Multi-scale Spatial-Spectral Attention Guided Fusion Network for Pansharpening

Yong Yang*
School of Computer Science and Technology, Tiangong University
Tianjin, China
greatyangy@126.com

Mengzhen Li*
School of Information Technology, Jiangxi University of Finance and Economics
Nanchang, China
me_limz@163.com

Shuying Huang†
School of Software, Tiangong University
Tianjin, China
shuyinghuang2010@126.com

Hangyuan Lu
College of Information Engineering, Jinhua Polytechnic
Jinhua, China
lhyhziee@163.com

Wei Tu
School of Big Data Science, Jiangxi Science and Technology Normal University
Nanchang, China
283299985@qq.com

Weiguo Wan
School of Software and Internet of Things Engineering, Jiangxi University of Finance and Economics
Nanchang, China
wanweiguo@jxufe.edu.cn

ABSTRACT

Pansharpening is to fuse high-resolution panchromatic (PAN) images with low-resolution multispectral (LR-MS) images to generate high-resolution multispectral (HR-MS) images. Most of the deep learning-based pansharpening methods did not consider the inconsistency of the PAN and LR-MS images and used simple concatenation to fuse the source images, which may cause spectral and spatial distortion in the fused results. To address this problem, a multi-scale spatial-spectral attention guided fusion network for pansharpening is proposed. First, the spatial features from the PAN image and spectral features from the LR-MS image are independently extracted to obtain the shallow features. Then, a spatial-spectral attention feature fusion module (SAFFM) is constructed to guide the reconstruction of spatial-spectral features by generating a guidance map to achieve the fusion of reconstructed features at different scales. In SAFFM, the guidance map is designed to ensure the spatial-spectral consistency of the reconstructed features. Finally, considering the difference between multiply scale features, a multi-level feature integration scheme is proposed to progressively achieve fusion of multi-scale features from different SAFFMs. Extensive experiments validate the effectiveness of the proposed network against other state-of-the-art (SOTA) pansharpening methods in both quantitative and qualitative assessments. The source code will be released at https://github.com/MELiMZ/ssaff.

CCS CONCEPTS

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

KEYWORDS

pansharpening, fusion network, guidance map, multi-scale fusion features

ACM Reference Format:
Yong Yang, Mengzhen Li, Shuying Huang, Hangyuan Lu, Wei Tu, and Weiguo Wan. 2023. Multi-scale Spatial-Spectral Attention Guided Fusion Network for Pansharpening. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3613814

1 INTRODUCTION

Remote sensing satellites enable rapid collection of data over a vast surface area, which can acquire information on natural resources and the ecological environment, offering substantial support for human society and scientific development. However, due to the technological constraints of remote sensing satellite sensors, individual sensors can only capture a very limited amount of radiant energy. Multispectral sensors with narrower bandwidths are capable of obtaining low spatial resolution multispectral (LR-MS) images, while panchromatic (PAN) sensors can obtain high spatial resolution grayscale PAN images [12]. To obtain images with both high spectral and high spatial resolutions, pansharpening technology is proposed to fuse the PAN and LR-MS images to generate high-resolution multispectral (HR-MS) images. HR-MS images are beneficial for improving the accuracy of high-level tasks, such as ground target detection and object classification. Therefore, pansharpening has become a research hotspot in remote sensing and multimodal information fusion due to its significant theoretical and practical importance [5].

In recent years, deep learning-based pansharpening techniques have shown significant growth and have been widely used [4, 8, 17, 30]. Convolutional neural networks (CNNs) based pansharpening methods [6] can generate high-quality HR-MS images by constructing a nonlinear mapping from the input to the output [5]. Most of

*Both authors contributed equally to this research.
†corresponding author.

MM '23, October 29–November 3, 2023, Ottawa, Canada

ACM ISBN 978-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3613814

the early methods directly utilized multiple convolutional layers to extract and fuse features from PAN and LR-MS images [11, 21, 26]. Due to the lack of consideration for feature consistency between two source images, the fusion results obtained by these methods often exhibit spatial and spectral distortion. In addition, the fused features at different levels contain a lot of complementary information. However, many pansharpening methods stack the same residual blocks or other modules, leading to the loss of complementary information in the reconstructed features.

In this paper, to address these problems, a multi-scale spatial-spectral attention guided fusion network for pansharpening is proposed. Firstly, two feature extraction branches with different structures are constructed to extract shallow spatial features and spectral features from the PAN and LR-MS images, respectively. Then, a spatial-spectral attention feature fusion module (SAFFM) is proposed to achieve the fusion of spatial-spectral features at different scales by designing an attention guidance map. To ensure the spatial-spectral information consistency of reconstructed features, a guidance map is designed based on the spatial features from the PAN image and the spectral features from the LR-MS image. Finally, a progressive multi-level feature integration scheme is proposed to integrate the output of the features by different SAFFMs to fully utilize the reconstructed features at different scales. We conduct extensive experiments to analyze the effectiveness of the proposed network and demonstrate its favorable performance.

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

- A multi-scale spatial-spectral attention guided fusion network is proposed by designing a three-stage structure to obtain HR-MS images with better spectral and spatial fidelity.
- To enhance the consistency of spatial-spectral information in the reconstructed fused features, a SAFFM is constructed by designing a spatial-spectral attention guidance map. Through the attention mechanism, the problem of spatial-spectral distortion is effectively solved.
- To improve the spatial detail quality of fused images, a multi-level feature integration scheme is designed by progressively fusing reconstructed features at different scales.

2 RELATED WORKS

2.1 CNN-based Pansharpening Methods

PNN [11] is the first pansharpening method based on CNN. This network can generate HR-MS images due to the excellent non-linear fitting and feature representation capabilities of CNN. Increasing the depth and width of the CNN is the most intuitive approach to further improve the performance of the network. However, the networks designed by simply stacking convolutional layers greatly increase computational costs and parameters and are more susceptible to overfitting. DRPNN [21] used residual blocks to prevent the gradient vanishing of deeper networks, thereby improving the performance of pansharpening. FusionNet [4] subtracted the upsampled LR-MS image from the PAN image and input the result into the residual network to calculate the required spatial information, thus effectively preserving the latent spatial and spectral information. DSCNN-RIE [25] proposed a residual information enhancement strategy, which integrates different levels of residual information and uses multi-scale convolutional filters to extract multi-scale spatial details, thus improving the feature extraction capability of the network. In the CNN-based pansharpening methods, neural networks are used to fuse high-resolution PAN and low-resolution LR-MS images, which improves the performance and becomes a promising technique for enhancing spatial and spectral resolution in remote sensing applications.

