Letter

Dual-Branch Multi-Level Feature Aggregation Network for Pansharpening

Gui Cheng, Zhenfeng Shao, Jiaming Wang,
Xiao Huang, and Chaoya Dang

Dear Editor,

In pansharpening task, the most existing deep-learning-based pansharpening methods fail to fully utilize the different level features, inevitably leading to spectral or spatial distortions. To address this challenge, in this letter, we propose a dual-branch multi-level feature aggregation network for pansharpening (DMFANet). The experimental results on the WorldView-II (WV-II) and QuickBird (QB) dataset confirmed the notable superiority of our method over the current state-of-the-art methods from quantitative and qualitative point of view. The source code is available at https://github.com/Gui-Cheng/DMFANet.

Introduction: Multispectral (MS) image with a wealth of spectral information has the potential to distinguish the surface materials and thus owns a broad remote sensing application. Due to the technical limitations, there exists a trade-off in remote sensing sensors between the spatial and spectral resolutions [1]. As a consequence, it is challenging to directly acquire images with high spatial and spectral resolution via a single sensor. However, the panchromatic (PAN) image with high spatial resolution and the corresponding multispectral (LRMS) image with low spatial resolution widely exist, which can not meet the needs of high-precision remote sensing applications to a certain degree. To address this challenge, the pansharpening technique is applied to integrate the spatial structure information from the PAN image and the spectral information from the LRMS image to generate the high-resolution multispectral (HRMS) image.

In the past few decades, numerous pansharpening methods have been proposed, which can be broadly divided into four major categories: 1) component substitution (CS)-based methods [2]; 2) multiresolution analysis (MRA)-based methods [3]; 3) hybrid methods [4]; 4) deep-learning-based methods [5].

In recent years, the CNN-based pansharpening methods have been developed and achieved promising results, such as PNN [6], MSDCNN [7], Pan-GAN [8], GTP-PNet [9], GPPNN [10]. However, some problems still remain to be solved. The most existing deep-learning-based pansharpening methods fail to fully utilize the different level features, inevitably leading to spectral or spatial distortions.

To address these challenges, a dual-branch multi-level feature aggregation network for pansharpening is proposed, called DMFANet. The main branch of DMFANet is the MS image multi-level feature extraction and aggregation branch to obtain the final HRMS image. Another branch is the PAN image feature extraction branch that provides high spatial structure information for the main branch. Specially, we conduct multi-level feature fusion throughout the whole network for better usage of the multi-level spectral and spatial information from MS image and PAN image. Inspired by the high efficient residual feature aggregation (RFA) framework [11], we also designed two RFA framework-based feature extraction modules for MS image and PAN image respectively, named MS image feature extraction module (MSFEM) and PAN image feature extraction module (PFEM). MSFEM aims to extract the spectral features from MS images, while the PFEM aims to extract spatial details from PAN images.

The main contributions of this study are summarized as follow: 1) We design a dual-branch network to fully extract the spectral features from MS image and spatial features from PAN image respectively. 2) We apply multi-level feature fusion throughout the whole network to take advantage of the multi-level effective information from PAN and MS images. 3) We design two high efficient feature extraction module, i.e., the MSFEM and PFEM.

Problem formulation: The target of our DMFANet is to extract the spectral features from MS image and spatial features from PAN image as much and as accurately as possible via a dual-branch network, fuse them at different feature levels, and aggregate fused features to make full use of the multi-level spectral and spatial information for generating promising fusion results. Fig. 1 presents the overall fusion framework of our DMFANet. We donate the LRMS image as I_{LRMS} and the corresponding PAN image as I_{PAN}. Our goal is to generate the HRMS image (I_{HRMS})

$$I_{\mathrm{HRMS}} = f((I_{\mathrm{LRMS}}, I_{\mathrm{PAN}}); \Theta) \tag{1}$$

where $f(\cdot)$ denotes the operation of our DMFANet, Θ refers to the trainable parameters of our network.

