Next, the obtained modality-aware feature maps of MS and PAN are jointly passed through K numbers of the core Normalization-based feature selection and restitution module, yielding the effective feature representation of the PAN modality. In each core module, the PAN feature is normalized and then integrated with the MS feature. Finally, we apply a convolution layer to transform the corrected feature of the final core module back to image space and then combine it with the Bibubic up-sampled input L as the output image.

3.2 Normalization-based feature selection and restitution module

As shown in Figure 2, normalization-based feature selection and restitution module consists of three phases: 1) consistent modality modulation phase, 2) feature selection phase and 3) feature restitution phase. To be specific, the first is responsible for modulating the input PAN features as the style of MS features by AdaIN operation, thus relieving the modality discrepancy. Then, the second employs the attention mechanism to select the effective part from the discarded features by the first stage while the third aims to restitute it as a compensation back to the normalized features by AdaIN, thus promoting to learn the consistent ones and further improving the feature representation.

Consistent modality modulation phase. As well recognized, since PAN and MS images captured in the same scene share the consistent information, they also have the modality-aware inconsistent information. Most of the existing pan-sharpening methods simply integrate the PAN and MS features together and then perform the next convolution operation, which is prone to result in the modality-aware discrepancy. To address this problem, inspired by style transformation [21], we employ the AdaIN operation to modulate the PAN features as the style of the MS modality features, thus enhancing the consistency of the matched PAN features with the MS feature distribution.

Taking a module for example, we denote the input MS feature and PAN feature by $F_{ms} \in \mathbb{R}^{h\times w\times c}$ and $F_p \in \mathbb{R}^{h\times w\times c}$ respectively, and the output by $\widetilde{F}^{+} \in \mathbb{R}^{h\times w\times c}$, where h, w, c denote the height, width, and number of channels, respectively. The PAN features are considered as guidance information to complement the MS features. To this end, we implement the modulation over the input PAN feature F_p. Specifically, we first try to reduce the modality discrepancy by performing Adaptive Instance Normalization as

$$F_t = \mathrm{AdaIN}(F_p) = \gamma(\frac{F_p - \mu(F_p)}{\sigma(F_p)}) + \beta, \tag{1}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and standard deviation computed across spatial dimensions independently for each channel and each *sample/instance* as

$$\begin{aligned} \mu_c(F_p) &= \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}(F_p)_{chw}, \\ \sigma_c(F_p) &= \sqrt{\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}((F_p)_{chw} - \mu_c(F_p))^2 + \epsilon}, \end{aligned} \tag{2}$$

where ϵ is the very small number in order to prevent the division denominator from being 0.

In terms of γ and β, as shown in Figure 2, we obtain them by the following two steps: 1) the input PAN feature F_p and MS feature F_{ms} are firstly concatenated and fed into the convolution layer C_1 to transform the channel of the concatenated feature back to the same as that of F_p

$$F_{pm} = C_1(Cat[F_p, F_{ms}]). \tag{3}$$

Then, the above feature F_{pm} is passed through two independent branches convolution layers C_3 and C_3 with 3×3 kernel to get two parameters $\delta\gamma$ and $\delta\beta$ as

$$\begin{aligned} \delta\beta &= C_3(F_{pm}), \\ \delta\gamma &= C_3(F_{pm}). \end{aligned} \tag{4}$$

2) we figure out the mean and standard deviation computed across spatial dimensions independently for each channel of the input MS feature F_{ms} as

$$\begin{aligned} \mu_c(F_{ms}) &= \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}(F_{ms})_{chw}, \\ \sigma_c(F_{ms}) &= \sqrt{\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}((F_{ms})_{chw} - \mu_c(F_{ms}))^2 + \epsilon}, \end{aligned} \tag{5}$$

Followed by above calculation, we integrate them to obtain the γ and β as

$$\begin{aligned} \beta &= \mu_c(F_{ms}) + \delta\beta, \\ \gamma &= \sigma_c(F_{ms}) + \delta\gamma. \end{aligned} \tag{6}$$

In this modulation way, the modality discrepancy will be relieved.