2.2 Attention Mechanism

Mnih et al. [13] embedded attention mechanisms into recurrent neural networks, enabling them to focus on key regions in images and extract visual features. This architecture allows the model to adaptively adjust attention based on contextual information. Xu et al. [23] introduced visual attention mechanisms in image captioning tasks, which can adaptively focus on key regions in images to generate more accurate textual descriptions. Vaswani et al. [18] proposed the Transformer attention mechanism, which is more suitable for building long-range dependency relationships. SE-Net [9] assigned weights to each channel and corrected the original features according to the weights of each feature map. The performance of the network is significantly boosted since this enhances the relevance of channels. Spatial attention is also crucial for computer vision tasks because it allows the model to focus on important regions within an image while ignoring irrelevant areas. CBAM [22] combined the channel attention mechanism and the spatial attention mechanism, which is compatible with mainstream network structures such as ResNet and DenseNet, making it easy to embed into existing computer vision networks. The attention mechanism can focus on important features in spatial and channel dimensions and suppress unnecessary features, and thus improve the nonlinear fitting ability of CNN.

3 THE PROPOSED METHOD

In this section, the framework of the proposed multi-scale spatial-spectral attention guided fusion network is first presented, as shown in Figure 1. Then, the network structure, its main components and the loss function are introduced in detail.

3.1 Framework

The proposed network consists of three stages: the feature extraction stage that extracts different input image features, the multi-scale spatial-spectral feature fusion stage that performs the fusion of different-scale features by constructing guidance maps , and the multi-level feature integration stage that progressively integrates multi-scale fused features.

First, in the feature extraction stage, considering the feature differences between two source images, two feature extraction branches with different structures are used to extract the initial features of the source images. One branch containing a deconvolution layer, a convolution layer and two ReLU layers is to achieve spatial upsampling and initial feature extraction of the LR-MS image $M(\in R^{C\times h\times w})$. The initial feature maps extracted from the LR-MS image are denoted as $F_m(\in R^{S\times H\times W})$. Here, C, h, w, and S represent the spectral number, height, width, and channel number of feature maps, respectively. The upsampling scale factor U is 4,

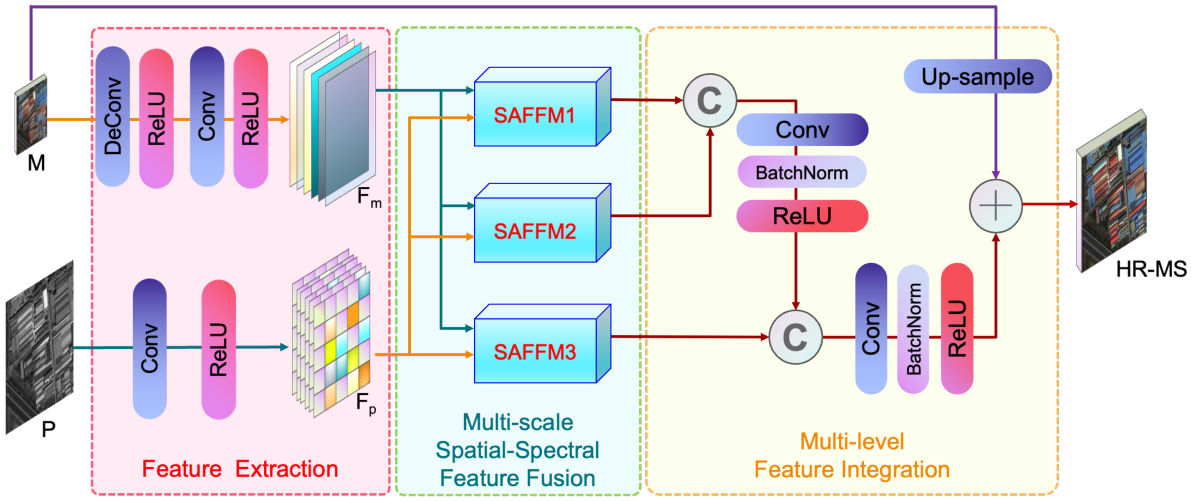


Figure 1: Flowchart of the proposed pansharpening network.

thus $H = Uh$ and $W = Uw$. Another branch containing a convolution layer and a ReLU layer is to extract the initial features of the PAN image $P(\in R^{1\times H\times W})$. The initial feature maps are denoted as $F_p(\in R^{S\times H\times W})$. The initial features F_m and F_p are obtained and simultaneously sent to the next feature fusion stage to achieve the fusion of two features at different scales.

Then, in the multi-scale spatial-spectral feature fusion stage, since the two images are obtained by two different sensors, the features of the two images are inconsistent in both spatial and spectral dimensions. If the features from two images are fused directly, there may be spatial and spectral distortion in the fusion results. Therefore, in this stage, three SAFFMs (SAFFM1, SAFFM2, and SAFFM3) are constructed by designing guidance maps to guide the reconstruction of spatial-spectral features and achieve the fusion of features at different scales. The structure of SAFFM, as shown in Figure 2, is described below.

Finally, to address the differences between multiple scale features, a multi-level feature integration scheme is designed to achieve the fusion of different scale features step by step and obtain the final fusion result.

3.2 SAFFM

Because the two images come from different sensors, PAN contains rich spatial features but lacks spectral features, while LR-MS images contain rich spectral features and some texture structures, resulting in inconsistent spatial and spectral features between them. Therefore, a SAFFM is constructed to realize feature fusion at different scales by designing a spatial-spectral attention guidance map.

