MMPN: Multi-supervised Mask Protection Network for Pansharpening

Changjie Chen[1*] , **Yong Yang**[2†] , **Shuying Huang**[3*] , **Wei Tu**[4] , **Weiguo Wan**[5] and **Shengna Wei**[2]
[1]School of Information Management, Jiangxi University of Finance and Economics, Nanchang, China
[2]School of Computer Science and Technology, Tiangong University, Tianjin, China
[3]School of Software, Tiangong University, Tianjin, China
[4]School of Mathematics and Computer Science, Jiangxi Science and Technology Normal University, Nanchang, China
[5]School of Software and Internet of Things Engineering, Jiangxi University of Finance and Economics, Nanchang, China
chencjpro@163.com, greatyangy@126.com, huangshuying@tiangong.edu.cn, 283299985@qq.com, wanweiguo@jxufe.edu.cn, 944961072@qq.com

Abstract

Pansharpening is to fuse a panchromatic (PAN) image with a multispectral (MS) image to obtain a high-spatial-resolution multispectral (HRMS) image. The deep learning-based pansharpening methods usually apply the convolution operation to extract features and only consider the similarity of gradient information between PAN and HRMS images, resulting in the problems of edge blur and spectral distortion in the fusion results. To solve this problem, a multi-supervised mask protection network (MMPN) is proposed to prevent spatial information from being damaged and overcome spectral distortion in the learning process. Firstly, by analyzing the relationships between high-resolution images and corresponding degraded images, a mask protection strategy (MPS) for edge protection is designed to guide the recovery of fused images. Then, based on the MPS, an MMPN containing four branches is constructed to generate the fusion and mask protection images. In MMPN, each branch employs a dual-stream multi-scale feature fusion module (DMFFM), which is built to extract and fuse the features of two input images. Finally, different loss terms are defined for the four branches, and combined into a joint loss function to realize network training. Experiments on simulated and real satellite datasets show that our method is superior to state-of-the-art methods both subjectively and objectively.

1 Introduction

Remote sensing images with high spatial and spectral resolution are widely used in various fields, such as change detection, rescue, navigation, and mapping [Deng *et al.*, 2022]. However, due to the limitations of the physical properties of satellite sensors, high-spatial-resolution multispectral (HRMS) images cannot be directly obtained. Therefore, researchers proposed to fuse the spatial information of the panchromatic (PAN) image and the spectral information of the low-spatial-resolution multispectral (LRMS) image through pansharpening methods to obtain HRMS images. Up to now, in the pansharpening task, some related issues, such as how to extract the spatial details of PAN images more accurately and keep the spatial and spectral consistency between the fused HRMS image and the source images, are still hot research topics.

At present, the existing pansharpening methods are mainly divided into four categories [Vivone *et al.*, 2020], i.e., component substitution (CS) methods [Tu *et al.*, 2001; Xu *et al.*, 2014; Aiazzi *et al.*, 2002], multi-resolution analysis (MRA) methods [Ghassemian, 2016], variational optimization (VO) methods [Lu *et al.*, 2021], and deep learning (DL)-based methods [Deng *et al.*, 2022; Peng *et al.*, 2022]. Among them, CS, MRA, and VO methods, called traditional methods, obtain HRMS images through filter estimation or sparse representation. Traditional pansharpening methods are simple to implement and have physical interpretabilities, but they rely on defined feature extraction methods and fusion rules to ensure the accuracy of results [Ghassemian, 2016].

Due to the powerful feature extraction ability of convolutional neural networks (CNNs), numerous DL-based pansharpening methods have been developed. PNN [Meng *et al.*, 2022] is the first CNN-based pansharpening method, which extracts and fuses the features from PAN and LRMS images. MSDCNN [Yuan *et al.*, 2018] proposed a multi-scale and multi-depth CNN that adopts multi-scale feature extraction and residual connection for the pansharpening. PCDRN [Yang *et al.*, 2020] proposed a cascaded progressive residual network by increasing the depth of the network to obtain more accurate spatial information. ColorGAN [Ozcelik *et al.*, 2020] provided a solution for colorizing PAN images, which constructs a generative adversarial network (GAN) to realize the generation of spectral information. TFNet [Liu *et al.*, 2020] proposed a two-stream fusion network, which extracts features of PAN and LRMS images by constructing two different branches. FusionNet [Deng *et al.*, 2022] proposed a network to estimate the injection details between LRMS and HRMS images. DSCNN [Yang *et al.*, 2021] proposed a

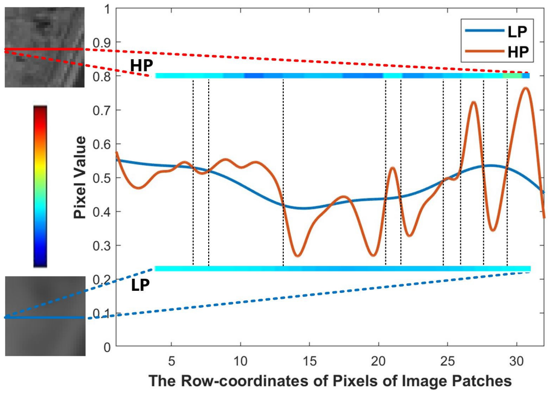


Figure 1: Pixel Distribution Curves of HP and LP.