Feature selection phase. As well recognized, normalization operation will inevitably discard some useful information of PAN features F_p by AdaIN. Targeting at above operation, it can be expressed as

$$R = F_p - F_t, \tag{7}$$

where R denotes the difference between the original input feature F_p and the normalized feature F_t. Regrading the information loss, we need to perform the feature selection over the discarded part R to distinguish the useful part. We propose to distill the useful part through masking the discarded R with the learned channel attention vector $\mathbf{a} = [a_1, a_2, \cdots, a_c]$ where the dimension c is the same as the F_p. Given the attention a, the selected useful part and the harmful part can be remarked as

$$\begin{aligned} R^{+}(:,:,k) &= a_k R(:,:,k), \\ R^{-}(:,:,k) &= (1 - a_k)R(:,:,k), \end{aligned} \tag{8}$$

where $R(:,:,k) \in \mathbb{R}^{h\times w}$ denotes the k^{th} channel of feature map R, $k = 1, 2, \cdots, c$. To implement the channel attention, we employ the SE-like attention network to produce the channel attention vector $\mathbf{a}$: 1) we first concatenate the modulated F_t and the input F_{ms} and then pass them through several convolutions to halve the channel dimension, 2) the channel-halved feature is pooled to the vector by global average pooling and then predict the attention vector a as

$$\mathbf{a} = sigmoid(C_1(GAP(C_3(Cat[F_t, F_{ms}])))), \tag{9}$$

where GAP indicates the global average pooling layer and sigmoid is the sigmoid activation function. Cat, C_1 and C_3 represent the concatenation operation by channel dimension, the convolution block with 1×1 kernel size and the convolution block with 3×3 kernel size respectively.

Feature restitution phase. After selecting out the useful part feature R^{+}, we can obtain the output feature $\widetilde{F}^{+}$ of the Normalization-based module by resistuting it to the style normalized feature F_t as

$$\widetilde{F}^{+} = F_t + R^{+}. \tag{10}$$

3.3 Contrastive learning strategy.

In order to facilitate the feature distillation, we enforce a contrastive learning constraint to close the distance between the restituted feature and the ground truth, and push the removed part away from the ground truth. In this way, the consistent features of PAN images are correctly selected and the inconsistent ones are filtered out, thus relieving the over-transferred artifacts in the process of

PAN-guided MS super-resolution. Given the restituted feature $\widetilde{F}^{+}$ with the useful part, the feature $\widetilde{F}^{-} = F_t + R^{-}$ with the selected harmful part R^{-} and the feature of ground truth F_H, the contrastive learning strategy can be written as

$$L_t = \frac{||\mathrm{Pool}(\widetilde{F}^{+}), \mathrm{Pool}(F_H)||_1}{||\mathrm{Pool}(\widetilde{F}^{-}), \mathrm{Pool}(F_H)||_1}, \quad (11)$$

where Pool(.) denotes the average pooling operation to avoid the distraction caused by spatial misalignment. In addition, to ensure the generated F_t being the consistent part of MS modality, we enforce a supervision loss between the output of F_t and the ground truth being passed through instance normalization (IN) layer as

$$L_c = ||\mathrm{IN}(F_t), \mathrm{IN}(F_H)||_1, \quad (12)$$

where F_H denotes the output feature of the ground truth being passed through the convolution block as PAN image.

3.4 Joint Training

As shown in Figure 2, we train the entire network in an end-to-end manner and the overall loss function consists of two parts: one for reconstructing the ground-truth MS image $L_g = \|f(L, P) - gt\|_1$ by $L1$ loss where $f(.)$ denotes the mapping function of our method, and the other for better distilling the consistent part and the inconsistent part between two modalities in the Normalization-based feature selection and restitution module, written as:

$$L = L_g + \lambda \sum_{b=1}^{K} L_t^b + L_c^b, \quad (13)$$

where L_t^b indicates the proposed Contrastive learning strategy for the b^{th} Normalization-based feature selection and restitution module (NSR) and K is the number of NSR modules. H is the ground truth MS image, and λ is the parameters to balance the two terms in the loss function. In our setting, λ is set as 0.1.

4 EXPERIMENTS AND RESULTS

4.1 Baseline methods

To show our proposed technique's efficacy, we compare it to the performance of several representative pan-sharpening algorithms: 1) five state-of-the-art deep-learning based methods, including PNN [40], PANNET [56], MSDCNN [58], SRPPNN [4], GPPNN [53] and BAM [65]; 2) five promising traditional methods, namely SFIM [36], Brovey [17], GS [33], IHS [19], and GFPCA [35].

4.2 Datasets and benchmark

Reduced resolution scene. 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 [48]. Specifically, given the MS image $H \in R^{M \times N \times C}$ and the PAN image $\widetilde{P} \in R^{rM \times rN \times b}$, 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.

Full resolution scenes. We construct an additional full-resolution real-world dataset of 200 samples over the newly selected GaoFen2 satellite in order to conduct the model generalization comparison. To be more specific, the additional dataset is generated using the full-resolution mode, which creates PAN and MS images in the manner described above without down-sampling, with PAN images having a resolution of 32×32 and MS images having a resolution of 128×128.