First, to improve the consistency of reconstructed features, the guidance map $GM(\in R^{S\times H\times W})$ is designed based on spatial features of P and spectral features of M. The spatial consistency of the fused features can be guided by the spatial attention weight map generated from F_p, and the spectral consistency of the fused features can be guided by the spectral attention weight map generated from F_m. Thus, GM can be obtained by multiplying the spatial attention weight map and the spectral attention weight map, thereby guiding the fusion of spatial-spectral features. In addition, to improve the feature representation ability of SAFFM, spatial consistency is established on multi-scale spatial features obtained by changing the size of convolutional kernels. Note that GMs in SAFFM1, SAFFM2, and SAFFM3 correspond to convolutional kernel sizes k=3, 5, and 7 for spatial feature extraction, respectively. The generation process of GM is defined as follows.

$$F_p^k = ReLU(\mathcal{BN}(Conv_{k\times k}(ReLU(\mathcal{BN}(Conv_{k\times k}(F_p)))))) \quad (1)$$

$$W_p^k = \sigma(Conv_{3\times 3}[MP(F_p^k); AP(F_p^k)]) \quad (2)$$

$$W_m = \sigma(MLP(AP(F_m)) + MLP(MP(F_m))) \quad (3)$$

$$GM^k = W_p^k \odot W_m \quad (4)$$

where $Conv_{k\times k}$ represents a convolutional operation with a convolutional kernel of size k (k=3, 5 or 7), which extracts features of different scales, $\mathcal{BN}$ denotes a Batch Normalization layer, and F_p^k represents feature maps containing features at different scales. W_p^k and GM^k represents the spatial attention weight map and GM corresponding to the convolutional kernel size k, respectively, and W_m is the spectral attention weight map. σ denotes the Sigmoid function, MP and AP denote the maximum pooling and average pooling, respectively, and $[\cdot;\cdot]$ denotes the concatenation.

Then, based on the generated GM, F_p is weighted and guided to reconstruct spectral features, while F_m is weighted and guided to reconstruct spatial features. Finally, the reconstructed features are concatenated and sent to a convolution layer including a convolution layer, a Batch Normalization layer and a ReLU layer to obtain the fusion feature maps F_{out}^k. The process of guiding feature reconstruction and feature fusion is defined as follows.

$$F_{out}^k = ReLU(\mathcal{BN}(Conv_{3\times 3}([GM^k \otimes F_p^k; GM^k \otimes F_m]))) \quad (5)$$

where $\otimes$ denotes element-by-element multiplication, and F_{out}^k (k=3, 5,or 7) is the output of SAFFM (SAFFM1, SAFFM2 or SAFFM3). The output of SAFFM1, SAFFM2, and SAFFM3 are sent to the multi-level feature integration stage to gradually achieve feature integration and obtain fusion results.

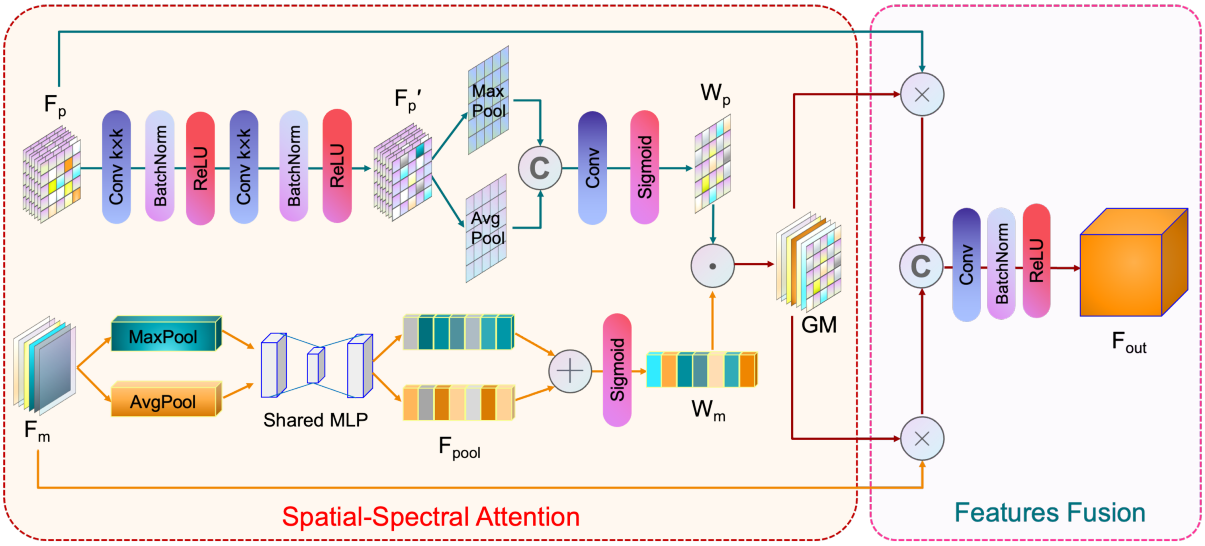


Figure 2: The structure of spatial-spectral attention feature fusion module (SAFFM). k=3,5,7.

3.3 Multi-level Feature Integration Stage

Because SAFFM1, SAFFM2, and SAFFM3 respectively employ spatial features of different scales to guide feature reconstruction, the fused features output by the three modules contain features of different scales. If these features are directly integrated using simple concatenation or addition operations, it may lead to edge blurring in the fused HR-MS image. Therefore, a multi-level feature integration scheme is designed. First, the outputs of SAFFM1 and SAFFM2 are concatenated and passed through a convolutional layer to obtain the first-level fusion features. Then, the first-level fusion features are fused with the output of SAFFM3 in the same way to obtain the second-level fusion features. The residual operation is constructed between the second-level features and the input MS image to obtain the final fusion result. The process of this stage is as follows.

$$F_1 = ReLU(\mathcal{BN}(Conv_{3\times3}([F^3_{out}, F^5_{out}]))) \tag{6}$$

$$F_2 = ReLU(\mathcal{BN}(Conv_{3\times3}([F_1, F^7_{out}]))) \tag{7}$$

$$FM = F_2 + UM \tag{8}$$

where F_1 and F_2 denotes the first and second level fusion feature maps. $FM(\in R^{C\times H\times W})$ denotes the fused HR-MS image, and UM denotes the upsampled LR-MS image.

3.4 Loss Function

The mean squared error (MSE) is used as the loss function of the proposed network.