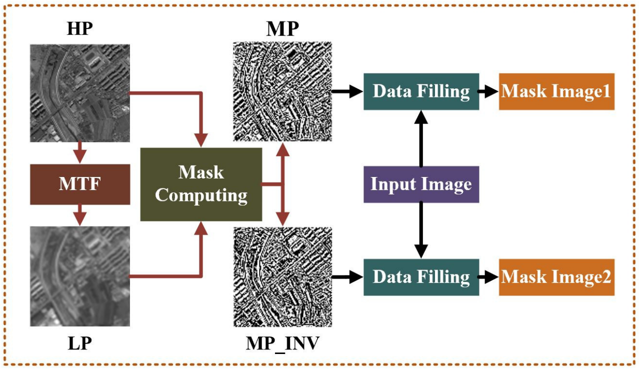


Figure 2: Structure of MPS.

dual-stream CNN for pansharpening by constructing an information complementation block to extract and enhance spatial details at different resolutions. ADKNet [Peng *et al.*, 2022] was proposed for pansharpening by building source-adaptive discriminative kernels from PAN and LRMS images. TDNet [Zhang *et al.*, 2022] proposed a fusion network with double-level, double-branch, and double-direction structures to fully exploit spectral information with MRA.

The DL-based pansharpening methods can achieve better objective performance indexes than the traditional methods. However, their results still have the problems of spectral distortion and edge blur compared to the ideal HRMS images [Deng *et al.*, 2022; Meng *et al.*, 2022]. The main reason is that the convolution operation in CNNs can capture abundant features by increasing the sizes of convolution kernels to expand the receptive field. However, the larger the convolution kernel, the greater the interference of adjacent pixels to the central pixel in the convolution area, so this operation will weaken the gradient information of image edges. In addition, most methods only consider the similarity of gradient information between PAN and HRMS images, but ignore the intensity change of adjacent areas at the edges between LRMS and HRMS images, which leads to spectral distortion in the fusion results. To solve the above problems, this paper first analyzes the change trend of intensity values between high-resolution images and corresponding degraded images, and proposes a mask-protected strategy (MPS) that is used to protect edges and guide the network to separately learn the features of different regions on both sides of image edges. Then, based on MPS, a multi-supervised mask-protection network (MMPN) is constructed to fuse PAN and LRMS images. Finally, to obtain better fusion results and increase the generalization of the network, a joint multi-supervised loss function is defined. The contributions of this work are as follows:

1) An MPS is proposed by defining a mask protection matrix to protect the edge information from the interference of neighborhood information in feature extraction.

2) Based on MPS, an MMPN is constructed for pansharpening, which contains four branches with the same structure for multiple learning tasks. This network can obtain fusion results with dual fidelity of spectral and spatial information.

3) In each branch of MMPN, a dual-stream multi-scale feature fusion module (DMFFM) is designed to better extract and fuse the features of two input images by constructing two encoders and one decoder for feature extraction and fusion, respectively.

4) To better train the network, a multi-supervised loss function based on multiple tasks is defined, in which a down-scale loss term is designed to improve the generalization of the network.

2 Definition of MPS

The current pansharpening methods mainly use the spatial information of PAN images to improve the resolution of LRMS images. However, LRMS images exist low contrast between adjacent target areas, that is, the difference of pixel values in the adjacent areas on both sides of the edges is small. Therefore, if only considering supplement details of PAN images for LRMS images without considering the change trend of pixel values in the neighborhoods, which may cause edge blur and spectral distortion.

To better show the change of pixel values between a high-resolution image and its corresponding low-resolution image, we take PAN image and its corresponding degraded image as an example. Figure 1 shows the change of pixel values of HP and LP images on a horizontal line, where HP and LP denote PAN image and its degraded version, respectively. It can be seen from Figure 1 that the intersection points of the two curves are about the positions where the pixel values change most steeply, or the positions where the gradient values are the largest, that is, the edge positions. If the pixel values of LP image are restored to those of HP image, on both sides of the edges, the pixel values of one side need to be raised, and those of another side need to be suppressed. Therefore, to realize the reconstruction from LP images to HP images, the pixel values of these two sides need different processing. Based on the above analysis, this paper proposes an MPS, which decomposes the areas on both sides of the intersection by constructing the mask protection matrices. In this way, the features of two sides need to be learned separately, and the image edges can be protected from being damaged.

In our network, an MPS is proposed to decompose an input image into two mask images, which contain adjacent regions with different pixel changes on both sides of the edges. The structure of MPS is shown in Figure 2. For pansharpening