4.3 Implementation details and metrics

All our networks are built in PyTorch on NVIDIA GeForce GTX 2080Ti GPU on a PC. During the training phase, Adam tunes them throughout 1000 epochs with a batch size of four. The initial learning rate is set at 8×10^{-4}. The learning rate is decayed by multiplying by 0.5 every 200 epochs. For reduced-resolution scene, several widely-used image quality assessment (IQA) metrics are adapted for performance measurement, including the PSNR, SSIM, SAM [22], ERGAS [2]. In addition, because there are no ground-truth MS images available for real-world full-resolution scenes, we utilize three widely-used no-reference IQA metrics to assess the model's performance: the spectral distortion index D_λ, the spatial distortion index D_S, the quality without reference (QNR).

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
BAM	41.3527	0.9671	0.0239	0.9932
Ours	**41.7113**	**0.9705**	**0.0223**	**0.9513**

4.4 Comparison with state-of-the-art methods

Evaluation on reduced-resolution scene. A summary of the assessment measures for three datasets is shown in Table 1, Table 2 and Table 4, where the values highlighted in red reflect the best

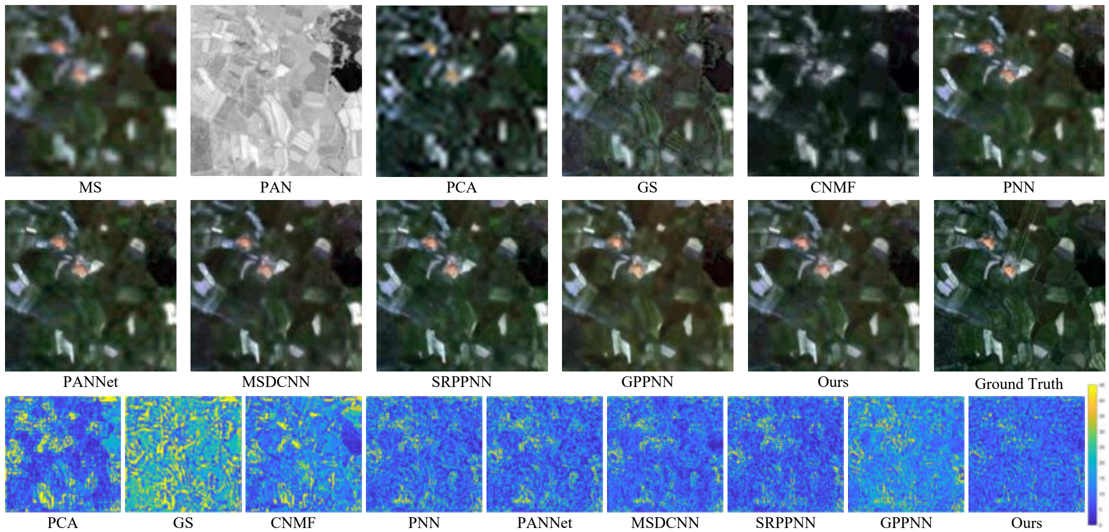

Figure 3: The visual comparisons between other pan-sharpening methods and our method on WorldView-II satellite.

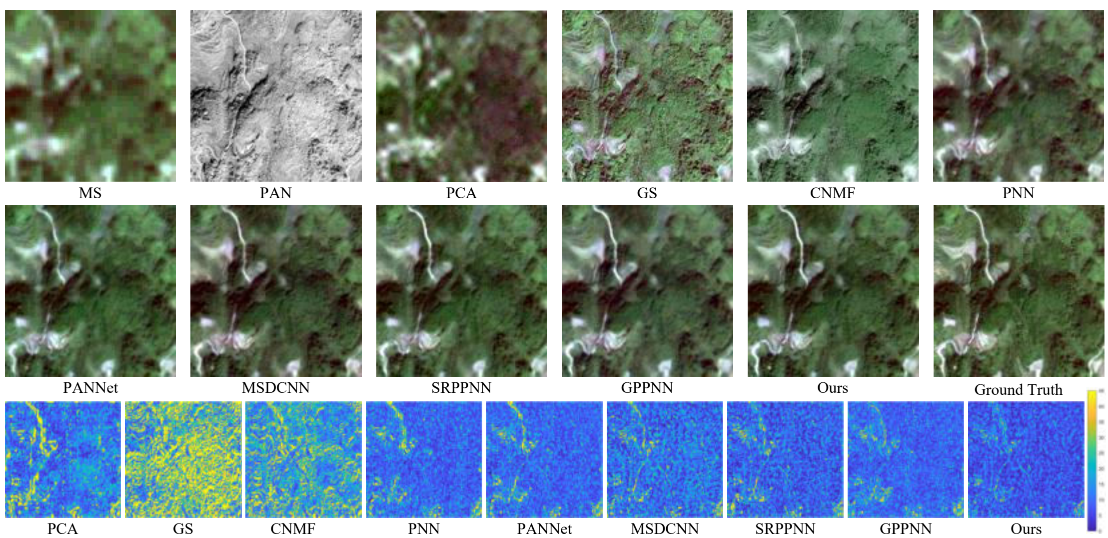

Figure 4: The visual comparisons between other pan-sharpening methods and our method on GaoFen2 satellite.