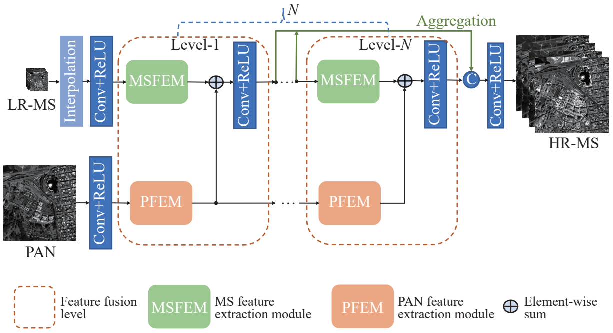


Fig. 1. The overall fusion framework of our DMFANet.

To be more specific, we extract spectral and spatial features from two branches and fuse them at different levels. We formulate the multi-level fusion function as follow:

$$D_i^{Fused} = H(D_i^{MS}, D_i^{PAN}) \tag{2}$$

where D_i^{Fused} represents the i-th level fusion features, $H(\cdot)$ denotes the feature fusion function, which represents element-wise sum operation, D_i^{MS} and D_i^{PAN} denote the i-th level features extracted from MS branch and PAN branch respectively, which can be formulated as follow:

$$D_i^{MS} = f_{MS}(D_{i-1}^{Fused}) \tag{3}$$

$$D_i^{PAN} = f_{\mathrm{PAN}}(D_{i-1}^{PAN}) \tag{4}$$

where the $f_{MS}(\cdot)$ and $f_{\mathrm{PAN}}(\cdot)$ represent the feature extraction functions of MS and PAN images, respectively.

Finally, an N-level fusion is conducted, with fused features aggregated. The generated HRMS image can be obtained by (5)

$$I_{\mathrm{HRMS}} = f_{\mathrm{conv}}(cat(D_1^{MS}, D_2^{MS}, \ldots, D_N^{MS})). \tag{5}$$

MS image feature extraction module: Despite that MS image contains rich spectral information, it is a challenging task to fully extract their spectral information. In this study, we propose an MS image feature extraction module (MSFEM) (Fig. 2) to complete this task. The proposed MSFEM combines the residual channel attention blocks (RCAB) and the RFA framework. The RCAB [12] integrates the channel attention into a residual block. The residual features are

Corresponding author: Zhenfeng Shao.

Citation: G. Cheng, Z. F. Shao, J. M. Wang, X. Huang, and C. Y. Dang, "Dual-branch multi-level feature aggregation network for pansharpening," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 11, pp. 2023–2026, Nov. 2022.

G. Cheng, Z. F. Shao, J. M. Wang, and C. Y. Dang are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: chenggui@whu.edu.cn; shaozhenfeng@whu.edu.cn; wjmecho@163.com; Chaoyadang99@163.com).

X. Huang is with the Department of Geosciences, University of Arkansas, Fayetteville, AR 72701 USA (e-mail: xh010@uark.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2022.105956

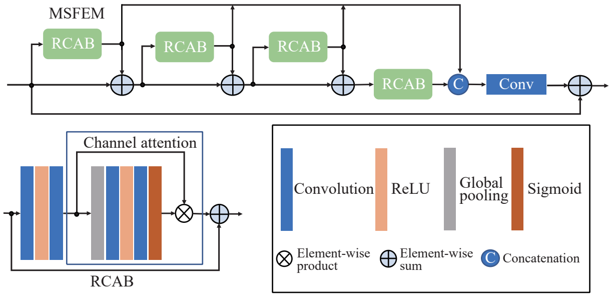
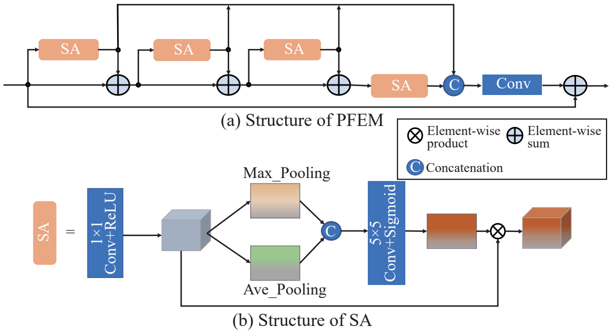


Fig. 2. The architecture of MSFEM.

first extracted by two convolutional layers. Then the channel attention block extracts the channel statistic among channels via a global pooling layer followed by two convolutional layers with a ReLU function and a Sigmoid function, respectively. Therefore, the MSFEM can better extract spectral features with an enhanced discriminative ability. We apply multi-level MSFEMs in our network.

