Unsupervised Attention-Based Generative Adversarial Network for Remote Sensing Image Fusion

Quanli Wang[1, 2], Qian Jiang[1, 2 [20000-0003-3097-0721]], Yuting Feng[1, 2], Shengfa Miao[1, 2], Huangqimei Zheng[1, 2], and Xin Jin[1, 2, *[0000-0003-2211-2006]]

[1] Engineering Research Center of Cyberspace, Yunnan University, Kunming 650000, China
[2] School of Software, Yunnan University, Kunming 650000, China
[*] Corresponding Author
xinxin_jin@163.com;

Abstract. Remote sensing image fusion combines single-band panchromatic (PAN) image with multi-spectral (MS) image to generate high quality fused image, also known as pan-sharpening. Most of the current methods suitable for remote sensing image fusion are supervised, which require proportional down-sampling of the original multi-spectral image as training image, and the original multi-spectral image as label image. This will result in poor performance of the model on full resolution images, so the unsupervised methods are more practical. Furthermore, most methods do not consider the differences between MS and PAN images and use the same modules to extract features, which results in some information loss. Therefore, we design an unsupervised attention-based generative adversarial network fusion framework (UAB-GAN), which can be trained directly on the datasets of unlabeled images. Specifically, the model framework consists of a generator and two discriminators. The generator employs different network modules with specific designs to extract unique modal features from PAN and MS images, respectively. Then two discriminators are designed to preserve the spectral and spatial information of different images. Additionally, we propose a unified loss function to integrate multi-scale spectral and spatial features without external data supervision. The effectiveness of the proposed method is demonstrated through experiments conducted on various datasets.

Keywords: Image Fusion, Generative Adversarial Network (GAN), Unsupervised Method, Remote Sensing Image.

1 Introduction

Generally speaking, remote sensing images usually come from multiple different sensors or platforms, and the data types include panchromatic (PAN) images, multi-spectral (MS) images, and hyper-spectral (HS) images. PAN images cover very rich spatial information and can record more texture and detailed information on the surface of objects, providing more accurate data for the identification and analysis of ground objects. MS images have greater advantages in terms of spectrum, because they contain data in multiple bands and can provide rich spectral information. They can be fused into a high-quality image through a series of algorithms and technologies to obtain richer and more accurate ground object information and feature information. And these algorithms that fuse two remotely sensed images are called pan-sharpening techniques. This fusion technology is widely used in fields such as earth observation, resource management, and environmental monitoring. These algorithms and techniques can be broadly divided into conventional and deep learning-based methods. Conventional fusion methods include component substitution-based (CS) image fusion methods[1][2][3], multi-resolution analysis-based (MRA) image fusion methods[4][5], and model-based image fusion methods.

Among CS-based methods, commonly used transformations include intensity-hue-saturation technique (IHS)[1], principal component analysis (PCA)[2], and Gram-Schmidt (GS)[3]. Among them, the IHS technology decomposes the image into three components: intensity, hue, and saturation, and then fuses the intensity information of the panchromatic image with the hue and saturation information of the multi-spectral image. PCA performs principal component transformation on the multi-spectral image, extracts the main features, and then fuses the panchromatic image with the principal components. The GS method replaces the corresponding bands of the multi-spectral image one by one with the bands of the panchromatic image to achieve image fusion.

MRA methods are widely used in the field of panchromatic and multi-spectral image fusion due to their ability to effectively preserve spectral information[4]. In MRA methods, the two types of images are converted to scale levels through specially designed conversion functions. This method can organically combine the spatial and spectral information of PAN images and MS images, thereby achieving comprehensive enhancement and optimization of remote sensing images. Common MRA methods include high-pass filter (HPF)[5], indusion[6] and à trous wavelet transform (ATWT)[7]. The HPF method enhances the spatial details of the image through filter design, thereby improving the resolution of the PAN image. The indusion method achieves adaptive fusion of PAN images and MS images by introducing weight parameters. The ATWT method uses the multi-scale characteristics of wavelet analysis to achieve multi-scale decomposition and fusion of images, and has good time-frequency locality. However, unreasonable models or parameters may cause the fused image to lose a lot of spatial information, thus affecting the fusion effect and quality.

Model-based image processing method refers to the use of mathematical models or physical models to describe the characteristics and attributes of images, and through the analysis and processing of these models to achieve image processing and enhancement[8]. The entire process of the model-based remote sensing image fusion method is as follows. First, interpretable mathematical models between input panchromatic images, multi-spectral images, and label images must be established. These models can describe the correlation and feature mapping relationships between different images. Then, the established model is parameterized by solving the optimization problem to maximize the quality and feature information of the fused image. A typical method is the correlated spatial detail (BDSD) model[9], which considers the correlation between PAN and MS images, thereby achieving efficient fusion of images.

However, these methods have some limitations. One of the major issues is their limited ability to resolve highly nonlinear mappings. Since the mapping relationship between PAN and MS images is usually non-linear, these methods may suffer from spectral distortion and loss of spatial details when handling complex fusion tasks.

2 Related Work

In recent years, some new image fusion methods based on deep learning have emerged[10][11][12]. These methods utilize techniques such as convolutional neural networks to better capture the complex relationships between images, thereby achieving higher quality image fusion. Deep learning is gradually recognized by researchers in the field of image processing, has become a mainstream method in many image processing tasks, and has achieved many breakthrough results in various fields. Feng et al.[13] proposed a multi-scale feature injection network for panchromatic sharpening, and achieved good results by using dynamic convolution modules and multi-level fusion modules. Wang et al.[14] proposed a new hybrid network for use in the field of remote sensing image fusion, and performed well in experiments on several datasets. Likewise, Transformer is suitable for PAN and MS image fusion. Liu et al.[15] designed a local moving window mechanism that can flexibly model images of different scales, which contributes to reducing the amount of computation.