values. On three satellite datasets, it is clearly shown that our technique outperforms all existing comparing algorithms in terms of all assessment metrics. With regard to the WorldView-II, GaoFen2 and WorldView-III datasets in particular, our strategy yields 0.26 dB, 0.25 dB and 0.10 dB improvements in PSNR compared to the second-best results obtained by using other methods. Other measurements, such as the PSNR, have shown comparable improvements to the PSNR over the last year. In comparison to existing deep learning-based approaches, we produce much superior outcomes, hence demonstrating the usefulness of our suggested strategy.

In addition, we exhibit the comparison of the visual results to testify the efficacy of our approach in Figure 3 and Figure 4 on representative samples of the WorldView-II and GaoFen2 datasets, respectively, in order to demonstrate the efficiency of our method. The MSE residual between the pan-sharpened findings and the ground truth is shown in the final row of the images. The spatial and spectral aberrations in our model are minimal in comparison to those of other competing techniques. It is simple to draw this conclusion based on the observation of MSE maps. Regarding the MSE residues, it has been observed that our suggested technique is more accurate than other comparison methods when compared to the ground truth. In this way, it can be concluded that our technique outperforms all existing competing pan-sharpening algorithms in terms of performance. In particular, we note that our suggested technique has finer-grained textures and coarser-grained structures when compared to previous methods, which is based on the amplified local areas we examined. For this reason, the closer the absolute

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
BAM	45.7419	0.9836	0.0134	0.6267
Ours	**47.3416**	**0.9893**	**0.0102**	**0.5476**

Table 3: Comparisons of FLOPs (G) and parameters number (M). "Param" denotes parameters number.

	PNN	PANNET	MSDCNN	SRPPNN	GPPNN	Ours
Param	0.0689	0.0688	0.2390	1.7114	0.1198	0.1229
FLOPs	1.1289	1.1275	3.9158	21.1059	1.3967	1.5375

error map is to a GT image, the more accurate the pan-sharpened result is.

Evaluation on full-resolution scene A pre-trained model built on GaoFen2 data is applied to some previously unseen full-resolution GaoFen2 satellite datasets in order to assess the performance of our network at full resolution and the generalization capabilities of the model. A quantitative comparison between representative CNN-based techniques and our solution is presented in the following Table 5. The lower D_λ, D_s and the higher QNR correspond to the better image quality. As demonstrated in Table 5, our proposed strategy outperforms existing conventional and deep learning-based methods on practically all indices, demonstrating that our method has better generalization ability than other methods.

4.5 Parameter numbers vs model performance

A more in-depth examination of the approaches is carried out by investigating their computational complexity, which is represented in Table 3 by the number of floating-point operations (FLOPs) and the number of parameters (in 10 M). Compared to other deep learning-based approaches, it can be observed that our network is able to create a decent trade-off and gets the greatest performance while using much fewer parameters and storage. 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.

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
BAM	30.3845	0.9188	0.0773	3.1679
Ours	**30.5355**	**0.9225**	**0.0747**	**3.1123**

4.6 Ablation experiments

To investigate the contribution of the devised components in our proposed network, we have conducted comprehensive ablation studies on the WorldView-II satellite dataset of the Pan-sharpening task. To be specific, the Normalization-based feature selection and restitution module and the contrastive learning loss in the optimization function are the two core designs. All the experimental results are measured by the widely-used IQA metrics, i.e., ERGAS [2], PSNR, SSIM, and SAM.

The Normalization module. To explore the positive impact of the proposed Normalization-based feature selection and restitution module, we experiment it by observing the network performance change through adding and removing it from the proposed method. The corresponding quantitative comparison is reported in Table 6. Observing the results from the first row of Table 6, it can be clearly figured out that the model performance has obtained considerable degradation when replacing the module from the network with the widely-used ResNet block to maintain the parameter consistence. It is because deleting it will result in the wrong influence of the PAN guidance over-transferring into the MS super-resolution process, thus leading to the modal discrepancy and further degrading the pan-sharpening results.