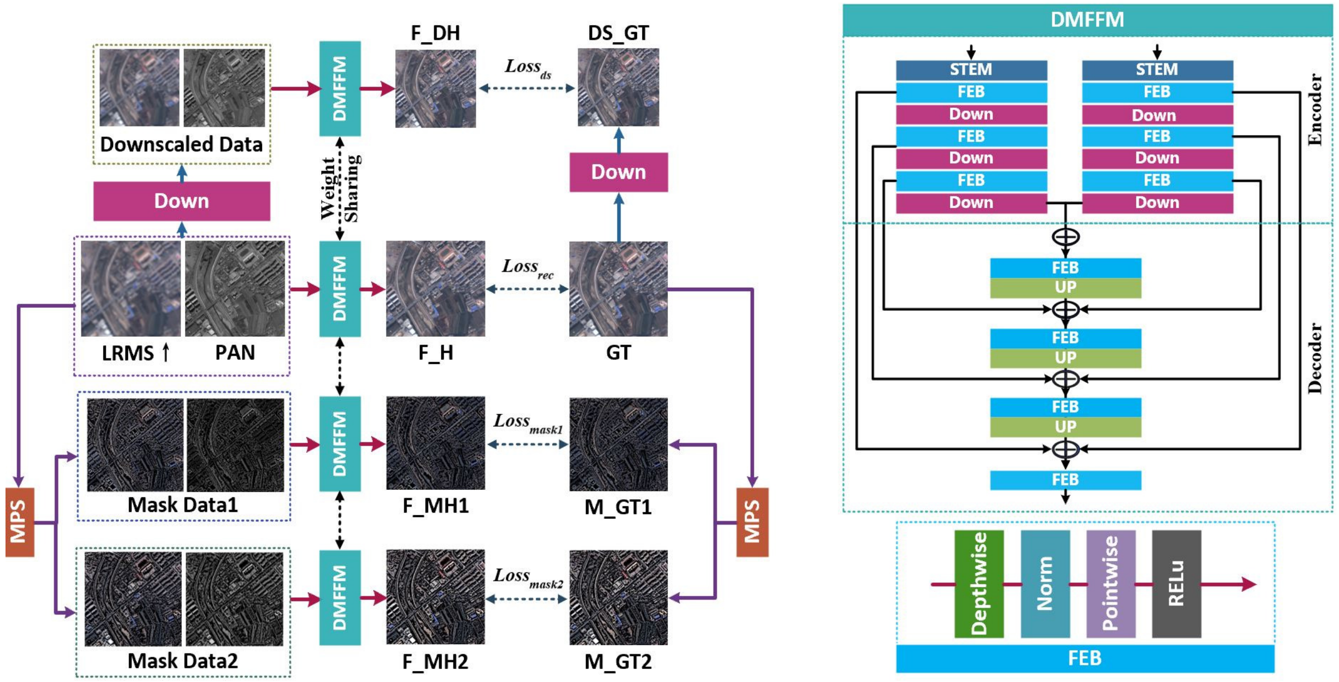


Figure 3: Overall architecture of MMPN. MPS denotes mask protection strategy, and STEM is a convolutional layer.

task, the proposed MPS is to protect the edges of the reconstructed HRMS image from being damaged. The calculation processes of MPS are as follows.

First, a PAN image Y^{HP} is blurred through a modulation transfer function (MTF) [Vivone *et al.*, 2020] to obtain a degraded image with low resolution, named Y^{LP} .

$$Y^{LP} = MTF\left(Y^{HP}\right) \tag{1}$$

For Y^{HP} and Y^{LP} , the pixels with same values are the intersection points. For other pixels, two mask protection matricesMP and $MP_INV$are defined by Formulas 2 and 3. If the pixel values of Y^{HP} are larger than those of Y^{LP} , the element of MP is set to 1, otherwise set 0.

$$MP\left(i,j\right) = \begin{cases} 1, & Y^{HP}\left(i,j\right) \geq Y^{LP}\left(i,j\right) \\ 0, & Y^{LP}\left(i,j\right) < Y^{LP}\left(i,j\right) \end{cases} \tag{2}$$

$$MP\_INV = 1 - MP \tag{3}$$

where (i, j) represent the pixel coordinates. Then, a data filling operation is performed on an input image (i.e., PAN, LRMS or HRMS) and two different mask protection matrices MP and MP_INV to obtain two mask images named $Mask\_data1$ and $Mask\_data2$, which contain complementary information of different areas. The data filling operation is defined as:

$$Mask\_data1 = Y^{In} * MP \tag{4}$$

$$Mask\_data2 = Y^{In} * MP\_INV \tag{5}$$

where $*$ represents an element-wise product.

Finally, the obtained mask data is used as the input of MMPN and to guide the network to separately learn the features of different regions on both sides of image edges.