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \|\mathcal{F}_\theta(P_i, M_i) - GT_i\|_2^2 \tag{9}$$

where $\mathcal{F}$ represents the proposed network, θ denotes all parameters in the network to be optimized, and $\|\cdot\|_2$ denotes the ℓ_2 norm. GT denotes the ground truth image, and N denotes the number of samples in the training set.

4 EXPERIMENTS

4.1 Experiment Settings

4.1.1 Datasets. All experiments were conducted on three satellite datasets, including Pléiades, IKONOS, and WorldView-3. The Pléiades and IKONOS satellites provide the PAN images and multispectral images with 4-bands, while the WorldView-3 satellite provides the PAN images and multispectral images with 8-bands. In the reduced-resolution experiments, the training set was generated by the Wald protocol [19] since ground truth images are not available. The original multispectral images were filtered by the modulation transfer function (MTF)-matched filter and downsampled to the reduced-resolution to construct the LR-MS images, while the PAN images were downsampled to the original multispectral image scale. The original multispectral images are treated as ground truth images. The resolution of the PAN and LR-MS images are 256×256 and 64×64, respectively. In the full-resolution experiments, the dataset was not down-sampled and the resolution of PAN and LR-MS images are 800×800 and 200×200, respectively .

4.1.2 Benchmarks. A number of SOTA pansharpening methods were compared to prove that the proposed network could generate HR-MS images with better spectral and spatial quality. These methods include: GSA [2], DRPNN [21], MSDCNN [26], PanColor-GAN [15], TFNet [10], FusionNet [4], PCDRN [24], DSCNN [25], and TDNet [28].

4.1.3 Evaluation Metrics. We used some image evaluation metrics commonly used in pansharpening to objectively evaluate the different methods. In the reduced-resolution experiments, the metrics include PSNR [14], RMSE [3], UIQI [20], $Q2^n$ [7], SAM [27], and ERGAS [29]. In the full-resolution experiments, the non-reference metrics include: D_λ, D_s, and QNR [3].

4.1.4 Implementation Details. Our network was trained and tested by PyTorch as the deep learning framework. The GPU device is an

Table 1: Quantitative evaluation of the most representative methods for the reduced-resolution Pléiades datasets.

Methods	PSNR(↑)	RMSE(↓)	UIQI(↑)	$Q2^n$(↑)	SAM(↓)	ERGAS(↓)	Time(s)
GSA[Aiazzi et al., 2007]	27.9760	13.9933	0.9530	0.8757	3.5046	3.5725	0.0283
DRPNN [Wei et al., 2017]	30.1216	11.1515	0.9680	0.8636	2.6809	2.9884	0.0424
MSDCNN [Yuan et al., 2018]	29.2428	12.2007	0.9629	0.8638	2.7117	3.2713	0.0295
PanColorGAN [Ozcelik et al., 2020]	23.1587	13.1564	0.9320	0.8031	7.8692	8.4645	0.0180
TFNet [Liu et al., 2020]	32.8707	8.5569	0.9805	0.9452	2.9289	2.0643	0.0201
FusionNet [Deng et al., 2020]	31.3848	10.4311	0.9794	0.9420	2.5486	2.5394	0.0142
PCDRN [Yang et al., 2020]	31.5914	9.7448	0.9767	0.9004	2.3488	2.6322	0.0153
DSCNN [Yang et al., 2022]	32.0029	9.0912	0.9787	0.9109	2.3126	2.4539	0.0273
TDNet [Zhang et al., 2022]	31.3226	10.7110	0.9117	0.9164	3.5946	2.6642	0.0124
Ours	**36.2279**	**5.4467**	**0.9914**	**0.9681**	**2.1979**	**1.3316**	0.0201

Table 2: Quantitative evaluation of the most representative methods for the reduced-resolution IKONOS datasets.

Methods	PSNR(↑)	RMSE(↓)	UIQI(↑)	$Q2^n$(↑)	SAM(↓)	ERGAS(↓)	Time(s)
GSA[Aiazzi et al., 2007]	27.0956	14.2779	0.9352	0.8363	4.6886	3.9925	0.0291
DRPNN [Wei et al., 2017]	26.9798	13.8740	0.9273	0.8154	4.0383	3.7677	0.0454
MSDCNN [Yuan et al., 2018]	27.1694	13.5968	0.9310	0.8239	4.0121	3.6587	0.0346
PanColorGAN [Ozcelik et al., 2020]	24.0976	12.4628	0.9309	0.8224	8.4360	6.0984	0.0332
TFNet [Liu et al., 2020]	30.9652	9.1658	0.9702	0.9227	3.2528	2.4465	0.0221
FusionNet [Deng et al., 2020]	28.1699	18.4607	0.9561	0.8763	4.4868	4.0201	0.0167
PCDRN [Yang et al., 2020]	28.8464	11.3740	0.9516	0.8650	3.5955	3.0352	0.0223
DSCNN [Yang et al., 2022]	29.0473	10.6432	0.9514	0.8646	3.4841	2.9886	0.0317
TDNet [Zhang et al., 2022]	29.6391	9.2735	0.8737	0.8783	3.7885	3.0256	0.0121
Ours	**33.8072**	**6.6281**	**0.9834**	**0.9508**	**2.3270**	**1.7352**	0.0202

Table 3: Quantitative evaluation of the most representative methods for the reduced-resolution WorldView-3 datasets.