The contrastive learning loss. The newly-designed contrastive learning loss aims to better distill the useful information and the harmful part from the discarded part by IN. In the second experiment of Table 6, we delete it to examine its effectiveness. The results in Table 6 demonstrate that removing it will degrade all metrics dramatically, indicating its significant role in our network. This is

Table 5: Evaluation on the real-world full-resolution scenes from GaoFen2 dataset. The best results are highlighted in bold.

Metrics	SFIM	GS	Brovey	IHS	GFPCA	PNN	PANNET	MSDCNN	SRPPNN	GPPNN	BAM	**Ours**
$D_\lambda\downarrow$	0.0822	0.0696	0.1378	0.0770	0.0914	0.0746	0.0737	0.0734	0.0767	0.0782	0.0755	**0.0672**
$D_s\downarrow$	**0.1087**	0.2456	0.2605	0.2985	0.1635	0.1164	0.1224	0.1151	0.1162	0.1253	0.1159	0.1115
QNR↑	0.8214	0.7025	0.6390	0.6485	0.7615	0.8191	0.8143	0.8251	0.8173	0.8073	0.8211	**0.8288**

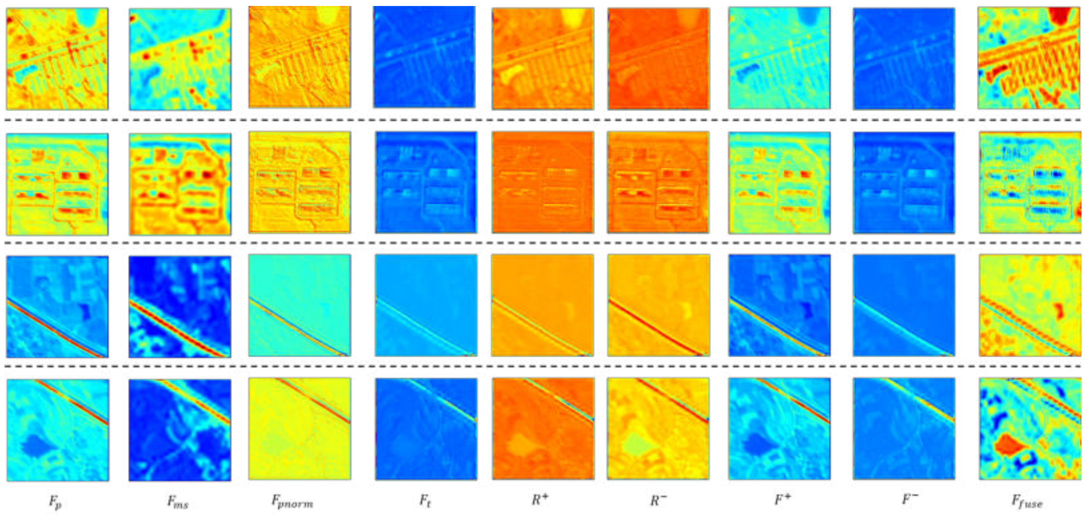

Figure 5: The Visualization of the immediate output feature maps.

Table 6: The results of ablation experiments of Normalization module "NSR" and the contrastive loss function "CL" over WorldView-II datasets. The best values are highlighted by the red bold.

Config	NSR	CL	PSNR↑	SSIM↑	SAM↓	ERGAS↓
(I)	×	✓	41.3781	0.9665	0.0248	1.0190
(II)	✓	×	41.6884	0.9689	0.0226	0.9521
(III)	✓	✓	41.7113	0.9705	0.0223	0.9513

due to its powerful ability to enable the module to filter out the inconsistent information and promote to learn the consistent ones.

In the last row of Table 6, we can clearly find that compared with the above variants, the best results can be obtained by combining all the above components. It further supports the claims above.

4.7 Visualization of the immediate output

To better understand how a Normalization-based feature selection and restitution module works, we visualize the intermediate feature maps of the first module of our pipeline in Figure 5. To be specific, we get each activation maps by averaging the feature maps along channels. As illustrated above, it shows the activation maps of input F_p, F_{ms}, the normalized feature F_{pnorm}, the modulated feature F_t and the effective part R^+, the discarded part R^- as well as the restituted feature $\widetilde{F}^+$, the fused feature F_{fuse} of $\widetilde{F}^+$ and F_{ms}, respectively. We see that after adding the consistent feature R^+, the contaminated feature $\widetilde{F}^+$ has the more powerful and informative capability while the discarded part R is the inconsistent part that contains the over-transferred PAN-modality unique information. It demonstrates the powerful capability of the core module.