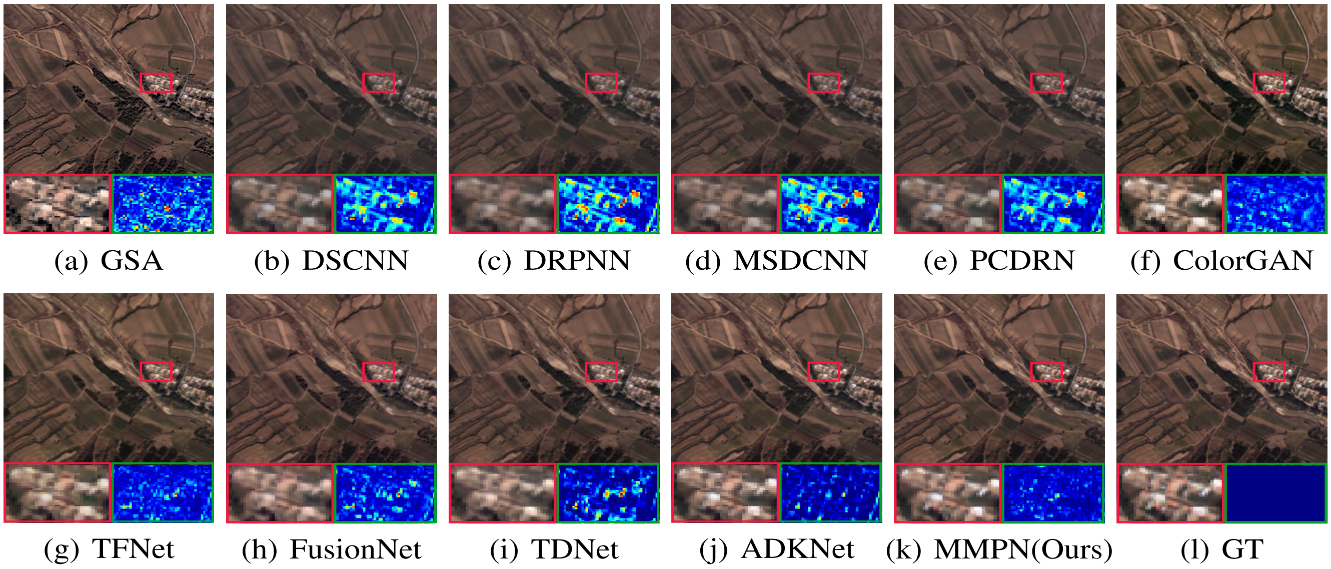
3 Proposed MMPN

In this section, aiming at the problems of edge blur and spectral distortion existing in the current pansharpening methods, an MMPN is proposed, as shown in Figure 3. First, the input source images and the ground truths (GT) are decomposed by the proposed MPS to obtain the mask data and the mask supervised images (M_GT1 and M_GT2), respectively. Then, four DMFFMs with weight sharing are constructed to realize four network branches, which perform the fusion of the source images, the two mask data, and the low-resolution version of the source images, respectively. Finally, a joint loss function containing four loss terms is defined to train the network.

3.1 Dual-stream Multi-scale Feature Fusion Module (DMFFM)

DMFFM is constructed to realize the fusion of two images, and its structure is shown in Figure 3. The DMFFM is designed as a neural network with dual stream inputs and multiscale feature extraction layers to fit the distribution of supervised images. To reduce the confusion of two image features, two encoders with same structure are designed to respectively extract features from two input images, and a decoder is constructed to reconstruct the image features and output the fusion result.

In the encoder, a STEM layer is set at the first layer of DMFFM to extract the initial features. A feature extraction block (FEB) is constructed to extract the features of different scales in the encoders. The structure of FEB is shown in Figure 3, and it is designed as a residual structure, which consists of a depth-wise convolution, a LayerNorm, a pointwise con-



(a) GSA (b) DSCNN (c) DRPNN (d) MSDCNN (e) PCDRN (f) ColorGAN

(g) TFNet (h) FusionNet (i) TDNet (j) ADKNet (k) MMPN(Ours) (l) GT

Figure 4: Fusion results on a simulated data from the IKONOS dataset.

volution, and a ReLU function. Since the MPS used in the network can protect the edge information of the image, the impact of increasing the size of a convolution kernel on the edge can be reduced. Therefore, the convolution kernels with a large receptive field are used to extract features in the spatial dimension. In FEB, the depth-wise convolution with kernel size of 7×7 is adopted to extract features of spatial dimension, and the point convolution is used to integrate channel features and reduce the numbers of feature channels. In the decoder, multiple FEBs are also used to fuse features of different scales from the encoders and output the fused image.

3.2 Joint Loss Function

Since four DMFFMs generate four different fused images, they need to be trained by different supervised images according to different input images. The output of a DMFFM can be defined as follows.

$$Y_{output} = DMFFM(Y_{input\_1}, Y_{input\_1} | W_\theta) \tag{6}$$

where W_θ represents the network parameters, Y_{input_1} and Y_{input_2} represent the different network inputs, and Y_{output} represents the network output. Corresponding to different branches, Y_{output} represents different fused results including the fused HRMS image Y_{F_H} , two fused mask data Y_{F_MH1} and Y_{F_MH2} , and the down-scaled fused image Y_{F_DH} . To guarantee the consistency between the fused results and supervise images, the loss items corresponding to the four branches in MMPN are defined as follows.

$$Loss_{rec} = |Y_{hrms} - Y_{F\_H}| \tag{7}$$

$$Loss_{mask1} = |Y_{M\_GT1} - Y_{F\_MH1}| \tag{8}$$

$$Loss_{mask2} = |Y_{M\_GT2} - Y_{F\_MH2}| \tag{9}$$

$$Loss_{ds} = |Y_{DS\_GT} - Y_{F\_DH}| \tag{10}$$

where $Loss_{rec}$, $Loss_{mask1}$, $Loss_{mask2}$, and $Loss_{ds}$ represent the loss terms corresponding to four branches. Y_{hrms} denotes the GT image, Y_{M_GT1} and Y_{M_GT2} represent mask images obtained by decomposing GT images using MPS, and Y_{DS_GT} denotes the down-scaled GT image. Therefore, the joint loss function used to supervise the MMPN is defined as follows.

$$Loss = Loss_{rec} + Loss_{mask1} + Loss_{mask2} + Loss_{ds} \tag{11}$$

4 Experiments

To validate the effectiveness of the proposed MMPN[1], subjective and objective experiments are conducted on the simulated and real satellite datasets, including IKONOS (4 bands), Pléiades (4 bands), and WorldView-3 (8 bands) In the simulated dataset, the LRMS images are generated according to the WALD protocol [Ghassemian, 2016] by using MTF and down-sampling operations on HRMS images. The sizes of LRMS and PAN images are 64×64 and 256×256, respectively. In real datasets, the sizes of LRMS and PAN images are 256×256 and 1024×1024, respectively.