PAN image feature extraction module: The proposed PAN image feature extraction module (PFEM) consists of 4 spatial attention (SA) blocks based on the RFA framework (Fig. 3(a)). The structure of the SA block is detailed in Fig. 3(b). The effectiveness of the spatial attention strategy that leads to the focus on the inter-spatial relationship of features has been verified in many tasks [13]. The SA block first extracts features by a 1 × 1 convolutional layer with a ReLU function. Then, an AvePooling layer and a MaxPooling layer are used to aggregate channel information. Finally, by concatenating these two kinds of features, a 5 × 5 convolutional layer with a Sigmoid function is applied to generate the spatial attention map. The combination of SA blocks and RFA framework results in better extraction of effective features among the spatial dimension from PAN image. In this study, we implement multi-level PFEMs.



(a) Structure of PFEM

(b) Structure of SA

Fig. 3. The architecture of PFEM.

Experimental setup: We perform experiments on WV-II and QB datasets with four MS bands: blue, green, red and NIR. The experimental LRMS and PAN image patches have the size of 64 × 64 × 4 and 256 × 256 × 1, respectively. For WV-II and QB datasets, the number of image patches from training, reduced-resolution testing, and full-resolution testing are 1254 and 308 120 and 80 400 and 200, respectively.

These experiments are conducted on a desktop with two NVIDIA GTX 2080Ti GPUs. Our proposed DMFANet and the deep-learning-based methods are implemented by PyTorch 1.5.1 library with Python 3.6.9. The Adam optimizer is applied to optimize the proposed method with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1e-8$. The learning rate is initialized to $1e-4$. We employ the mean squared error (MSE) as the loss function. All the deep-learning-based methods are trained on reduced-resolution dataset. In the following experiments, our proposed DMFANet is based on 5 fusion levels. We set $N = 5$ in these experiments.

To verify the performance of our DMFANet, we conduct reduced-resolution testing based on Wald's protocol [14] and full-resolution testing. We take both qualitative and quantitative evaluation on these two types of testing. We compare our method with eight mainstream fusion algorithms, including three widely used traditional pansharpening algorithms, i.e., Brovey [2], Gram-Schmidt (GS), MTF-GLP [3], and five deep-learning-based methods, i.e., MSDCNN [7], PNN [6], DIRCNN [15], GPPNN [10], MUCNN [16].

We apply six widely used metrics, i.e., PSNR, SSIM, ERGAS, SAM, UIQI, SCC, for the reduced-resolution testing. For the full-resolution testing, the quality of no reference (QNR) index is utilized to characterize the fusion performance. The QNR consists of two parts: the spectral distortion index (D_λ) and spatial distortion index (D_s).

Results from the WV-II dataset: We firstly present qualitative and quantitative testing results on the WV-II dataset from the aspects of reduced-resolution and full-resolution to demonstrate the performance of each method.

The qualitative testing results under the reduced-resolution are shown in Fig. 4 . Intuitively speaking, our proposed DMFANet presents the highest consistency with the referenced HRMS image. Obviously, the Brovey and GS suffer from spectral distortion, while the MTF_GLP suffer great spatial distortion, resulting blurring details. Compared to traditional methods that suffer from notable spectral and spatial distortion, the comparison deep-learning-based methods are able to better preserve spatial information, but also suffer from a little spectral distortion. Instead, our DMFANet can largely preserve the spectral distribution and spatial structure, thanks to the introduced MS features extraction branch that learns the spectral information and the PAN features extraction branch that preserves the spatial structure detail. The above results demonstrate that our DMFANet not only reconstructs more accurate spectral distribution but also generate reasonable spatial structure details, outperforming other selected methods.