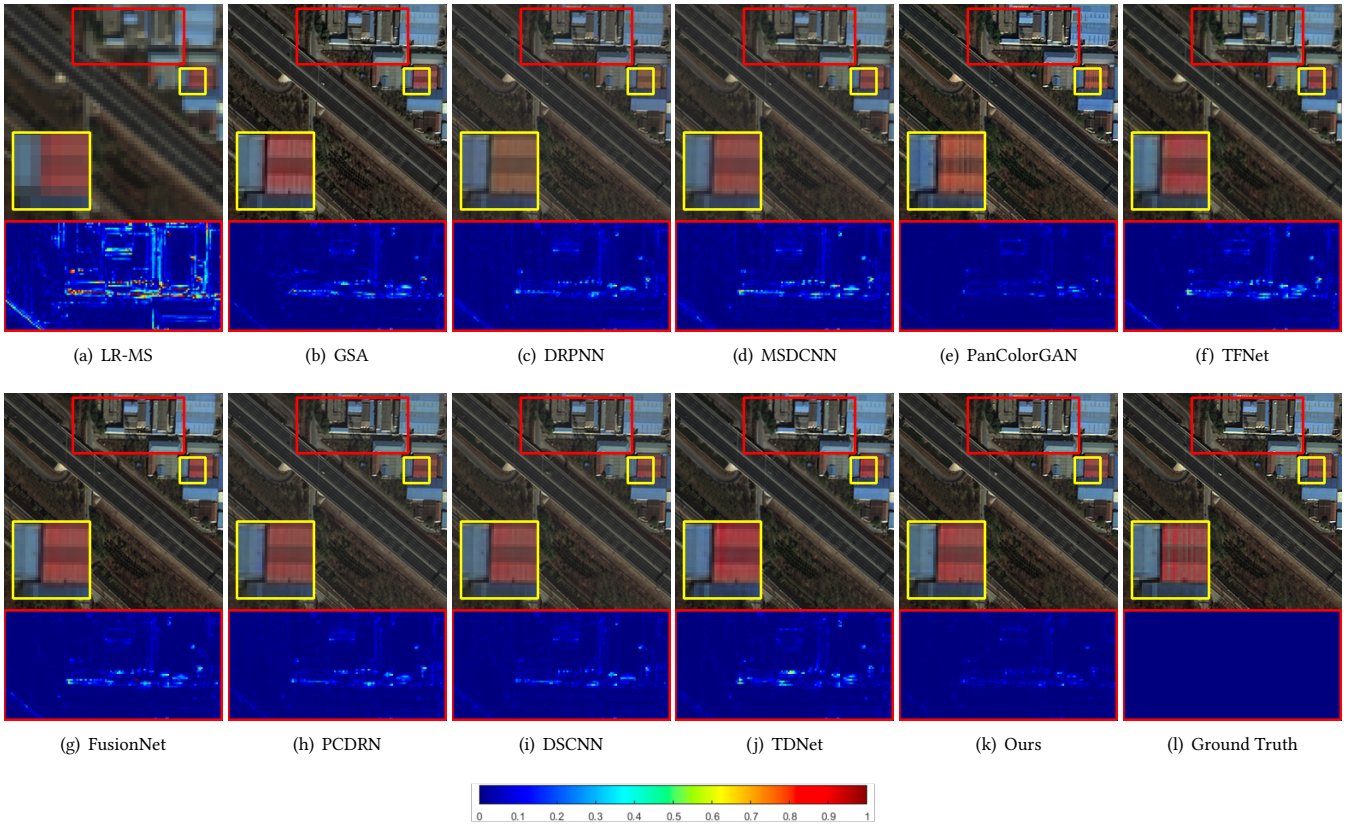
Methods	PSNR(↑)	RMSE(↓)	UIQI(↑)	$Q2^n$(↑)	SAM(↓)	ERGAS(↓)	Time(s)
GSA[Aiazzi et al., 2007]	27.2705	14.1655	0.9358	0.8688	6.6912	4.3425	0.0459
DRPNN [Wei et al., 2017]	25.6762	19.5398	0.9224	0.8152	7.9339	6.1439	0.0572
MSDCNN [Yuan et al., 2018]	25.2981	19.8476	0.9171	0.8033	7.7049	6.2930	0.0469
PanColorGAN [Ozcelik et al., 2020]	25.3901	13.9954	0.9149	0.8302	8.3228	5.6929	0.0258
TFNet [Liu et al., 2020]	28.2917	12.9021	0.9047	0.8921	5.3725	3.7714	0.0241
FusionNet [Deng et al., 2020]	27.2539	15.1852	0.8999	0.8818	5.1426	4.5063	0.0197
PCDRN [Yang et al., 2020]	27.8324	13.2502	0.9460	0.8681	5.3713	4.0870	0.0184
DSCNN [Yang et al., 2022]	28.1630	13.0162	0.9472	0.8707	5.0543	3.9331	0.0325
TDNet [Zhang et al., 2022]	30.1461	12.4702	0.8504	0.8936	5.3467	3.7361	0.2931
Ours	**31.8543**	**10.0735**	**0.9743**	**0.9370**	**4.3165**	**2.6150**	0.0222

NVIDIA GeForce RTX 3090. Adam was used as the optimizer. The initial learning rate was set to 0.003, and multiplied by 0.5 per 100 epochs.

4.2 Reduced-Resolution Assessment

Tables 1, 2, and 3 show the average quantitative evaluation metrics and test time for different pansharpening methods on the reduced-resolution Pléiades, IKONOS, and WorldView-3 satellite datasets, respectively. The best results are shown in bold. It can be seen that the proposed network is optimal in all evaluation metrics, which proves that the network is effective. In particular, ERGAS can objectively reflect the spatial-spectral quality of HR-MS images, while SAM can measure the degree of spectral distortion. Compared with other pansharpening methods, the ERGAS and SAM metrics of the proposed network are closer to the ideal values. Therefore, the proposed network can generate HR-MS images with better spatial-spectral quality. Moreover, the proposed network can create a decent trade-off and achieve the best performance while using less time.

In addition, to demonstrate that the proposed network is more effective from a subjective point of view, the visualization results of different methods are shown in Figure 3. From the enlarged yellow box, it can be observed that the red building edges produced by the GSA and TFNet methods lose partial spectral information. The DRPNN, PanColorGAN, and DSCNN models exhibit varying degrees of color deviation. The MSDCNN, FusionNet, PCDRN, and



(a) LR-MS (b) GSA (c) DRPNN (d) MSDCNN (e) PanColorGAN (f) TFNet

(g) FusionNet (h) PCDRN (i) DSCNN (j) TDNet (k) Ours (l) Ground Truth

Figure 3: Visual comparisons in natural colors and MSE residual maps of the most representative methods on the WorldView-3 dataset.

TDNet models have lost more spatial information, such as stripes on the red buildings.

The MSE residual maps for the red-boxed portion of each image in Figure 3 are shown at the bottom. The fewer residuals in the MSE residual maps, the better the fusion results. The proposed method contains the minimum residuals. This further demonstrates the better performance of the proposed network.