In the experiments, the performance of the proposed MMPN is compared with those of state-of-the-art methods. The comparison methods include GSA [Aiazzi *et al.*, 2007], DRPNN [Wei *et al.*, 2017], MSDCNN [Yuan *et al.*, 2018], PCDRN [Yang *et al.*, 2020], ColorGAN [Ozcelik *et al.*, 2020], TFNet [Liu *et al.*, 2020], DSCNN [Yang *et al.*, 2021], FusionNet [Deng *et al.*, 2022], ADKNet [Peng *et al.*, 2022], and TDNet [Zhang *et al.*, 2022]. The benchmark images are obtained by the polynomial kernel method (EXP) [Aiazzi *et al.*, 2002]. Significantly, all deep learning-based methods are retrained using the same datasets for fairness, and tested on the environment of NVIDIA GeForce RTX 3090 and INTEL 11700K.

[1] The code is available at github.com/sharpeningNN/MMPN.

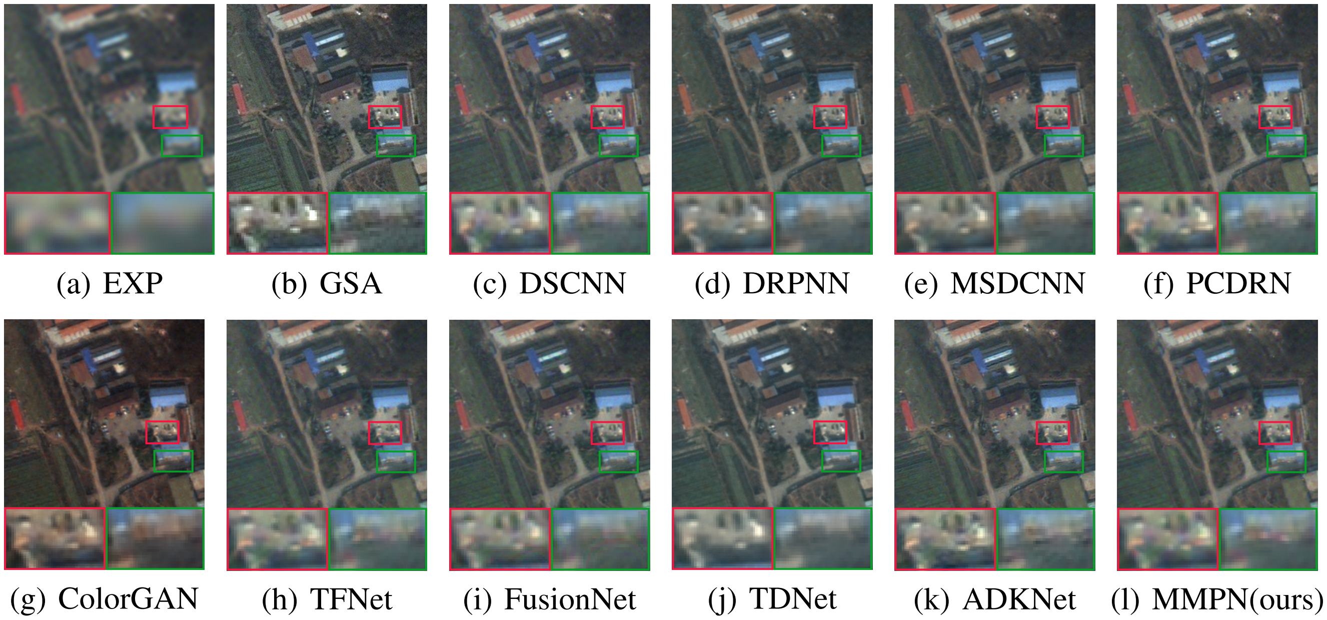
Sensors	Methods	PSNR(↑)	RMSE(↓)	UIQI(↑)	Q_2^n(↑)	SAM(↓)	ERGAS(↓)	Time(s)(↓)
Pléiades	EXP	26.9511	15.7011	0.9257	0.8161	3.1093	3.9761	0.0075
	GSA	27.9760	13.99327	0.9530	0.8757	3.5046	3.5725	0.0283
	DSCNN	32.0029	9.0192	0.9787	0.9109	2.3126	2.4539	0.0273
	DRPNN	30.1216	11.1515	0.9680	0.8636	2.6809	2.9884	0.0424
	MSDCNN	29.2428	12.2007	0.9629	0.8638	2.7117	3.2713	0.0295
	PCDRN	31.5914	9.7448	0.9767	0.9004	2.3488	2.6322	0.0153
	ColorGAN	23.1587	13.1564	0.9320	0.8031	7.8692	8.4645	0.0180
	TFNet	32.8707	8.5569	0.9805	0.9452	2.9289	2.0643	0.0201
	FusionNet	31.3848	10.4311	0.9794	0.9420	2.5486	2.5394	0.0142
	TDNet	31.3226	10.7110	0.9117	0.9164	3.5946	2.6642	0.0124
	ADKNet	35.4673	5.5952	0.9898	0.9648	2.2762	1.4621	0.0353
	MMPN(ours)	**37.2155**	**4.8362**	**0.9930**	**0.9726**	**1.9946**	**1.1946**	0.0464
IKONOS	EXP	26.0662	15.5442	0.8919	0.7523	3.4198	4.1350	0.0079
	GSA	27.0956	14.2779	0.9352	0.8363	4.6886	3.9925	0.0291
	DSCNN	29.0473	10.6432	0.9514	0.8646	3.4841	2.9886	0.0317
	DRPNN	26.9798	13.8740	0.9273	0.8154	4.0383	3.7677	0.0454
	MSDCNN	27.1694	13.5968	0.9310	0.8239	4.0121	3.6587	0.0346
	PCDRN	28.8464	11.3740	0.9516	0.8650	3.5955	3.0352	0.0223
	ColorGAN	24.0976	12.4628	0.9309	0.8224	8.4360	6.0984	0.0332
	TFNet	30.9652	9.1658	0.9702	0.9227	3.2528	2.4465	0.0221
	FusionNet	28.1699	18.4607	0.9561	0.8763	4.4868	4.0201	0.0167
	TDNet	29.6391	9.2735	0.8737	0.8783	3.7885	3.0256	0.0121
	ADKNet	33.3930	6.9367	0.9818	0.9477	2.4315	1.8290	0.0380
	MMPN(ours)	**34.6915**	**5.9038**	**0.9865**	**0.9572**	**2.1394**	**1.5529**	0.0471
WorldView-3	EXP	23.7930	20.9737	0.8540	0.6528	5.5129	0.0149	0.0149
	GSA	27.2705	14.1655	0.9358	0.8688	6.6912	4.3425	0.0459
	DSCNN	28.1630	13.0162	0.9472	0.8707	5.0543	3.9331	0.0325
	DRPNN	25.6762	19.5398	0.9224	0.8152	7.9339	6.1439	0.0572
	MSDCNN	25.2981	19.8476	0.9171	0.8033	7.7049	6.2930	0.0469
	PCDRN	27.8324	13.2502	0.9460	0.8681	5.3713	4.0870	0.0184
	ColorGAN	25.3901	13.9954	0.9149	0.8302	8.3228	5.6929	0.0258
	TFNet	28.2917	12.9021	0.9047	0.8921	5.3725	3.7714	0.0241
	FusionNet	27.2539	15.1852	0.8999	0.8818	5.1426	4.5063	0.0197
	TDNet	30.1461	12.4702	0.8504	0.8936	5.3467	3.7361	0.2931
	ADKNet	31.4306	10.5339	0.9725	0.9336	4.4753	2.7267	0.0564
	MMPN(ours)	**33.3068**	**7.4569**	**0.9808**	**0.9506**	**3.8399**	**2.2621**	0.0761