Most of these methods are based on supervised learning. To this end, the original PAN and MS images are usually down-sampled into low-resolution image pairs at a certain proportion, and then the low-resolution image (LRMS) pairs are used as input to the network model, and the original MS images are used Labeled images (GT) are used for training. This method is also called the Wald protocol[16]. However, due to the use of down-sampled low-resolution image pairs during the training process, when the trained network is tested on a high-resolution test set, the test results will usually be significantly different. Therefore, the unsupervised remote sensing image fusion methods have more practical significance in specific applications. Ni et al.[17] proposed an unsupervised fusion model based on a learnable degradation process. ZeRGAN[18] is a novel unsupervised method that does not require pre-training. Zhou et al.[19] proposed cycle-consistent unsupervised generative adversarial networks for pan-sharpening.

In addition, most methods do not consider the differences between MS and PAN images and use the same modules to extract features, which results in some information loss. Based on the above reasons, this work proposes an unsupervised attention-based generative adversarial network, which can be trained directly on a dataset of unlabeled images. First, our generator designs two different convolutional attention modules for different types of input images, namely the convolutional block attention module (CBAM)[20] and the panchromatic convolutional block attention module (CBAM-P). These modules include both channel attention and spatial attention. Secondly, the addition of skip connections can prevent gradient disappearance and ensure the stability of the model training process. Then, a dual discriminator design is adopted to focus on the spectral and spatial information of different images. Finally, the algorithm uses multiple loss functions to ensure that the network can generate high-quality fused images (HRMS). The main contributions of this paper are as follows.

1) We design an unsupervised framework that can learn directly on a dataset of unlabeled images. This model can generate high-quality fused images without downsampling the original image to obtain labels.

2) Considering that PAN and MS images are rich in different information, we innovatively designed a differentiated dual-stream feature extraction module based on multiple attention. And use dual discriminators to improve the information retention ability of the network.

3) We introduce a unified non-reference loss function to ensure that the network generates high-quality fused images. Experiments show that our method has better results compared with common conventional methods and unsupervised methods.

The remaining chapters of this article are organized as follows. Section 3 describes the proposed UAB-GAN method in detail. Section 4 describes the experimental results of different methods and performs ablation experiments. The entire article is summarized in Section 5.

3 Proposed Method

Based on existing unsupervised fusion methods and research field development trends, this paper designs an overall architecture based on generative adversarial networks. The overall model includes a generator module based on differential feature extraction and two discriminator modules to maintain the spectral and spatial characteristics of the generated image. The specific structure of the model is shown in Figure 1. First of all, since PAN and MS images have different focuses of information, it is feasible to design specific feature extraction modules for different images. This method is conducive to obtaining feature information for subsequent operations. Then, the feature integration and restoration stage uses a CBAM with channel attention and spatial attention. This channel-spatial dimension integration method is more conducive to image reconstruction. In addition, the up-sampled original MS image is connected to the final output in a skip layer connection. This can maintain the stability of the training process and improve fused image quality. Finally, the design idea of the dual discriminator comes from the Pan-GAN network[21]. Based on the designed loss function, the two discriminators focus on the spectral and spatial information of the image.

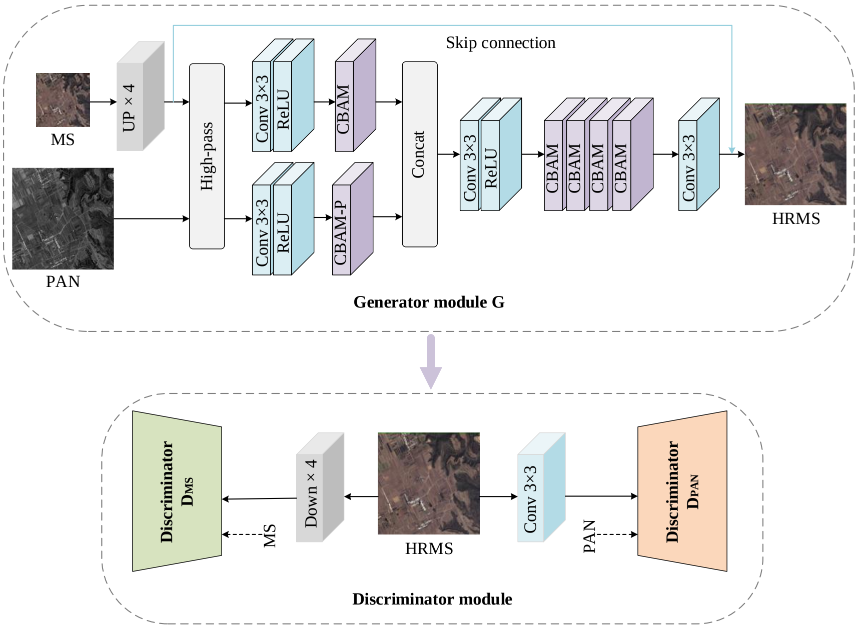


Fig. 1. Structural diagram of UAB-GAN network.

3.1 Generator

The structure of the generator is shown in Figure 1. Specifically, the generator G can be divided into a dual-stream feature extraction stage and a feature integration and image restoration stage. According to the research of PanNet[22], training the generator in the high-frequency domain after high-pass filter (HPF) processing is beneficial to preserving the spatial characteristics of the image. In the dual-stream feature extraction stage, there is a convolution layer containing a ReLU activation function, and the size of the convolution kernel is uniformly set to 3×3. The study found that spectral information and spatial information exist in both MS and PAN images. In addition, considering the richness of spectral information of MS images, the CBAM is applied in the algorithm for feature extraction of MS images. However, the spatial information of PAN images is more comprehensive. Based on this thinking, the CBAM-P was designed for information extraction of PAN images. The specific designs of the two modules are shown in Figure 2.