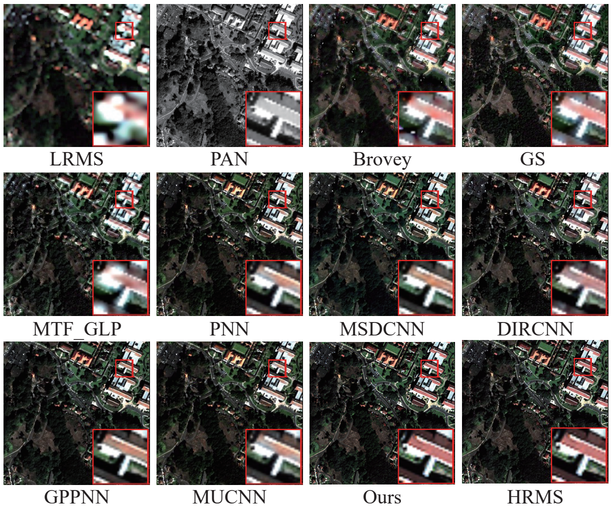


Fig. 4. The qualitative testing results from comparison methods under the reduced-resolution on the WV-II dataset.

The qualitative testing results under the full-resolution (Fig. 5) also demonstrate that our method leads to better spectral information preservation, evidenced by the clearer texture details. For example, we can observe that the spectral distribution of the land is largely consistent with the LRMS image and the spatial structure details of the land are similar to the ones in PAN image.

We further provide the quantitative testing results on WV-II dataset (Table 1). Table 1 displays that DMFANet achieves the best average values on PSNR, SSIM, SAM, ERGAS, SCC, and UIQI, indicating that the fused results generated by our method are most consistent with the reference HRMS image from the aspects of spectral distribution and spatial structure details. Compared with no-reference metrics, our DMFANet ranks the second on D_s and the third on QNR. Nevertheless, it is notable that our qualitative results are better than MTF_GLP, PNN, and GS as shown in Fig. 5. The no-reference metrics apply PAN and interpolated LRMS images as references to quantify spatial and spectral distortion. However, the spectral distribution in interpolated LRMS image and the spatial structure in PAN image can be different from that of the real HRMS image, leading to the fact that our method does not obtain the best performance on these three no-reference metrics.

Results from the QB dataset: To further validate the effectiv-

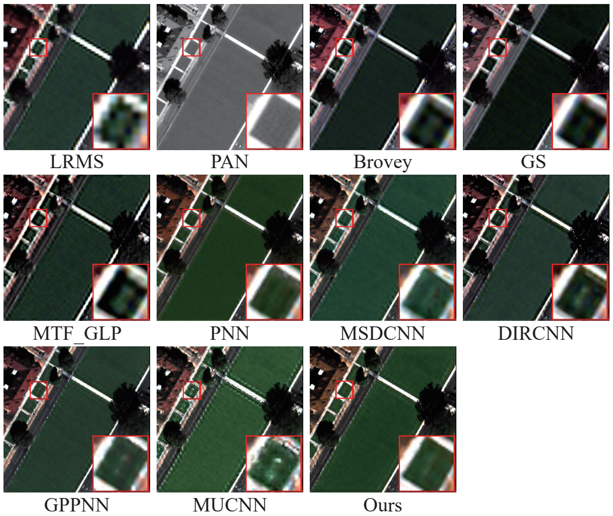


Fig. 5. The qualitative testing results from the comparison methods under the full-resolution on the WV-II dataset.

eness of DMFANet, we conduct comparison experiments on the QB dataset. Fig. 6 shows the qualitative testing results under the reduced-resolution. Compared to traditional methods, ours DMFANet presents a well spectral preservation. The results of MTF_GLP suffer a great spatial distortion. Similarly, the results of the deep-learning based methods such as MSDCNN PNN and MUCNN cannot preserve the spectral information well. For the spatial information preservation, our proposed DMFANet can rebuilt the spatial texture of building, outperforming the comparison methods. The quantitative results (Table 2) under the reduced-resolution testing also illustrate the best performance of methods among the comparison methods. Furthermore the qualitative comparison results under full-resolution in Fig. 7 demonstrate that our proposed DMFANet are most similar to the LRMS image in term of spectral feature and PAN image in terms of spatial feature. Therefore, both the qualitative results and the quantitative results have shown that our proposed DMFANet achieves the best performance compared with the selected competing methods.