Table 1: Average quantitative results on the simulated data from Pléiades, IKONOS, and WorldView-3

4.1 Experiments on Simulated Dataset

As shown in Figure 4, the subjective fusion images of all comparison methods on a pair of images from IKONOS dataset can be observed. From the figure, we can find that since the result of EXP is obtained by directly interpolating the LRMS image, it is the most blurred compared with those of other methods due to the lack of spatial details. The results obtained by other deep learning-based methods have the problems of edge blurring and spectral distortion. For example, the results of DSCNN, FusionNet, MSDCNN, DRPNN, PCDRN, TDNet, ADKNet, and TFNet have more blurry edges and lower brightness compared to GT. The result of GSA has serious spectral distortion, and our result is the closest to GT. To show the difference between fusion results and GT more clearly, we calculate the residual maps between them and GT, and show an enlarged local area and its corresponding residual map below each result. From the residual maps, it can be clearly seen that the results of comparison methods have obvious spectral distortion and lose some spatial details. The residual map of our result has the least residual information, which also indicates the effectiveness of the proposed method.

To further compare the performance of each method, the quantitative evaluation results are summarized in Table 1. The evaluation metrics [Deng *et al.*, 2022], including the peak signal to noise ratio (PSNR), the root mean square error (RMSE), the relative average spectral error (RASE), the universal image quality index (UIQI), Q_2^n, the spectral angle mapper (SAM), and the erreur relative globale adimensionnelle de synthèse (ERGAS), are used to objectively evaluate the performance of different methods. For the objective evaluation experiments, the best results are marked with bold


(a) EXP (b) GSA (c) DSCNN (d) DRPNN (e) MSDCNN (f) PCDRN

(g) ColorGAN (h) TFNet (i) FusionNet (j) TDNet (k) ADKNet (l) MMPN(ours)

Figure 5: Fusion results on a real data from the Pléiades dataset.

Methods	$D_\lambda(\downarrow)$	$D_s(\downarrow)$	QNR(↑)	Time($s\downarrow$)	Methods	$D_\lambda(\downarrow)$	$D_s(\downarrow)$	QNR(↑)	Times($s\downarrow$)
EXP	0.0026	0.2055	0.7924	0.0906	EXP	0.0034	0.0933	0.9036	0.1770
GSA	0.1347	0.1619	0.7259	0.4150	GSA	0.0872	0.1195	0.8053	0.5991
DSCNN	0.0242	**0.0263**	0.9501	0.1540	DSCNN	0.0306	0.0426	0.9283	0.1603
DRPNN	0.0533	0.0363	0.9123	0.3083	DRPNN	0.0887	0.0869	0.8341	0.4398
MSDCNN	0.0434	0.0462	0.9123	0.1907	MSDCNN	0.0872	0.0944	0.8287	0.3519
PCDRN	0.0375	0.0382	0.9256	0.0925	PCDRN	0.0602	0.0650	0.8791	0.0996
ColorGAN	0.0434	0.0618	0.8975	0.0334	ColorGAN	0.1126	0.1138	0.7892	0.0509
TFNet	0.0318	0.0310	0.9381	0.0310	TFNet	0.0163	0.0464	0.9206	0.0435
FusionNet	0.0403	0.0285	0.9328	0.0289	FusionNet	0.0625	0.0468	0.8953	0.0361
TDNet	0.1008	0.1006	0.8093	0.0637	TDNet	0.0661	0.0870	0.8534	0.0846
ADKNet	0.0143	0.0371	0.9489	0.1032	ADKNet	0.0260	0.0277	0.9471	0.1121
MMPN(ours)	**0.0100**	0.0379	**0.9525**	0.1962	MMPN(ours)	**0.0154**	**0.0217**	**0.9632**	0.2181