Specifically, there are two kinds of attention in both CBAM and CBAM-P, namely channel attention and spatial attention, but there are differences in the specific order. CBAM uses the original channel-spatial attention arrangement, while CBAM-P adopts a new spatial-channel attention arrangement to enhance the extraction ability of spatial features. In addition, both modules add residual connections, which can effectively prevent gradient disappearance.

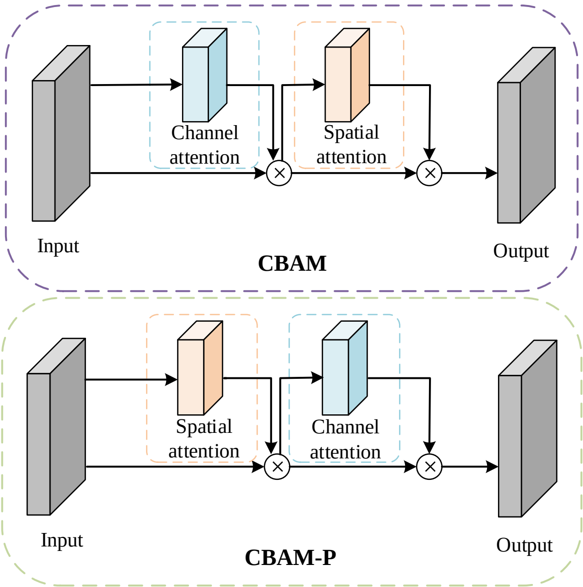


Fig. 2. Convolutional Block Attention Module (CBAM) and Panchromatic Convolutional Block Attention Module (CBAM-P)

In the channel attention module, the input features are pooled simultaneously by two types of pooling, namely average pooling (AvgPool) and maximum pooling (MaxPool), which will obtain two feature maps of specific sizes. Then the two feature maps pass through the multi-layer perceptron (MLP) respectively, and then the result vectors are added. Then the weights are mapped to a specific range through the activation function, and finally multiplied with the original features in the channel dimension to obtain the output feature map. The specific process is shown in Formula 1.

$$H_C(I) = S(MLP(AvgPool(I)) + MLP(MaxPool(I))) \tag{1}$$

where *I* represents the input feature and *S* represents the activation function.

The design idea of spatial attention is generally similar. The input features are first subjected to two pooling operations, and the obtained feature maps are spliced in the channel dimension. Then spatial weights are generated through convolution and activation functions, and finally the processed feature maps are obtained. The specific process is shown in Formula 2. And figure 3 shows two attention modules.

$$H_S(I) = S(Conv7\times7([AvgPool(I); MaxPool(I)])) \tag{2}$$

where *I* represents the input feature. *S* represents the activation function. *Conv*7×7 represents a convolution operation with a convolution kernel size of 7×7.

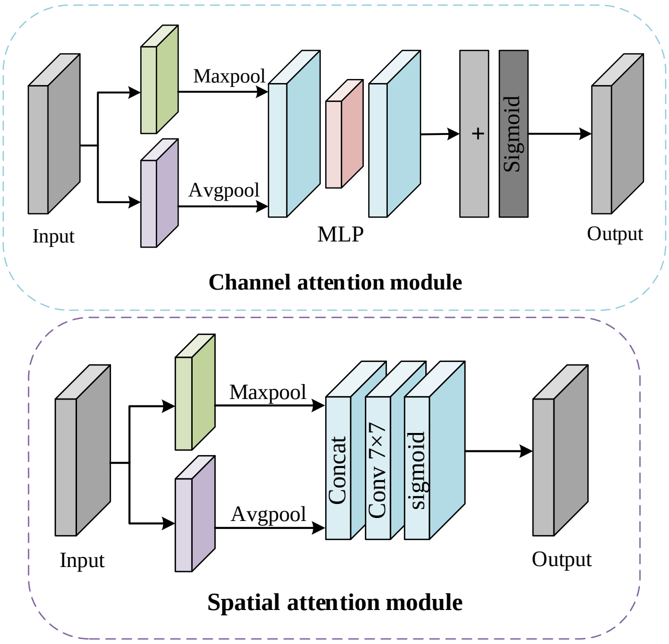


Fig. 3. Channel attention module and spatial attention module

3.2 Discriminator

The model uses two discriminators for training, which can effectively improve the retention ability of spatial and spectral information. The specific structure is shown in Figure 4. The discriminator that takes input MS image pairs is called D_{MS}, and the other discriminator that takes input PAN image pairs is called D_{PAN}. The design principle of the discriminator is similar to that of pix2pix[23], consisting of different numbers of convolutional layers, and the size of the convolutional kernel is 4×4. The input image pairs in D_{MS} are the original MS image and the down-sampled HRMS image, respectively, and the spectral information is balanced through dynamic training. The input image pairs in DPAN are the original PAN image and the converted PAN image, which retain spatial information through dynamic training. The converted PAN image is obtained from the HRMS image after passing through a 3×3 convolution layer. At the same time, a deeper network structure is designed for larger PAN images. The discriminator adopts a fully convolutional layer design, which can make the training process more stable and show better tolerance to test images of different sizes.

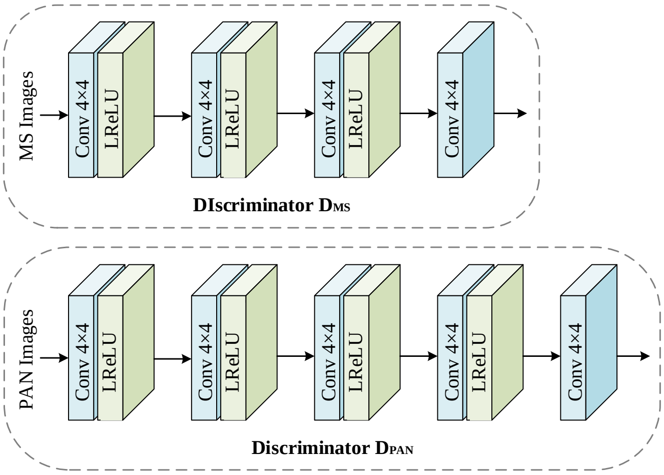


Fig. 4. Structure diagram of two discriminators

3.3 Loss Function

The overall loss function of UAB-GAN consists of two parts, namely the loss of the generator G, and the loss of the discriminators D_{MS} and D_{PAN}. The specific composition of the loss function will be described below.