4.3 Full-Resolution Assessment

Pre-trained models based on reduced-resolution datasets are applied to previously unseen full-resolution datasets to evaluate the performance of our network at full-resolution and the generalization abilities of the models. Table 4 shows the average quantitative evaluation metrics and test time of different methods on the full-resolution Pléiades and WorldView-3 satellite datasets, respectively. The best results are shown in bold. QNR is a crucial indicator for evaluating fused images at full-resolution as it allows for a comprehensive assessment of both spatial and spectral distortion. In comparative analyses of QNR metrics, the proposed network demonstrates superior performance compared to other pansharpening methods.

Figure 4 shows the visual comparison of the fusion results of full-resolution images. The EXP image is obtained by upsampling the LR-MS image to the size of the PAN image using the EXP interpolation method [1]. From the enlarged area in the yellow box, it can be seen that the proposed method generates the HR-MS image with richer spectral information and clearer sidewalk contours. In addition, from the enlarged area in the red box, it can be seen that the pansharpened results of GSA, MSDCNN, TFNet, and FusionNet have artifacts, while the pansharpened images of methods such as DRPNN, PanColorGAN, and TDNet lose partial spectral information. The proposed method not only fully preserves the spectral information but also has high spatial-spectral consistency in the pansharpened images.

4.4 Ablation Study

To validate the effectiveness of the proposed network, we compare the performance of the algorithm with and without important modules such as SAFFM, multi-scale convolutional kernels, and progressive feature integration, on the reduced-resolution WorldView-3 satellite dataset. To explore the positive impact of the SAFFM, we conducted experiments on it by observing the network performance change by adding and removing SAFFM from the proposed network. The results are presented in Table 5, which shows that the performance of the network without SAFFM decreases compared to that of the proposed network with SAFFM. This is because simply

Table 4: Quantitative evaluation of the most representative methods for the full-resolution Pléiades and WorldView-3 datasets.

Methods	Pléiades				WorldView-3			
	$D_\lambda(\downarrow)$	$D_s(\downarrow)$	QNR(↑)	Time(s)	$D_\lambda(\downarrow)$	$D_s(\downarrow)$	QNR(↑)	Time(s)
GSA[Aiazzi et al., 2007]	0.1347	0.1619	0.7259	0.4150	0.0872	0.1195	0.8053	0.5991
DRPNN [Wei et al., 2017]	0.0533	0.0363	0.9123	0.3083	0.0887	0.0869	0.8341	0.4398
MSDCNN [Yuan et al., 2018]	0.0434	0.0462	0.9123	0.1907	0.0872	0.0944	0.8287	0.3519
PanColorGAN [Ozcelik et al., 2020]	0.0434	0.0618	0.8975	0.0334	0.1126	0.1138	0.7892	0.0509
TFNet [Liu et al., 2020]	0.0318	0.0310	0.9381	0.0310	**0.0163**	0.0464	0.9206	0.0435
FusionNet [Deng et al., 2020]	0.0403	0.0285	0.9328	0.0289	0.0625	0.0468	0.8953	0.0361
PCDRN [Yang et al., 2020]	0.0375	0.0382	0.9256	0.0925	0.0602	0.0650	0.8791	0.0996
DSCNN [Yang et al., 2022]	0.0242	**0.0263**	0.9501	0.1540	0.0306	0.0426	0.9283	0.1603
TDNet [Zhang et al., 2022]	0.1008	0.1006	0.8093	0.0637	0.0661	0.0870	0.8534	0.0846
Ours	**0.0103**	0.0399	**0.9502**	0.1750	0.0205	**0.0176**	**0.9623**	0.1779

(a) PAN (b) EXP (c) GSA (d) DRPNN (e) MSDCNN (f) PanColorGAN

(g) TFNet (h) FusionNet (i) PCDRN (j) DSCNN (k) TDNet (l) Ours

Figure 4: Visual comparisons of the fusion results of a full-resolution image from WorldView-3 dataset.

Table 5: Ablation experiments are conducted for the SAFFM.

	PSNR(↑)	RMSE(↓)	UIQI(↑)	$Q2^n$(↑)	SAM(↓)	ERGAS(↓)
w/o SAFFM	31.2148	10.7938	0.9717	0.9315	4.4408	2.7677
w/ SAFFM	**31.8543**	**10.0735**	**0.9743**	**0.9370**	**4.3165**	**2.6150**

Table 6: Ablation experiments are conducted for multi-scale convolution kernels.

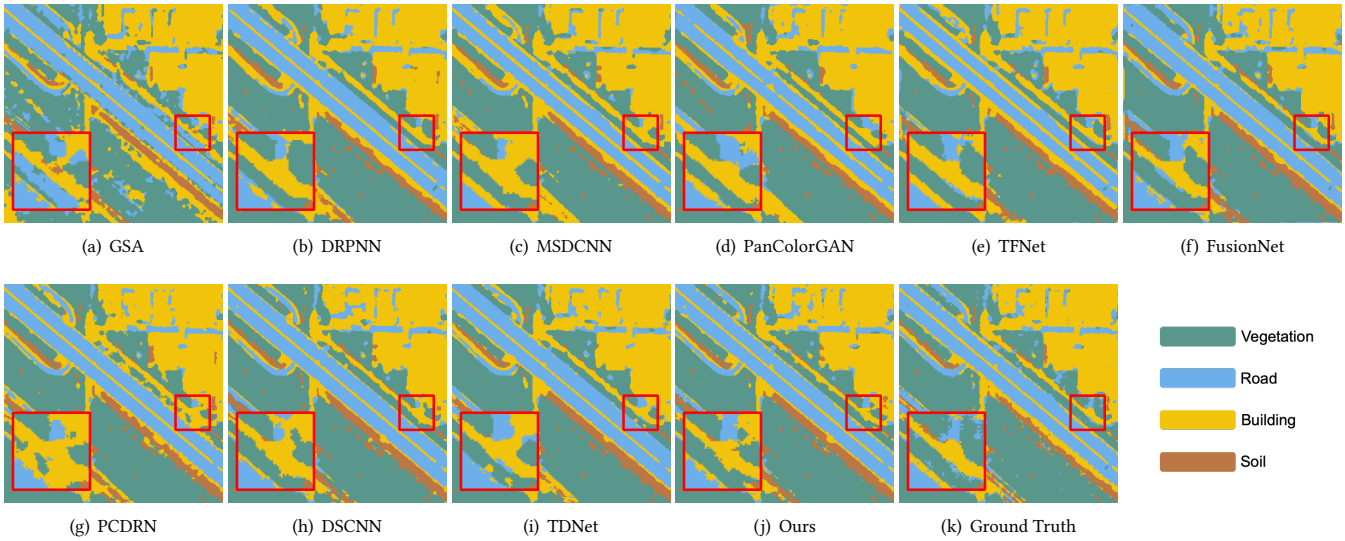
	PSNR(↑)	RMSE(↓)	UIQI(↑)	$Q2^n$(↑)	SAM(↓)	ERGAS(↓)
w/o Muti-scale	31.6345	10.3377	0.9734	0.9354	4.3685	2.6756
w/ Muti-scale	**31.8543**	**10.0735**	**0.9743**	**0.9370**	**4.3165**	**2.6150**