Table 2: Average quantitative results on the real data from Pléiades (4 bands, left) and WorldView-3 (8 bands, right)

fonts, and the second-best results are underlined. In addition, the test time of all methods is given to compare the computational cost of each method. From Table 1, we can see that our method achieves the best results in all three datasets, which means that the proposed MMPN has a better fitting performance than other methods.

In terms of test time, since MMPN uses many convolution kernels with large size, the computational time is increased compared with other deep learning-based methods. However, due to the GPU support, MMPN still can achieve real-time performance.

4.2 Experiments on Real Dataset

Figure 5 shows the fusion results on a pair of images on the Pléiades dataset. To more clearly observe the difference of spectral and detail information between different fusion results, two small regions are selected and amplified. As can be seen from the figure, the result of EXP has the least spatial details. The result of GSA method has the problem of over-injection of details, but lacks of spectral information. Compared with other results, the result of ColorGAN displays a distinct color cast. It can be clearly observed from the enlarged areas that our result has more abundant spectral information than those of other comparison methods.

Due to the absence of GT, some non-reference quantitative metrics including D_λ , D_s , and QNR [Wald, 2000] are used to assess the similarity of spectral and spatial details between fusion images and source images. D_λ measures the spectral similarity between the fusion results and LRMS images, D_s measures the spatial similarity between the fusion results and PAN images, and QNR calculates the overall similarity through D_λ and D_s . As shown in Table 2, the results of EXP method have the lowest D_λ values, which indicates the EXP method maintains the best spectral information.

	Accuracy(↑)		Accuracy(↑)
EXP	0.4589	GSA	0.5294
DSCNN	0.7552	DRPNN	0.5964
MSDCNN	0.6581	PCDRN	0.7516
ColorGAN	0.6000	TFNet	0.7107
FusionNet	0.5488	TDNet	0.6478
ADKNet	0.7410	MMPN(ours)	**0.7739**

Table 3: Classification experiment performance of Figure 4.

	MPS	DB	PSNR(↑)	SAM(↓)	ERGAS(↓)
Model 1	×	×	34.6452	2.2031	1.9614
Model 2	✓	×	**37.2561**	2.0432	1.2590
Model 3	✓	✓	37.2155	**1.9946**	**1.1949**

Table 4: Ablation experiment of different branches on the simulated data from Pléiades.

The results obtained by the traditional GSA method have lower QNR values that those of most comparison methods, which indicates that GSA method obtains worser pansharpening performance. Among the deep learning-based methods, the proposed MMPN has the highest QNR indexes on 4-channels (Pléiades) and 8-channels (WorldView-3) datasets, which also indicates the advanced performance in dual fidelity of spectral and spatial information.

In summary, the proposed MMPN achieves advanced performance in both subjective visual effects and quantitative metrics compared with other methods on different datasets.

4.3 Application Experiment

To evaluate the application property of all comparison methods, image classification experiments are implemented for all fusion results. Referring to the reference [Lu *et al.*, 2021], the ENVI tool is utilized for classification. The GT images in the simulated dataset are first fed to the classification model to obtain classification results, which are regarded as reference images for the classification results of other methods. The classification accuracy value is used to quantitatively evaluate the classification performance of different methods. Table 3 shows the classification results of Figure 4. As shown in Table 3, the proposed MMPN achieves the best classification accuracy than other comparison methods.

4.4 Ablation Experiment

In this section, some ablation experiments are conducted to demonstrate the effectiveness of MPS, downscale branch (DB, top branch in Figure 3), and DMFFM in MMPN. In the experiments of validating MPS and DB, three models containing different components are tested. Model 1 contains only one DMFFM, without MPS and DB. Model 2 contains three branches, using MPS but not DB. Model 3 denotes the proposed MMPN with MPS and DB. Table 4 shows the quantitative indexes obtained by the three models on the simulate dataset. The performance of model 2 is significantly improved compared with that of model 1, which indicates that MPS effectively protects the spatial features from being corrupted. The PSNR value obtained by model 3 is slightly lower than that of model 2, but the difference is not significant. This is because DB can be regarded as a data enhancement, which can increase the generalization of the network, but leads to a slight reduction in the fitting effect for a certain distribution. To verify that DB can improve the generalization of the network, the model 2 without DB is compared with model 3 on the real dataset , as shown in Table 5. It can be seen from Table 5 that the indicators obtained by model 3 with DB are all better than that of model 2. Therefore, the performance of DB is verified.