5 CONCLUSION

In this paper, we propose a Normalization-based Feature Selection and Restitution mechanism, which is capable of filtering out the inconsistent features and promoting to learn the consistent ones. To the best of our knowledge, this is the first attempt to introduce the normalization mechanism into pan-sharpening to explicitly address the modality discrepancy. Extensive experiments validate the effectiveness of the proposed network and the favorable generalization ability to real-world full-resolution scenes against other state-of-the-art methods.

In the future, we will investigate the feasibility of incorporating our proposed NSR into other existing pan-sharpening algorithms to enhance their performance.

ACKNOWLEDGMENTS

This work was supported by 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] Bruno Aiazzi, Stefano Baronti, and Massimo Selva. 2007. Improving component substitution pansharpening through multivariate regression of MS + Pan data. *IEEE Transactions on Geoscience and Remote Sensing* 45, 10 (2007), 3230–3239.

[2] L. Alparone, L. Wald, J. Chanussot, C. Thomas, P. Gamba, and L. M. Bruce. 2007. Comparison of Pansharpening Algorithms: Outcome of the 2006 GRS-S Data Fusion Contest. *IEEE Transactions on Geoscience and Remote Sensing* 45, 10 (2007), 3012–3021.

[3] Coloma Ballester, Vicent Caselles, Laura Igual, Joan Verdera, and Bernard Rougé. 2006. A variational model for P+ XS image fusion. *International Journal of Computer Vision* 69, 1 (2006), 43–58.

[4] Jiajun Cai and Bo Huang. 2021. Super-Resolution-Guided Progressive Pansharpening Based on a Deep Convolutional Neural Network. *IEEE Transactions on Geoscience and Remote Sensing* 59, 6 (2021), 5206–5220.

[5] Xiangyong Cao, Yang Chen, Qian Zhao, Deyu Meng, Yao Wang, Dong Wang, and Zongben Xu. 2015. Low-Rank Matrix Factorization under General Mixture Noise Distributions. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 1493–1501.

[6] Xiangyong Cao, Xueyang Fu, Danfeng Hong, Zongben Xu, and Deyu Meng. 2021. PanCSC-Net: A Model-Driven Deep Unfolding Method for Pansharpening. *IEEE Transactions on Geoscience and Remote Sensing* (2021), 1–13.

[7] Xiangyong Cao, Lin Xu, Deyu Meng, Qian Zhao, and Zongben Xu. 2017. Integration of 3-dimensional discrete wavelet transform and Markov random field for hyperspectral image classification. *Neurocomputing* 226 (2017), 90–100.

[8] Xiangyong Cao, Zongben Xu, and Deyu Meng. 2019. Spectral-Spatial Hyperspectral Image Classification via Robust Low-Rank Feature Extraction and Markov Random Field. *Remote. Sens.* 11, 13 (2019), 1565.

[9] Xiangyong Cao, Jing Yao, Zongben Xu, and Deyu Meng. 2020. Hyperspectral Image Classification With Convolutional Neural Network and Active Learning. *IEEE Transactions on Geoscience and Remote Sensing* 58, 7 (2020), 4604–4616.

[10] Xiangyong Cao, Feng Zhou, Lin Xu, Deyu Meng, Zongben Xu, and John Paisley. 2018. Hyperspectral Image Classification With Markov Random Fields and a Convolutional Neural Network. *IEEE Transactions on Image Processing* 27, 5 (2018), 2354–2367.

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

[12] Chen Chen, Yeqing Li, Wei Liu, and Junzhou Huang. 2015. SIRF: Simultaneous Satellite Image Registration and Fusion in a Unified Framework. *IEEE Transactions on Image Processing* 24, 11 (2015), 4213–4224.

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

[14] Ying Fu, Zhiyuan Liang, and Shaodi You. 2021. Bidirectional 3D Quasi-Recurrent Neural Network for Hyperspectral Image Super-Resolution. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14 (2021), 2674–2688.

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

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

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

[18] Juan Mario Haut, Mercedes E. Paoletti, Javier Plaza, Jun Li, and Antonio Plaza. 2018. Active Learning With Convolutional Neural Networks for Hyperspectral Image Classification Using a New Bayesian Approach. *IEEE Transactions on Geoscience and Remote Sensing* 56, 11 (2018), 6440–6461.

[19] R. Haydn, G. W. Dalke, J. Henkel, and J. E. Bare. 1982. 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 (1982), 571–577.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[21] Xun Huang and Serge Belongie. 2017. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 1510–1519. https://doi.org/10.1109/ICCV.2017.167

[22] A. F. Goetz J. R. H. Yuhas and J. M. Boardman. 1992. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm. *Proc. Summaries Annu. JPL Airborne Geosci. Workshop* (1992), 147–149.