Ablation study: To verify the effectiveness of each strategy in our proposed method, we perform the ablation experiments on WV-II dataset. Table 3 records the results of four variants of DMFANet. In the following, we make a detail analysis of each strategy.

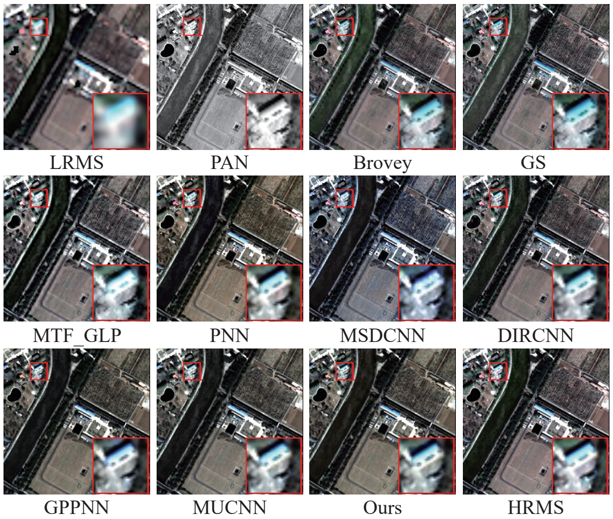


Fig. 6. The qualitative testing results from the comparison methods under the reduced-resolution on the QB dataset.

1) Multi-level feature fusion: To evaluate the best feature fusion level, we perform comparison experiments with the feature fusion level from 1 to 12 based on our proposed DMFANet (Table 4). Through experiments, it can be seen that by increasing the fusion level from 1 to 5, the performance of pansharpening is notably improved. However, the performance of pansharpenging will decrease when the fusion level continues to increase. The reason is that the input MS image and PAN image can already be finely integrated by 5 fusion levels, keeping increasing the fusion level makes the training inefficient.

2) Aggregation structure: To confirm the effectiveness of aggregation structure, we compare the performance of DMFANet and DMFANet without aggregation structure. From the results in the first line at Table 3 , we observe that the performance of DMFANet reduces when the aggregation structure is discarded. For example, the reductions in SSIM and ERGAS are 0.011 and 0.026. The results prove that the aggregation structure contributes to the performance of DMFANet.

3) Dual-branch structure: To verify the superiority of the dual-branch structure, we compare the performance of the model with dual-branch structure and the model with single-branch structure under the same setting of other parameters. The model with single-

Table 1. Quantitative Testing Results on WV-II Dataset

Method	PSNR	SSIM	SAM	ERGAS	SCC	UIQI	D_λ	D_s	QNR
Brovey	18.818	0.623	0.236	19.284	0.760	0.573	0.075	0.090	0.841
GS	19.690	0.619	0.240	18.806	0.796	0.582	0.029	0.076	0.897
MTF-GLP	19.508	0.655	0.264	20.117	0.824	0.622	0.029	0.044	0.929
PNN	22.905	0.797	0.114	12.597	0.901	0.760	0.044	0.077	0.883
MSDCNN	23.213	0.809	0.115	12.394	0.907	0.773	0.058	0.084	0.864
DIRCNN	22.421	0.764	0.167	13.411	0.887	0.721	0.045	0.103	0.858
GPPNN	23.108	0.815	0.106	12.487	0.904	0.777	0.053	0.112	0.843
MUCNN	23.477	0.825	0.107	11.913	0.914	0.790	0.046	0.103	0.857
Ours	23.912	0.840	0.104	11.264	0.921	0.804	0.046	0.069	0.888