1) QNR loss: QNR loss comes from a non-reference quality index, which includes spectral information component D_{λ}, spatial information component Ds and non-reference index QNR. The QNR index is used by PGMAN in the construction of the non-reference loss function to minimize the information loss of the generated image[24]. This article follows the same design principle, and the specific design is shown in Equation 3.

$$L_Q = 1 - QNR \tag{3}$$

The calculation process of QNR is shown in Equation 4.

$$QNR = (1 - D_{\lambda})(1 - D_S) \tag{4}$$

The processes of spectral information component D_{λ} and spatial information component Ds are shown in Equations 5 and 6.

$$D_{\lambda} = \sqrt{\frac{2}{N(N-2)} \sum_{x=1}^{N} \sum_{y=1}^{N} \left| Q(K_x, K_y) - Q(L_x, L_y) \right|} \tag{5}$$

$$D_S = \sqrt{\frac{1}{N} \sum_{x=1}^{N} \left| Q(K_x, P) - Q(L_x, P') \right|} \tag{6}$$

where *K* represents the fused HRMS image. *L* represents the original MS image. *P* represents the original PAN image. *P′* represents the image after HRMS spectral degradation. *N* is the number of bands. *x* and *y* represent the current frequency band. *Q* represents the image quality index.

2) Adversarial loss: The focuses of the discriminators D_{MS} and D_{PAN} are different. D_{MS} focuses on the retention of spectral information, while D_{PAN} focuses on the retention of spatial information. During the dynamic training process, the information preservation ability of the generator is enhanced. To ensure that the fusion image as close to the real thing as possible can be generated. The loss function of generator G is shown in Equation 7.

$$L_G = \frac{1}{M}\sum_{m}^{M} -\alpha D_{MS}(K'^{(m)}) - \beta D_{PAN}(P'^{(m)}) + L_Q(K^{(m)}, L^{(m)}, P^{(m)}) \tag{7}$$

where *M* is the sample size. *K* represents the fused HRMS image. *K′* represents the image after HRMS down-sampling. *α* and *β* are hyper-parameters.

The loss functions of DMS and DPAN are shown in Equations 8 and 9.

$$L_{D_{MS}} = \frac{1}{M}\sum_{m=1}^{M}\left\|K'^{(m)} - L^{(m)}\right\|_F^2 \tag{8}$$

$$L_{D_{PAN}} = \frac{1}{M}\sum_{m=1}^{M}\left\|P'^{(m)} - P^{(m)}\right\|_F^2 \tag{9}$$

where $\| \ \|_F$ represents the Frobenius norm matrix operation.

4 Experiments and Results

4.1 Experiment Details

Experiments are coed on two datasets to evaluate the superiority of the proposed method, including WorldView II (WV2) and QuickBird (QB). The divisions of training data, validation data and test data of the two data sets are 1135/59/59 and 451/22/22 respectively. Validation data is used for full-resolution experimental evaluation, and test data is used for low-resolution experimental evaluation. When constructing the training set, the Wald protocol [16]is used to generate training image pairs, and the original MS image is cropped into image patches of 256 × 256 pixels and used as a reference image (GT). The corresponding PAN image size is 256 × 256, and the LRMS image size is 64×64.

We selected eight comparative methods to verify the excellence of the proposed method, including four classic fusion methods and four representative unsupervised methods based on deep learning. The four classic fusion methods are Brovey[25], MTF_GLP_HPM[26], PCA[2] and SFIM[27]. Four representative deep learning-based unsupervised methods include PANGAN[21], LDPNet[17], ZeRGAN[18] and

UCGAN[19].In addition, we selected three commonly used non-reference quality indicators D_{λ}, *Ds* and QNR for full resolution quality evaluation[28]. The following reference indicators are selected: peak signal-to-noise ratio (PSNR)[29], relative global error (ERGAS)[30], correlation coefficient (CC)[31], spectral angle mapping (SAM)[32] and universal image quality index (UIQI)[33].

Our UAB-GAN is implemented in PyTorch version 1.4.0, and the corresponding GPU is NVIDIA GeForce GTX 1080Ti -11GB. The batch size was set to 8 in order to produce optimal experimental results. The algorithm uses the Adam optimizer to optimize the loss during model training, and the learning rate is set to 0.0001.

Table 1. Reference numerical results on the QB dataset.

Method	PSNR↑	ERGAS↓	CC↑	SAM↓	UIQI↑
Brovey[25]	26.6485	5.5547	0.7795	0.0509	0.6949
MTF_GLP_HPM[26]	24.9313	6.0709	0.8577	0.0536	0.7330
PCA[2]	24.7955	6.8604	0.6974	0.1064	0.6337
SFIM[27]	26.5900	5.2108	0.8462	**0.0498**	0.7235
LDPNet[17]	23.6202	6.7961	0.7639	0.1233	0.6486
UCGAN[19]	28.0187	4.2700	0.8938	0.0634	0.7895
PANGAN[21]	27.2667	4.7514	0.8670	0.0519	0.7502
ZeRGAN[18]	24.4477	6.8690	0.7561	0.0922	0.6462
UAB-GAN	**28.3682**	**4.2261**	**0.9075**	0.0978	**0.8114**
Ideal	+∞	0	1	0	1

Table 2. Non-reference numerical results on the QB dataset.