[23] Junjun Jiang, Jiayi Ma, Chen Chen, Zhongyuan Wang, Zhihua Cai, and Lizhe Wang. 2018. SuperPCA: A Superpixelwise PCA Approach for Unsupervised Feature Extraction of Hyperspectral Imagery. *IEEE Transactions on Geoscience and Remote Sensing* 56, 8 (2018), 4581–4593.

[24] Junjun Jiang, Jiayi Ma, and Xianming Liu. 2020. Multilayer Spectral-Spatial Graphs for Label Noisy Robust Hyperspectral Image Classification. *IEEE Transactions on Neural Networks and Learning Systems* (2020), 1–14.

[25] Junjun Jiang, Jiayi Ma, Zheng Wang, Chen Chen, and Xianming Liu. 2019. Hyperspectral Image Classification in the Presence of Noisy Labels. *IEEE Transactions on Geoscience and Remote Sensing* 57, 2 (2019), 851–865.

[26] Junjun Jiang, He Sun, Xianming Liu, and Jiayi Ma. 2020. Learning Spatial-Spectral Prior for Super-Resolution of Hyperspectral Imagery. *IEEE Transactions on Computational Imaging* 6 (2020), 1082–1096.

[27] Kui Jiang, Zhongyuan Wang, Peng Yi, and Junjun Jiang. 2018. A Progressively Enhanced Network for Video Satellite Imagery Superresolution. *IEEE Signal Processing Letters* 25, 11 (2018), 1630–1634.

[28] Kui Jiang, Zhongyuan Wang, Peng Yi, Junjun Jiang, Guangcheng Wang, Zhen Han, and Tao Lu. 2019. GAN-Based Multi-level Mapping Network for Satellite Imagery Super-Resolution. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*. 526–531.

[29] Kui Jiang, Zhongyuan Wang, Peng Yi, Junjun Jiang, Emily Xiao, and Yuan Yao. 2018. Deep Distillation Recursive Network for Remote Sensing Imagery Super-Resolution. *Remote Sensing* 10 (10 2018), 1700.

[30] Kui Jiang, Zhongyuan Wang, Peng Yi, Guangcheng Wang, Tao Lu, and Junjun Jiang. 2019. Edge-Enhanced GAN for Remote Sensing Image Superresolution. *IEEE Transactions on Geoscience and Remote Sensing* 57, 8 (2019), 5799–5812.

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

[32] P Kwarteng and A Chavez. 1989. Extracting spectral contrast in Landsat Thematic Mapper image data using selective principal component analysis. *Photogrammetric Engineering and remote sensing* 55, 339-348 (1989), 1.

[33] C.A. Laben and B.V. Brower. 2000. Process for Enhancing the Spatial Resolution of Multispectral Imagery Using Pan-Sharpening. *US Patent 6011875A* (2000).

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

[35] W. Liao, H. Xin, F. V. Coillie, G. Thoonen, and W. Philips. 2017. 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*.

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

[37] Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping Zhang. 2020. DDcGAN: A Dual-Discriminator Conditional Generative Adversarial Network for Multi-Resolution Image Fusion. *IEEE Transactions on Image Processing* 29 (2020), 4980–4995.

[38] Jiayi Ma, Wei Yu, Chen Chen, Pengwei Liang, Xiaojie Guo, and Junjun Jiang. 2020. Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion. *Information Fusion* 62 (2020), 110–120.

[39] SG Mallat. 1989. A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11, 7 (1989), 674–693.

[40] Giuseppe Masi, Davide Cozzolino, Luisa Verdoliva, and Giuseppe Scarpa. 2016. Pansharpening by Convolutional Neural Networks. *Remote Sensing* 8, 7 (2016).

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

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

[43] Vijay P. Shah, Nicolas H. Younan, and Roger L. King. 2008. An Efficient Pan-Sharpening Method via a Combined Adaptive PCA Approach and Contourlets. *IEEE Transactions on Geoscience and Remote Sensing* 46, 5 (2008), 1323–1335.

[44] Xin Tian, Yuerong Chen, Changcai Yang, Xun Gao, and Jiayi Ma. 2020. A Variational Pansharpening Method Based on Gradient Sparse Representation. *IEEE Signal Processing Letters* 27 (2020), 1180–1184.

[45] Xin Tian, Yuerong Chen, Changcai Yang, and Jiayi Ma. 2021. Variational Pansharpening by Exploiting Cartoon-Texture Similarities. *IEEE Transactions on Geoscience and Remote Sensing* (2021), 1–16.