Table 2. Quantitative Testing Results on QB Dataset

Method	PSNR	SSIM	SAM	ERGAS	SCC	UIQI	D_λ	D_s	QNR
Brovey	23.903	0.780	0.102	7.119	0.877	0.698	0.044	0.098	0.862
GS	25.022	0.780	0.102	6.896	0.897	0.699	0.028	0.093	0.882
MTF-GLP	25.469	0.788	0.114	6.665	0.924	0.717	0.024	0.048	0.930
PNN	27.621	0.850	0.103	4.847	0.948	0.779	0.068	0.079	0.859
MSDCNN	27.361	0.839	0.101	5.019	0.949	0.771	0.068	0.073	0.864
DIRCNN	28.123	0.846	0.085	4.697	0.949	0.772	0.0295	0.0591	0.913
GPPNN	29.102	0.876	0.073	4.078	0.960	0.811	0.048	0.066	0.889
MUCNN	29.246	0.878	0.076	4.104	0.957	0.812	0.044	0.065	0.894
Ours	29.316	0.881	0.071	4.058	0.961	0.815	0.041	0.066	0.897

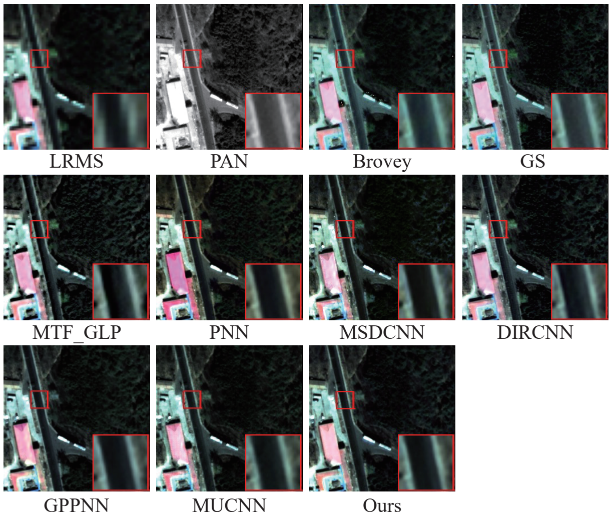


Fig. 7. The qualitative testing results from the comparison methods under the full-resolution on the QB dataset.

Table 3. The Experimental Results of Ablation Study

Aggregation structure	Dual-branch	MSFEM	PFEM	SSIM	SAM	ERGAS	SCC	UIQI
×	√	√	√	0.829	0.108	11.290	0.915	0.798
√	×	√	√	0.826	0.110	11.561	0.917	0.790
√	√	×	√	0.815	0.109	12.062	0.909	0.779
√	√	√	×	0.830	0.108	11.273	0.911	0.796
√	√	√	√	**0.840**	**0.104**	**11.264**	**0.921**	**0.804**

branch structure only contains the MS image multi-level feature extraction and aggregation branch as mentioned in DMFANet. The results in the second line at Table 3 show that the dual-branch structure can significantly improve the performance of pansharpening.

4) MSFEM: To verify the effectiveness of MSFEM, we replace the RCAB with convolutional blocks with the same filter size and other parameters setting in MSFEM and conduct experiments. Through the experimental results in the third line at Table 3, the performance of each metric significantly reduces, especially for the reduction in ERGAS which is 0.798, indicating the spectral distortion increases. The results prove that the MSFEM contributes to the spectral feature extraction.

5) PFEM: Similarly, we replace the SA block with convolutional blocks with the same filter size and other parameters setting in PFEM. The experimental results were recorded in the fourth line at Table 3, we can see that the reductions in SSIM and SCC are 0.01 and 0.01, indicting more spatial distortion. The results confirm that the PFEM contributes to the spatial feature extraction.

Table 4. The Evaluation of Different Fusion Levels Based on DMFANet

Levels	1	2	3	4	5	6	7	8	9	10	11	12
PSNR	22.18	22.84	23.15	23.49	**23.91**	23.74	23.61	23.60	23.50	23.54	23.35	23.15
SSIM	0.776	0.806	0.815	0.821	**0.840**	0.837	0.830	0.831	0.825	0.823	0.816	0.810

Conclusion: In this letter, we propose a dual-branch multi-level feature aggregation network for pansharpening, called DMFANet. Our network consists of two branches designed by the residual feature aggregation framework. The purpose of our DMFANet is to extract the spectral distribution features and spatial structure features in an efficient and comprehensive manner via a dual-branch network, fuse them at multi-levels, and finally aggregate each fused feature, thus taking full advantage of the complementary information for generating promising fusion results. Such a design allows not only the approximation to the HRMS reference image in terms of spectral distribution but also the reconstruction of reasonable spatial structure details. The experimental results from WV-II and QB datasets demonstrate the notable superiority of our method over the current state-of-the-art methods from quantitative and qualitative point of view.