Method	Dλ↓	DS↓	QNR↑
Brovey[25]	0.0230	0.2659	0.7174
MTF_GLP_HPM[26]	0.0238	0.3562	0.6288
PCA[2]	0.0232	0.2652	0.7180
SFIM[27]	0.0230	0.2689	0.7146
LDPNet[17]	0.0217	0.2743	0.7100
UCGAN[19]	0.0233	0.2553	0.7276
PANGAN[21]	0.0236	0.2854	0.6980
ZeRGAN[18]	0.0234	0.2516	0.7310
UAB-GAN	**0.0214**	**0.2501**	**0.7340**
Ideal	0	0	1

4.2 Experimental Results

The numerical results of all methods with and without reference on the QB data set are shown in Table 1 and Table 2 respectively. The data reflects that the UAB-GAN method achieved good results in both test experiments. The best value is marked in bold and the next best value is underlined. Specifically, the UAB-GAN method has advantages in all non-reference indicators and most reference indicators, indicating that the generated fusion image is rich in information features and has little difference from the features of the label image. Since some conventional methods place more emphasis on the calculation weight of image pixels, making the spectral characteristics of the generated image more obvious, the UAB-GAN method is slightly lower than the SFIM method and the Brovey method in terms of SAM index.

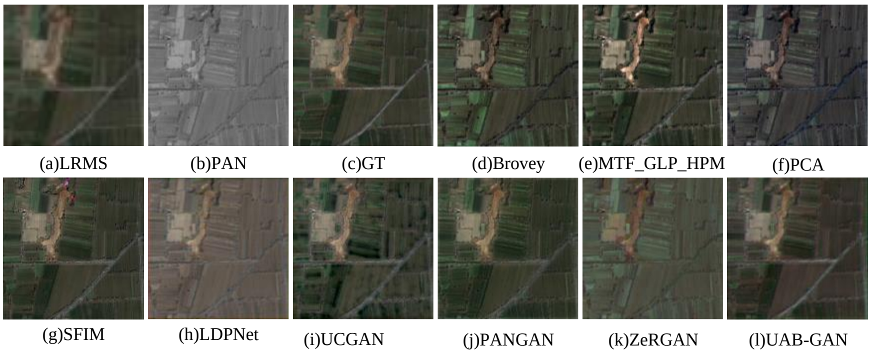


Fig. 5. Fusion results of different methods on the QB dataset.

The fusion results on the QB dataset are shown in Figure 5, where 5(a) is the up-sampled LRMS image, 5(b) and 5(c) are the PAN and original MS images respectively. There are obvious spatial feature differences in the result maps of the Brovey method and the PCA method. The fused images of the UCGAN and PANGAN methods have spatial distortion problems. Although this problem is less severe with other methods, there is still a bias in spectral information. For example, light spots appear in the result image of the SFIM method. The color information of the image generated by the LDPNet method and the ZeRGAN method is quite different from the original image, which is specifically reflected in the color difference of fields and roads. The UAB-GAN method has the best spatial and color information recovery effect.

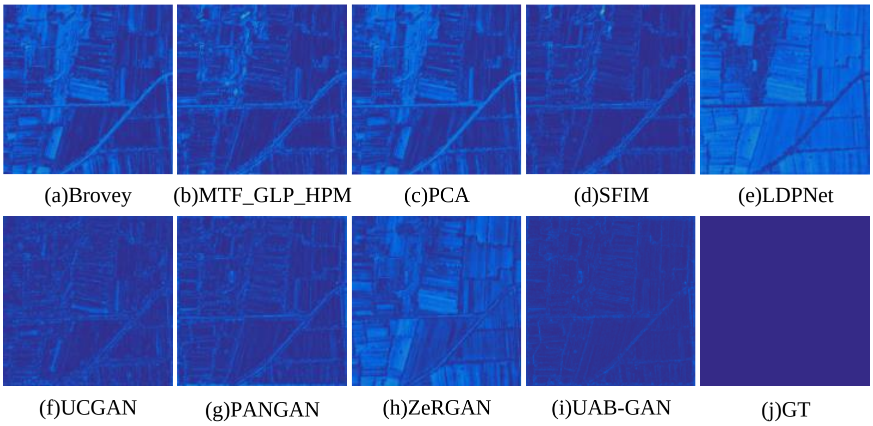
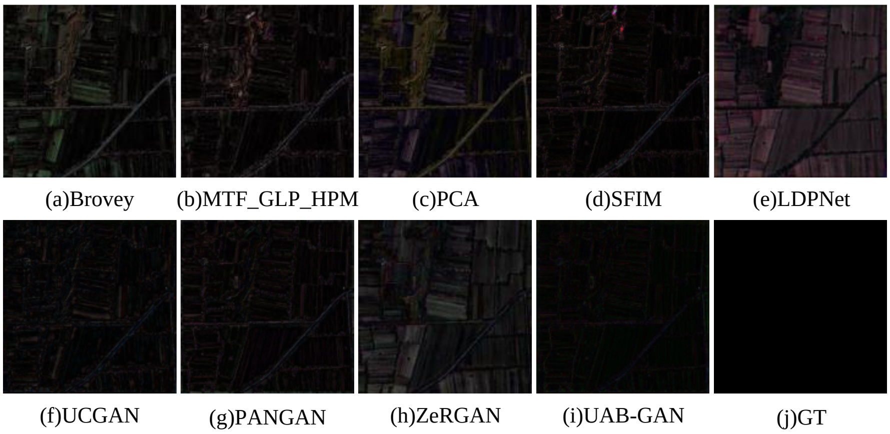


Fig. 6. Spatial error maps of different methods on the QB dataset.