[46] Xin Tian, Kun Li, Zhongyuan Wang, and Jiayi Ma. 2021. VP-Net: An Interpretable Deep Network for Variational Pansharpening. *IEEE Transactions on Geoscience and Remote Sensing* (2021), 1–16.

[47] Gemine Vivone, Luciano Alparone, Jocelyn Chanussot, Mauro Dalla Mura, Andrea Garzelli, Giorgio A Licciardi, Rocco Restaino, and Lucien Wald. 2014. A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing* 53, 5 (2014), 2565–2586.

[48] Lucien Wald, Thierry Ranchin, and Marc Mangolini. 1997. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogrammetric Engineering and Remote Sensing* 63 (11 1997), 691–699.

[49] Xinya Wang, Jiayi Ma, and Junjun Jiang. 2021. Hyperspectral Image Super-Resolution via Recurrent Feedback Embedding and Spatial-Spectral Consistency Regularization. *IEEE Transactions on Geoscience and Remote Sensing* (2021), 1–13.

[50] Xiao Wu, Ting-Zhu Huang, Liang-Jian Deng, and Tian-Jing Zhang. 2021. Dynamic Cross Feature Fusion for Remote Sensing Pansharpening. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 14687–14696.

[51] Zhong-Cheng Wu, Ting-Zhu Huang, Liang-Jian Deng, Jin-Fan Hu, and Gemine Vivone. 2021. VO+Net: An Adaptive Approach Using Variational Optimization and Deep Learning for Panchromatic Sharpening. *IEEE Transactions on Geoscience and Remote Sensing* (2021), 1–16.

[52] Han Xu, Jiayi Ma, Zhenfeng Shao, Hao Zhang, Junjun Jiang, and Xiaojie Guo. 2021. SDPNet: A Deep Network for Pan-Sharpening With Enhanced Information Representation. *IEEE Transactions on Geoscience and Remote Sensing* 59, 5 (2021), 4120–4134.

[53] Shuang Xu, Jiangshe Zhang, Zixiang Zhao, Kai Sun, Junmin Liu, and Chunxia Zhang. 2021. Deep Gradient Projection Networks for Pan-sharpening. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1366–1375.

[54] Keyu Yan, Man Zhou, Liu Liu, Chengjun Xie, and Danfeng Hong. 2022. When Pansharpening Meets Graph Convolution Network and Knowledge Distillation. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–15. https://doi.org/10.1109/TGRS.2022.3168192

[55] Gang Yang, Man Zhou, Keyu Yan, Aiping Liu, Xueyang Fu, and Fan Wang. 2022. Memory-Augmented Deep Conditional Unfolding Network for Pan-Sharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1788–1797.

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

[57] Qiangqiang Yuan, Yancong Wei, Xiangchao Meng, Huanfeng Shen, and Liangpei Zhang. 2018. 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 (2018), 978–989.

[58] Q. Yuan, Y. Wei, X. Meng, H. Shen, and L. Zhang. 2018. 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 (2018), 978–989.

[59] Hao Zhang and Jiayi Ma. 2021. GTP-PNet: A residual learning network based on gradient transformation prior for pansharpening. *ISPRS Journal of Photogrammetry and Remote Sensing* 172 (2021), 223–239.

[60] Man Zhou, Xueyang Fu, Jie Huang, Feng Zhao, Aiping Liu, and Rujing Wang. 2022. Effective Pan-Sharpening With Transformer and Invertible Neural Network. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–15. https://doi.org/10.1109/TGRS.2021.3137967

[61] Man Zhou, Jie Huang, Yanchi Fang, Xueyang Fu, and Aiping Liu. 2022. Pan-Sharpening with Customized Transformer and Invertible Neural Network. AAAI Press.

[62] Man Zhou, Zeyu Xiao, Xueyang Fu, Aiping Liu, Gang Yang, and Zhiwei Xiong. 2021. Unfolding Taylor's Approximations for Image Restoration. In *NeurIPS*.

[63] Man Zhou, Keyu Yan, Jie Huang, Zihe Yang, Xueyang Fu, and Feng Zhao. 2022. Mutual Information-Driven Pan-Sharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1798–1808.

[64] Man Zhou, Keyu Yan, Jinshan Pan, Wenqi Ren, Qiaokang Xie, and Xiangyong Cao. 2022. Memory-augmented Deep Unfolding Network for Guided Image Super-resolution. *ArXiv* abs/2203.04960 (2022).

[65] Tian-Jing Zhang Xiaoxu Jin Zi-Rong Jin, Liang-Jian Deng. 2021. BAM: Bilateral Activation Mechanism for Image Fusion. *Proceedings of the 29th ACM International Conference on Multimedia (ACM MM)* (2021), DOI: 10.1145/3474085.3475571.