	$D_\lambda(\downarrow)$	$D_s(\downarrow)$	QNR(↑)
w/o DB (Model 2)	0.0158	0.0409	0.9476
w/ DB (Model 3)	**0.0100**	**0.0379**	**0.9525**

Table 5: Ablation experiment of different branches on the real data from Pléiades.

	PSNR(↑)	SAM(↓)	ERGAS(↓)
ConvBlock	30.1345	3.5217	2.8361
ResBlock	32.9013	2.9013	2.0312
FEB	**34.6452**	**2.2031**	**1.9614**

Table 6: Ablation experiment of different network structure in DMFFM on the simulated data from Pléiades.

In DMFFM, FEB plays a key role. To prove the performance of FEB in DMFFM, the ablation experiments using two network structures to replace FEB are performed. The first structure uses convolutional block (ConvBlock) to replace FEB, and another structure uses a residual block (ResBlock) to replace FEB. As shown in Table 6, we can find that the DMFFM with FEB obtains the best performance than other network structures, which indicate the effectiveness of FEB.

5 Conclusion

In this paper, firstly, based on the analysis of the gray value changes between the high-resolution image and the corresponding degraded image, an MPS is proposed to protect the spatial information from being destroyed in the feature learning of the network. Then, based on this strategy, an MMPN containing four branches with the same structure is proposed to realize the learning of multiple fusion tasks, and each task is realized through the constructed DMFFM. Finally, different loss items are defined for the four task branches and combined into a joint loss function for network training. Experimental results show that our method achieves better performance than other advanced methods in terms of subjective vision and objective indicators.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62072218, No. 61862030, and No. 62261025) and the Project of the Education Department of Jiangxi Province under Grant YC2021-B131.

Contribution Statement

Changjie Chen and Shuying Huang contributed equally to this work. Corresponding author: Yong Yang.

References

[Aiazzi *et al.*, 2002] Bruno Aiazzi, Luciano Alparone, Stefano Baronti, and Andrea Garzelli. Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis. *IEEE Transactions on geoscience and remote sensing*, 40(10):2300–2312, 2002.

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

[Deng *et al.*, 2022] Liangjian Deng, Gemine Vivone, Mercedes E Paoletti, Giuseppe Scarpa, Jiang He, Yongjun Zhang, Jocelyn Chanussot, and Antonio Plaza. Machine learning in pansharpening: A benchmark, from shallow to deep networks. *IEEE Geoscience and Remote Sensing Magazine*, 10(3):279–315, 2022.

[Ghassemian, 2016] Hassan Ghassemian. A review of remote sensing image fusion methods. *Information Fusion*, 32:75–89, 2016.

[Liu *et al.*, 2020] Xiangyu Liu, Qingjie Liu, and Yunhong Wang. Remote sensing image fusion based on two-stream fusion network. *Information Fusion*, 55:1–15, 2020.

[Lu *et al.*, 2021] Hangyuan Lu, Yong Yang, Shuying Huang, Wei Tu, and Weiguo Wan. A unified pansharpening model based on band-adaptive gradient and detail correction. *IEEE Transactions on Image Processing*, 31:918–933, 2021.

[Meng *et al.*, 2022] Xiangchao Meng, Nan Wang, Feng Shao, and Shutao Li. Vision transformer for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022.

[Ozcelik *et al.*, 2020] Furkan Ozcelik, Ugur Alganci, Elif Sertel, and Gozde Unal. Rethinking cnn-based pansharpening: Guided colorization of panchromatic images via gans. *IEEE Transactions on Geoscience and Remote Sensing*, 59(4):3486–3501, 2020.

[Peng *et al.*, 2022] Siran Peng, Liang-Jian Deng, Jin-Fan Hu, , and Yuwei Zhuo. Source-adaptive discriminative kernels based network for remote sensing pansharpening. *International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 1283–1289, 2022.

[Tu *et al.*, 2001] Teming Tu, Shunchi Su, Hsuenchyun Shyu, and Ping S Huang. A new look at ihs-like image fusion methods. *Information fusion*, 2(3):177–186, 2001.

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

[Wald, 2000] Lucien Wald. Quality of high resolution synthesised images: Is there a simple criterion? In *Third conference" Fusion of Earth data: merging point measurements, raster maps and remotely sensed images"*, pages 99–103. SEE/URISCA, 2000.

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

[Xu *et al.*, 2014] Qizhi Xu, Bo Li, Yun Zhang, and Lin Ding. High-fidelity component substitution pansharpening by the fitting of substitution data. *IEEE Transactions on Geoscience and Remote Sensing*, 52(11):7380–7392, 2014.

[Yang *et al.*, 2020] Yong Yang, Wei Tu, Shuying Huang, and Hangyuan Lu. Pcdrn: Progressive cascade deep residual network for pansharpening. *Remote Sensing*, 12(4):676, 2020.

[Yang *et al.*, 2021] Yong Yang, Wei Tu, Shuying Huang, Hangyuan Lu, Weiguo Wan, and Lixin Gan. Dual-stream convolutional neural network with residual information enhancement for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021.

[Yuan *et al.*, 2018] Qiangqiang Yuan, Yancong Wei, Xiangchao Meng, Huanfeng Shen, and Liangpei Zhang. A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(3):978–989, 2018.

[Zhang *et al.*, 2022] Tianjiang Zhang, Liangjian Deng, Tingzhu Huang, Jocelyn Chanussot, and Gemine Vivone. A triple-double convolutional neural network for panchromatic sharpening. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.