(a)Brovey (b)MTF_GLP_HPM (c)PCA (d)SFIM (e)LDPNet

(f)UCGAN (g)PANGAN (h)ZeRGAN (i)UAB-GAN (j)GT

Fig. 7. Spectral error maps of different methods on the QB dataset.

In order to intuitively demonstrate the spectral and spatial restoration effects of different fused images, the differences of the different images were visualized. The spatial error map is mapped into blue space, and the spectral error map is in black. Correspondingly, the smaller the difference from the original image, the closer the error image is to a pure color image. The contours in Figure 6(a), (c), (e), and (h) are more obvious, indicating that the spatial recovery effect is poor. The difference between Figure 6(f) and (i) is small, indicating that the spatial features are well preserved. In terms of spectral error, the conclusions obtained are basically consistent with the previous ones. The error figure 7(i) of UAB-GAN is closer to black, indicating that this method best retains spectral information.

Table 3. Reference numerical results on the WV2 dataset.

Method	PSNR↑	ERGAS↓	CC↑	SAM↓	UIQI↑
Brovey[25]	28.0123	7.8024	0.8377	**0.1066**	0.7111
MTF_GLP_HPM[26]	25.5589	9.8919	0.8662	0.1218	0.7202
PCA[2]	27.0186	8.4640	0.8375	0.2172	0.7043
SFIM[27]	27.6404	8.2301	0.8444	0.1107	0.7051
LDPNet[17]	25.6351	10.7552	0.8355	0.2341	0.6755
UCGAN[19]	28.4956	6.9677	0.8902	0.1503	0.7510
PANGAN[21]	25.8711	8.2515	0.8569	0.1240	0.6763
ZeRGAN[18]	26.2008	9.3745	0.8076	0.1718	0.6598
UAB-GAN	**29.6372**	**6.9228**	**0.9265**	0.1831	**0.8026**
Ideal	+∞	0	1	0	1

Table 4. Non-reference numerical results on the WV2 dataset.

Method	Dλ↓	DS↓	QNR↑
Brovey[25]	0.0131	0.2164	0.7737
MTF_GLP_HPM[26]	0.0143	0.3216	0.6693
PCA[2]	0.0134	0.2122	0.7776
SFIM[27]	0.0132	0.2042	0.7858
LDPNet[17]	0.0143	0.2527	0.7369
UCGAN[19]	0.0137	0.2437	0.7461
PANGAN[21]	0.0142	0.2296	0.7598
ZeRGAN[18]	0.0134	0.2209	0.7689
UAB-GAN	**0.0119**	**0.1858**	**0.8048**
Ideal	0	0	1

Similarly, experiments conducted on the WV2 data set also verified the superiority of the proposed UAB-GAN method, and various numerical results are shown in Table 3 and Table 4. From the result analysis, the UAB-GAN method has the best effect, followed by the UCGAN method, which also shows that the unsupervised method based on the generative adversarial network has a broad application space. Specifically, the PSNR of the UAB-GAN method is numerically higher than the UCGAN method by 1.1416dB. It is better than the UCGAN method on UIQI by 0.0516. Compared with other conventional methods, the advantages are more prominent.

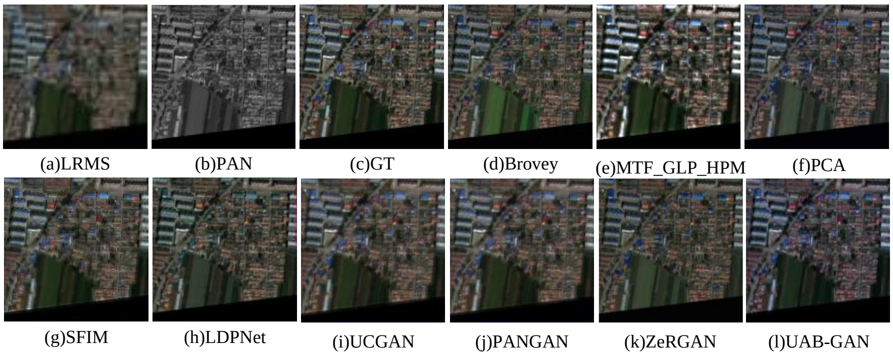
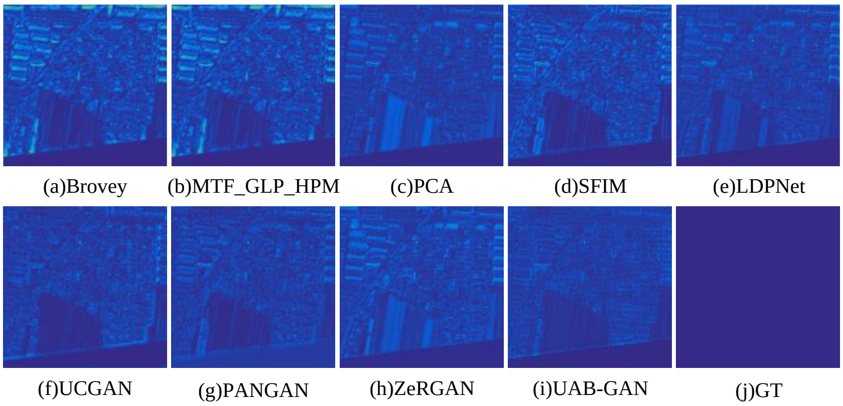

(a)LRMS (b)PAN (c)GT (d)Brovey (e)MTF_GLP_HPM (f)PCA
(g)SFIM (h)LDPNet (i)UCGAN (j)PANGAN (k)ZeRGAN (l)UAB-GAN

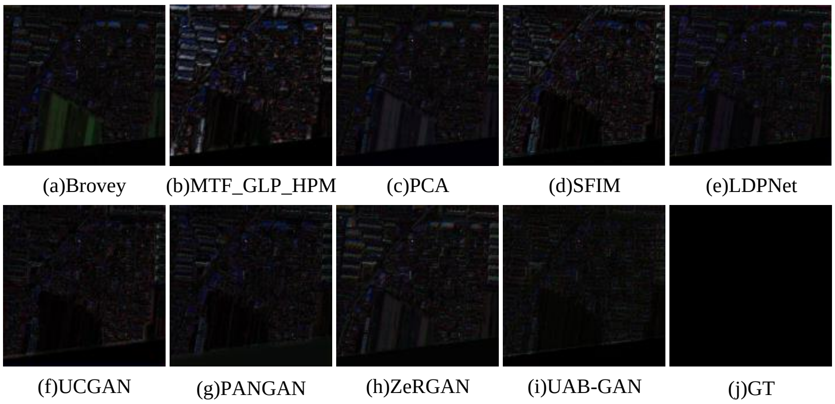
Fig. 8. Fusion results of different methods on the WV2 dataset.