concatenating different input features will lead to high inconsistency in the spatial-spectral features of the fused features, which will reduce the quality of HR-MS images.

Table 7: Ablation experiments are conducted for progressive feature integration.

	PSNR(↑)	RMSE(↓)	UIQI(↑)	$Q2^n$(↑)	SAM(↓)	ERGAS(↓)
w/o progressive intregation	31.3001	10.8348	0.9714	0.9308	4.4481	2.7853
w/ progressive intregation	**31.8543**	**10.0735**	**0.9743**	**0.9370**	**4.3165**	**2.6150**

The second set of ablation experiments used 3×3 convolutional kernels for calculating spatial attention weights in all SAFFMs, to verify the necessity of multi-scale convolution. The objective metrics are shown in Table 6, which indicates that using multi-scale convolution kernels achieves better performance compared to using the same kernel for all SAFFM. This demonstrates that convolution kernels of different sizes have different abilities to extract features at different scales, and using convolution kernels of multiple scales can more fully extract the spatial features.



(a) GSA (b) DRPNN (c) MSDCNN (d) PanColorGAN (e) TFNet (f) FusionNet

(g) PCDRN (h) DSCNN (i) TDNet (j) Ours (k) Ground Truth

Figure 5: Classification results of the images in Figure 3.

Table 8: Quantitative evaluation of the classification results of the reduced-resolution WorldView-3 dataset.

Methods	OA(%)(↑)	KC(↑)
GSA[Aiazzi et al., 2007]	71.1380	0.5859
DRPNN [Wei et al., 2017]	75.5707	0.6499
MSDCNN [Yuan et al., 2018]	76.7853	0.6658
PanColorGAN [Ozcelik et al., 2020]	76.3748	0.6611
TFNet [Liu et al., 2020]	75.2075	0.6449
FusionNet [Deng et al., 2020]	77.8107	0.6836
PCDRN [Yang et al., 2020]	78.3096	0.6900
DSCNN [Yang et al., 2022]	78.7506	0.6970
TDNet [Zhang et al., 2022]	77.1500	0.6723
Ours	**80.1819**	**0.7167**

To demonstrate the effectiveness of progressive multi-level feature integration, the output from all SAFFMs are directly concatenated for feature integration. The objective metrics are shown in Table 7, which shows progressive multi-level feature integration can better integrate multi-scale fusion features compared to directly concatenating them. This demonstrates that the progressive multi-level feature integration strategy can fully utilize the reconstructed features at different scales.

4.5 Classification Experiments

High-quality HR-MS images can provide more powerful data support in a variety of ground analysis tasks. To verify that the proposed network can improve the accuracy of image classification, the HR-MS images generated in reduced-resolution experiments are classified using the support vector machine-based classification algorithm. Taking the fused images in Figure 3 as an example, their classification results are shown in Figure 5. From the enlarged red box, it can be seen that the classification map of the proposed method is most similar to that of the ground truth image, especially in the left-bottom area, where the division between the path and roadside vegetation is more distinct and orderly.

To quantitatively evaluate the classification results, objective metrics are used to compare different methods, including overall accuracy (OA) and kappa coefficient (KC) [16]. Table 8 shows the objective metrics of HR-MS images generated by different methods in the classification experiments and the best results are shown in bold. From the table, we can see that the proposed network can achieve the best classification results in all metrics, which proves the effective performance of our pansharpening method.

5 CONCLUSION

In this paper, we propose a novel multi-scale spatial-spectral attention guided fusion network to address the spatial-spectral distortion in HR-MS images and improve the quality of pansharpening. The input features from different modalities are fused via a guidance map within the SAFFM, which significantly improves the consistency of spatial-spectral information in the fused features. The progressive integration of multi-scale fused features further enhances the spatial quality of the fused image. Extensive experiments on multiple satellite datasets demonstrate the effectiveness and excellent generalization of the proposed network. In the future, we will investigate the feasibility of incorporating our proposed SAFFM into other multimodal information fusion algorithms to improve their performance.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (No. 62072218, No. 61862030, and 62261025), and by the Project of the Education Department of Jiangxi Province (No. GJJ2201330).

REFERENCES

[1] Bruno Aiazzi, Luciano Alparone, Stefano Baronti, and Andrea Garzelli. 2002. Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis. *IEEE Transactions on Geoscience and Remote Sensing* 40, 10 (2002), 2300–2312. https://doi.org/10.1109/TGRS.2002.803623

[2] 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. https://doi.org/10.1109/TGRS.2007.901007

[3] Luciano Alparone, Bruno Aiazzi, Stefano Baronti, Andrea Garzelli, Filippo Nencini, and Massimo Selva. 2008. Multispectral and panchromatic data fusion assessment without reference. *Photogrammetric Engineering & Remote Sensing* 74, 2 (2008), 193–200.