Acknowledgments: This work was supported in part by the National Natural Science Foundation of China (42090012), 03 Special Research and 5G Project of Jiangxi Province in China (20212ABC03A09), and the Open Grants of the State Key Laboratory of Severe Weather (2021LASW-A17).

References

[1] H. Shen, X. Meng, and L. Zhang, "An integrated framework for the spatio-temporal-spectral fusion of remote sensing images," *IEEE Trans. Geoscience and Remote Sensing*, vol. 54, no. 12, pp. 7135–7148, Dec. 2016.

[2] A. R. Gillespie, A. B. Kahle, and R. E. Walker, "Color enhancement of highly correlated images. II. Channel ratio and "chromaticity" transformation techniques," *Remote Sensing Environment*, vol. 22, no. 3, pp. 343–365, Aug. 1987.

[3] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "MTF-tailored multiscale fusion of high-resolution MS and PAN imagery," *Photogrammetric Engineering & Remote Sensing*, vol. 72, no. 5, pp. 591–596, 2006.

[4] S. A. Valizadeh and H. Ghassemian, "Remote sensing image fusion using combining IHS and Curvelet transform," in *Proc. 6th Int. Symposium Telecomm.*, Nov. 2012, pp. 1184−1189.

[5] B. Xiao, B. Xu, X. Bi, and W. Li, "Global-feature encoding U-Net (GEU-Net) for multi-focus image fusion," *IEEE Trans. Image Processing*, vol. 30, pp. 163–175, 2020.

[6] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sensing*, vol. 8, no. 7, p. 594, Jul. 2016. DOI: 10.3390/rs8070594.

[7] Q. Yuan, Y. Wei, X. Meng, H. Shen, and L. Zhang, "A multiscale and multidepth convolutional neural network for remote sensing imagery Pan-Sharpening," *IEEE J. Selected Topics Applied Earth Observations and Remote Sensing*, vol. 11, no. 3, pp. 978–989, Mar. 2018.

[8] J. Ma, W. Yu, C. Chen, P. Liang, X. Guo, and J. Jiang, "PAN-GAN: An unsupervised PAN-sharpening method for remote sensing image fusion," *Information Fusion*, vol. 62, pp. 110–120, 2020.

[9] H. Zhang and J. Ma, "GTP-PNet: A residual learning network based on gradient transformation prior for pansharpening," *ISPRS J. Photogrammetry and Remote Sensing*, vol. 172, pp. 223–239, 2021.

[10] S. Xu, J. Zhang, Z. Zhao, K. Sun, J. Liu, and C. Zhang, "Deep gradient projection networks for PAN-sharpening," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2021, pp. 1366–1375.

[11] J. Liu, W. Zhang, Y. Tang, J. Tang, and G. Wu, "Residual feature aggregation network for image super-resolution," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2020, pp. 2359–2368.

[12] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. European Conf. Computer Vision*, 2018, pp. 286–301.

[13] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. European Conf. Computer Vision*, 2018, pp. 3–19.

[14] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogrammetric Engineering and Remote Sensing*, vol. 63, no. 6, pp. 691–699, 1997.

[15] M. Jiang, H. Shen, J. Li, Q. Yuan, and L. Zhang, "A differential information residual convolutional neural network for pansharpening," *ISPRS J. Photogrammetry and Remote Sensing*, vol. 163, pp. 257–271, 2020.

[16] Y. Wang, L.-J. Deng, T.-J. Zhang, and X. Wu, "SS-Conv: Explicit spectral-to-spatial convolution for pansharpening," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 4472–4480.