Figure 8 shows some test results of WV2 images obtained by different algorithms. The images of the LDPNet and ZeRGAN methods cannot retain the color information in the original image well, and obvious color distortion appears in the test image. The grass color is brighter in the restored image by Brovey method, but it is different from the original image. Fusion images from other methods have problems with spatial details, manifesting as architectural distortion and blurring. As shown in Figure 8(l), the architectural color of fused image is closer to the original image, and the spatial details are well performed. Comprehensive analysis shows that UCGAN performs satisfactorily on the WV2 data set, and the image is well restored in terms of space and color.


(a)Brovey (b)MTF_GLP_HPM (c)PCA (d)SFIM (e)LDPNet

(f)UCGAN (g)PANGAN (h)ZeRGAN (i)UAB-GAN (j)GT

Fig. 9. Spatial error maps of different methods on the WV2 dataset.


(a)Brovey (b)MTF_GLP_HPM (c)PCA (d)SFIM (e)LDPNet

(f)UCGAN (g)PANGAN (h)ZeRGAN (i)UAB-GAN (j)GT

Fig. 10. Spectral error maps of different methods on the WV2 dataset.

Same as the experiments on the QB data set, Figures 9 and 10 show the spatial error maps and spectral error maps of different methods on the WV2 data set. The error of the traditional method is more obvious, which is also verified in the error image. Among unsupervised-based methods, UCGAN has smaller spatial errors, but it has a certain gap compared with UAB-GAN in terms of spectral feature retention. Since the ZeRGAN method does not require pre-training, it is not as good as other unsupervised methods in terms of fused image quality, and there are obvious differences in the error maps. Overall, the fusion results of the UAB-GAN method perform excellently in both spatial and spectral aspects.

4.3 Ablation Experiments

To verify the rationality and effectiveness of each module of the UAB-GAN method, this section conducts corresponding ablation experiments. Correspondingly, the single discriminator generative adversarial network is used as the basic network. In order to obtain more credible experimental data, the experiments set the same number of parameters, and all ablation experiments were performed on the same data set (WV2).

1) Discriminator design: An ablation experiment was conducted on the dual discriminator design, and the experimental data results are shown in Table 4-5. Based on the experimental results, it can be concluded that compared with the generative adversarial network composed of a single discriminator, the generative adversarial network composed of dual discriminators is more competitive in the method proposed in this article. Experimental data shows that the design of dual discriminators in UAB-GAN is effective, and the spectral quality and spatial details of the generated images are significantly improved.

Table 5. Discriminator ablation study.

Single-D	Dual-D	PSNR↑	ERGAS↓	CC↑	SAM↓	UIQI↑
√	×	28.5727	8.2718	0.8994	0.2020	0.7635
×	√	29.6372	6.9228	0.9265	0.1831	0.8026

2) Feature extraction strategy: To confirm the effectiveness of the dual-stream feature extraction module design, relevant ablation experiments were conducted. The experimental data results are shown in Table 6. We describe the network that uses residual blocks to build a dual-stream feature extraction module as UAB-GAN1, and the network that uses CBAM to build a dual-stream feature extraction module as UAB-GAN2. The proposed UAB-GAN uses a mixture of CBAM and CBAM-P to build a dual-stream feature extraction module. Experimental results show that UAB-GAN1 has poor performance because it cannot integrate local and global information. UAB-GAN2 has achieved better results due to the use of attention module, but is slightly worse in terms of comprehensive error. UAB-GAN achieved the best results overall, indicating that this way of designing specific convolutional attention modules for different images is desirable.

Table 6. Ablation research on different feature extraction strategies.

Model	PSNR↑	ERGAS↓	CC↑	SAM↓	UIQI↑
UAB-GAN	29.6372	6.9228	0.9265	0.1831	0.8026
UAB-GAN1	28.4955	8.0687	0.9145	0.2110	0.7795
UAB-GAN2	29.2419	8.3147	0.9364	0.2017	0.8147

5 Conclusion

In this article, an unsupervised framework is designed that can learn directly on a data set of unlabeled images without the need to down-sample the original image to obtain the label. This model is designed based on a generative adversarial network. We use a dual-stream feature extraction module based on different attention to extract differential feature information from PAN and MS images respectively. And a dual discriminator is developed to preserve the input spectral and spatial information when performing fusion. Experiments on data sets of different resolutions show that the combined design model achieves better results without increasing computational complexity. In addition, the study found that using a network structure of two discriminators can achieve satisfactory results, and the differential feature extraction ability of the model is further enhanced after designing specific convolutional attention for different images. Although UAB-GAN has achieved improvements in many indicators, compared with supervised methods, there is still room for improvement in down-scale images. In the future, we will optimize the model and improve overall performance.