[4] Liang-Jian Deng, Gemine Vivone, Cheng Jin, and Jocelyn Chanussot. 2020. Detail injection-based deep convolutional neural networks for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing* 59, 8 (2020), 6995–7010. https://doi.org/10.1109/TGRS.2020.3031366

[5] Liang-Jian Deng, Gemine Vivone, Mercedes E Paoletti, Giuseppe Scarpa, Jiang He, Yongjun Zhang, Jocelyn Chanussot, and Antonio Plaza. 2022. Machine Learning in Pansharpening: A benchmark, from shallow to deep networks. *IEEE Geoscience and Remote Sensing Magazine* 10, 3 (2022), 279–315. https://doi.org/10.1109/MGRS.2022.3187652

[6] Xueyang Fu, Wu Wang, Yue Huang, Xinghao Ding, and John Paisley. 2020. Deep multiscale detail networks for multiband spectral image sharpening. *IEEE Transactions on Neural Networks and Learning Systems* 32, 5 (2020), 2090–2104. https://doi.org/10.1109/TNNLS.2020.2996498

[7] Andrea Garzelli and Filippo Nencini. 2009. Hypercomplex quality assessment of multi/hyperspectral images. *IEEE Geoscience and Remote Sensing Letters* 6, 4 (2009), 662–665. https://doi.org/10.1109/LGRS.2009.2022650

[8] Meiqi Gong, Jiayi Ma, Han Xu, Xin Tian, and Xiao-Ping Zhang. 2022. D2TNet: A ConvLSTM Network With Dual-Direction Transfer for Pan-Sharpening. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–14. https://doi.org/10.1109/TGRS.2022.3169134

[9] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, Salt Lake City, Utah, 7132–7141. https://doi.org/10.48550/arXiv.1709.01507

[10] Xiangyu Liu, Qingjie Liu, and Yunhong Wang. 2020. Remote sensing image fusion based on two-stream fusion network. *Information Fusion* 55 (2020), 1–15. https://doi.org/10.1016/j.inffus.2019.07.010

[11] Giuseppe Masi, Davide Cozzolino, Luisa Verdoliva, and Giuseppe Scarpa. 2016. Pansharpening by convolutional neural networks. *Remote Sensing* 8, 7 (2016), 594. https://doi.org/10.3390/rs8070594

[12] Xiangchao Meng, Huanfeng Shen, Huifang Li, Liangpei Zhang, and Randi Fu. 2019. Review of the pansharpening methods for remote sensing images based on the idea of meta-analysis: Practical discussion and challenges. *Information Fusion* 46 (2019), 102–113. https://doi.org/10.1016/j.inffus.2018.05.006

[13] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. 2014. Recurrent models of visual attention. *Advances in neural information processing systems* 27 (2014). https://proceedings.neurips.cc/paper_files/paper/2014/file/09c6c3783b4a70054da74f2538ed47c6-Paper.pdf

[14] Zahra Hashemi Nezhad, Azam Karami, Rob Heylen, and Paul Scheunders. 2016. Fusion of hyperspectral and multispectral images using spectral unmixing and sparse coding. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9, 6 (2016), 2377–2389. https://doi.org/10.1109/JSTARS.2016.2528339

[15] Furkan Ozcelik, Ugur Alganci, Elif Sertel, and Gozde Unal. 2020. Rethinking CNN-based pansharpening: Guided colorization of panchromatic images via GANs. *IEEE Transactions on Geoscience and Remote Sensing* 59, 4 (2020), 3486–3501. https://doi.org/10.1109/TGRS.2020.3010441

[16] Pouria Ramzi, Farhad Samadzadegan, and Peter Reinartz. 2013. Classification of hyperspectral data using an AdaBoostSVM technique applied on band clusters. *IEEE journal of selected topics in applied earth observations and remote sensing* 7, 6 (2013), 2066–2079. https://doi.org/10.1109/JSTARS.2013.2292901

[17] Wei Tu, Yong Yang, Shuying Huang, Weiguo Wan, Lixin Gan, and Hangyuan Lu. 2022. MMDN: Multi-Scale and Multi-Distillation Dilated Network for Pansharpening. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–14. https://doi.org/10.1109/TGRS.2022.3179449

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[19] 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, 6 (1997), 691–699.

[20] Zhou Wang and Alan C Bovik. 2002. A universal image quality index. *IEEE signal processing letters* 9, 3 (2002), 81–84. https://doi.org/10.1109/97.995823

[21] Yancong Wei, Qiangqiang Yuan, Huanfeng Shen, and Liangpei Zhang. 2017. Boosting the accuracy of multispectral image pansharpening by learning a deep residual network. *IEEE Geoscience and Remote Sensing Letters* 14, 10 (2017), 1795–1799. https://doi.org/10.1109/LGRS.2017.2736020

[22] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision*. Springer Link, Munich, Germany, 3–19. https://doi.org/10.48550/arXiv.1807.06521

[23] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. PMLR, 2048–2057.

[24] Yong Yang, Wei Tu, Shuying Huang, and Hangyuan Lu. 2020. PCDRN: Progressive cascade deep residual network for pansharpening. *Remote Sensing* 12, 4 (2020). https://doi.org/10.3390/rs12040676

[25] Yong Yang, Wei Tu, Shuying Huang, Hangyuan Lu, Weiguo Wan, and Lixin Gan. 2022. Dual-Stream Convolutional Neural Network With Residual Information Enhancement for Pansharpening. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–16. https://doi.org/10.1109/TGRS.2021.3098752

[26] 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. https://doi.org/10.1109/JSTARS.2018.2794888

[27] Roberta H Yuhas, Alexander FH Goetz, and Joe W Boardman. 1992. 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*.

[28] Tian-Jiang Zhang, Liang-Jian Deng, Ting-Zhu Huang, Jocelyn Chanussot, and Gemine Vivone. 2022. A Triple-Double Convolutional Neural Network for Panchromatic Sharpening. *IEEE Transactions on Neural Networks and Learning Systems* (2022), 1–14. https://doi.org/10.1109/TNNLS.2022.3155655

[29] Yongjun Zhang, Chi Liu, Mingwei Sun, and Yangjun Ou. 2019. Pan-sharpening using an efficient bidirectional pyramid network. *IEEE Transactions on Geoscience and Remote Sensing* 57, 8 (2019), 5549–5563. https://doi.org/10.1109/TGRS.2019.2900419

[30] 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* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 3365–3374. https://doi.org/10.1145/3503161.3547774