Acknowledgments. This study is supported by the Yunnan Fundamental Research Projects (Nos. 202201AT070112, 202301AW070007, 202301AU070210, 202401AT070470, and 202301AT070407), National Natural Science Foundation of China (Nos. 62101481 and 62261060), Major Scientific and Technological Project of Yunnan Province (No. 202202AD080002), and National Double Tops Construction Fund of China (C619300A020),

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Dou W, Chen Y.: An improved IHS image fusion method with high spectral fidelity. The Int Archiv of the Photogramm, Rem Sensing and Spat Inform Sciences 37, 1253-1256 (2008)
2. Kwarteng P, Chavez A.: Extracting spectral contrast in Landsat Thematic Mapper image data using selective principal component analysis. Photogramm. Eng. Remote Sens **55**(1), 339-348 (1989)
3. Maurer T.: How to pan-sharpen images using the gram-schmidt pan-sharpen method–A recipe. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 40, 239-244 (2013)
4. Alparone L, Aiazzi B, Baronti S, et al.: Sharpening of very high resolution images with spectral distortion minimization. IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No. 03CH37477) 1, 458-460 (2003)
5. Chavez P, Sides S C, Anderson J A.: Comparison of three different methods to merge multiresolution and multispectral data-Landsat TM and SPOT panchromatic. Photogrammetric Engineering and remote sensing **57**(3), 295-303 (1991)
6. Khan M M, Chanussot J, Condat L, et al.: Indusion: Fusion of multispectral and panchromatic images using the induction scaling technique. IEEE Geoscience and Remote Sensing Letters **5**(1), 98-102 (2008)

7. Shensa M J.: The discrete wavelet transform: wedding the a trous and Mallat algorithms. IEEE Transactions on signal processing **40**(10), 2464-2482 (1992)
8. Zhao Z, Bai H, Zhu Y, et al.: DDFM: denoising diffusion model for multi-modality image fusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8082-8093 (2023)
9. Zhong S, Zhang Y, Chen Y, et al.: Combining component substitution and multiresolution analysis: A novel generalized BDSD pansharpening algorithm. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **10**(6), 2867-2875 (2017)
10. Jin X, Zhang P, Jiang Q, et al.: F-UNet++: Remote Sensing Image Fusion Based on Multi-purpose Adaptive Shuffle Attention and Composite Multi-Input Reconstruction Network. IEEE Transactions on Instrumentation and Measurement 72, 1-15 (2022)
11. Chen Y, Dai X, Chen D, et al.: Mobile-former: Bridging mobilenet and transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5270-5279 (2022)
12. Ma J, Tang L, Fan F, et al.: SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. IEEE/CAA Journal of Automatica Sinica **9**(7), 1200-1217 (2022)
13. Feng Y, Jin X, Jiang Q, et al.: MPFINet: A Multilevel Parallel Feature Injection Network for Panchromatic and Multispectral Image Fusion. Remote Sensing **14**(23), 6118 (2022)
14. Wang Q, Jin X, Jiang Q, et al.: DBCT-Net: A dual branch hybrid CNN-transformer network for remote sensing image fusion. Expert Systems with Applications 233, 120829 (2023)
15. Liu Z, Lin Y, Cao Y, et al.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision, pp: 10012-10022 (2021)
16. Wald L, Ranchin T, Mangolini M.: Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. Photogrammetric engineering and remote sensing **63**(6), 691-699 (1997)
17. Ni J, Shao Z, Zhang Z, et al.: LDP-Net: An unsupervised pansharpening network based on learnable degradation processes. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15, 5468-5479 (2022)
18. Diao W, Zhang F, Sun J, et al.: ZeRGAN: Zero-reference GAN for fusion of multispectral and panchromatic images. IEEE Transactions on Neural Networks and Learning Systems 34(11), 8195-8209 (2023)
19. Zhou H, Liu Q, Weng D, et al.: Unsupervised cycle-consistent generative adversarial networks for pan sharpening. IEEE Transactions on Geoscience and Remote Sensing 60, 1-14 (2022)
20. Woo S, Park J, Lee J Y, et al.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp: 3-19 (2018)
21. Ma J, Yu W, Chen C, et al.: Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion. Information Fusion 62, 110-120 (2020)
22. Yang J, Fu X, Hu Y, et al.: PanNet: A deep network architecture for pan-sharpening. In: Proceedings of the IEEE international conference on computer vision, pp: 5449-5457 (2017)
23. Isola P, Zhu J Y, Zhou T, et al.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp: 1125-1134 (2017)
24. Zhou H, Liu Q, Wang Y.: PGMAN: An unsupervised generative multiadversarial network for pansharpening. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14, 6316-6327 (2021)

25. Chang C I.: An information-theoretic approach to spectral variability, similarity, and discrimination for hyperspectral image analysis. IEEE Transactions on information theory **46**(5), 1927-1932 (2000)
26. Vivone G, Restaino R, Dalla Mura M, et al.: Contrast and error-based fusion schemes for multispectral image pansharpening. IEEE Geoscience and Remote Sensing Letters **11**(5), 930-934 (2013)
27. Liu J G.: Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details. International Journal of remote sensing **21**(18), 3461-3472 (2000)
28. Alparone L, Aiazzi B, Baronti S, et al.: Multispectral and panchromatic data fusion assessment without reference. Photogrammetric Engineering & Remote Sensing **74**(2), 193-200 (2008)
29. Nezhad Z H, Karami A, Heylen R, et al.: 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), 2377-2389 (2016)
30. Wald L.: 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". SEE/URISCA, pp: 99-103 (2000)
31. Palsson F, Sveinsson J R, Benediktsson J A, et al.: Classification of pansharpened urban satellite images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **5**(1), 281-297 (2011)
32. Vivone G, Alparone L, Chanussot J, et al.: A critical comparison among pansharpening algorithms[J]. IEEE Transactions on Geoscience and Remote Sensing **53**(5), 2565-2586 (2014)
33. Wang Z, Bovik A C.: A universal image quality index. IEEE signal processing letters **9**(3), 81-84 (2002)