Information Fusion 91 (2023) 31–46

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

UPanGAN: Unsupervised pansharpening based on the spectral and spatial loss constrained Generative Adversarial Network

Qizhi Xu [a], Yuan Li [a,*], Jinyan Nie [b], Qingjie Liu [b], Mengyao Guo [a]

[a] School of Mechatronical Engineering, Beijing Institute of Technology, Beijing, China
[b] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China

ARTICLE INFO

Keywords:
Pansharpening
Image fusion
Convolutional neural network
Generative Adversarial Network

ABSTRACT

It is observed that, in most of the CNN-based pansharpening methods, the multispectral (MS) images are taken as the ground truth, and the downsampled panchromatic (Pan) and MS images are taken as the training data. However, the trained models from the downsampled images are not suitable for the pansharpening of the MS images with rich spatial and spectral information at their original spatial resolution. To tackle this problem, a novel iterative network based on spectral and textural loss constrained Generative Adversarial Network (GAN) is proposed for pansharpening. First, instead of directly outputting the fused imagery, the GAN focuses on generating the mean difference image. The input of the GAN is a good initial difference image, which will make the network work better. Second, the coarse-to-fine fusion framework is designed to generate the fused imagery. It uses two optimized discriminators to distinguish the generated images, and performs multi-level fusion processing on PAN and MS images to generate the best pansharpening image in full resolution. Finally, the well-designed loss functions are embedded into both the generator and the discriminators to accurately preserve the fidelity of the fused imagery. We validated our method by the images from QuickBird, GaoFen-2 and WorldView-2 satellites. The experimental results demonstrated that the proposed method obtained a better fusion performance than the state-of-the-art methods in both visual comparison and quantitative evaluation.

1. Introduction

To date, numerous optical earth observation satellites, such as QuickBird, GeoEye-1, GaoFen-2 and WorldView-3, have been launched to simultaneously acquire the panchromatic (Pan) and multispectral (MS) imagery. Since the Pan and MS sensors have to make a fundamental tradeoff between spatial and spectral resolution, the spatial resolution of the MS imagery is lower than that of the Pan imagery [1]. However, the MS imagery with the same spatial resolution as that of the Pan imagery, is more desirable in many applications, such as map updating and urban investigation. As a consequence, pansharpening (PS) technology has been developed to spatially enhance the MS imagery by injecting the high spatial details from the Pan imagery to the MS imagery.

Existing pansharpening methods can be divided into five categories: (1) the component substitution (CS) methods; (2) the multi-resolution analysis (MRA) methods; (3) the variation optimization (VO) methods; (4) the Generative Adversarial Networks (GAN)-free deep learning methods; (5) the GAN-based deep learning methods. With the development of deep learning technology, several network frameworks are used for pansharpening. From the perspective of network framework, we divide deep learning-based methods into the GAN-free and GAN-based pansharpening methods.

1.1. The CS methods

The CS methods [2–4] project the pixel value of the MS imagery into a new feature space by a matrix transformation, then totally or partially substitute the first projected component of MS imagery by the Pan imagery, and finally inversely project these components to sharpen the MS imagery. Intensity-hue-saturation (IHS), principal component analysis (PCA), Brovey transform (BT) and Gram–Schmidt (GS) all belong to this category.

In 2004, a new fast IHS method (FIHS) is proposed [5]. In addition to its characteristics of fast calculation, this method also extends the traditional three-order transformation to the transformation of any order. But the FIHS has the same problem as the IHS algorithm, it can cause spectral distortion. Adaptive IHS (AIHS) [6] approximates the error of linear combination of panchromatic image and multispectral

* Corresponding author.
E-mail addresses: qizhi@bit.edu.cn (Q. Xu), liyuansme@bit.edu.cn (Y. Li), niejinyan@buaa.edu.cn (J. Nie), qingjie.liu@buaa.edu.cn (Q. Liu), myguo@mail.buct.edu.cn (M. Guo).

https://doi.org/10.1016/j.inffus.2022.10.001
Received 10 October 2021; Received in revised form 19 September 2022; Accepted 1 October 2022
Available online 8 October 2022

image to calculate the fusion ratio of different channels. Improved AIHS (IAIHS) [7] uses the gradient information of the multispectral image to assign different weight coefficients to different channels, so that different details can be injected into each channel. PCA [8] is a statistical method that can transform multivariate data with related variables into unrelated variables. The advantage of PCA algorithm is that it is suitable for multi-band image fusion. The fused image has high spatial resolution, good detail information and little influence from noise. However, the disadvantage is that the spectrum distortion is serious.

The BT [9] fusion method is to transform the MS image using color standardization transform, and then product the three bands of the MS image with the high resolution Pan image. The advantage of BT is that the overall structure information is maintained well, but there will be spectral distortion. GS [10] fusion method can not only preserve the spatial details of Pan image well, but also fast. Pansharpening by IHS, PCA, BT and GS technologies have high fidelity in spatial detail information, and these methods have the advantages of fast speed and easy implementation. However, these methods still have limitations. They work well when there is a strong correlation between high resolution panchromatic images and low resolution multispectral images. However, they were unable to account for local differences caused by spectral mismatches between panchromatic and multispectral images. Therefore, the fusion image will produce obvious spectral distortion.

1.2. The MRA methods

Multi-resolution analysis (MRA) algorithm is to transform the source image to obtain the expression coefficient, and then reverse transform the expression coefficient to obtain the final image fusion result. Most of the methods based on MRA use wavelet transform and curvilinear transform. Commonly used models for extracting spatial details include: wavelet, discrete wavelet, addictive wavelet luminance proportional (AWLP), Guided Filter, Curvelet, Contourlet and Laplacian Pyramids, etc.

Wavelet transform has good time-frequency local analysis characteristics and can represent the characteristic information of image in horizontal, vertical and diagonal directions. Ranchin T. et al. [11] first proposed an image fusion algorithm based on discrete wavelet transform in 1993 and achieved good fusion effect. However, the quality of the fused image is affected by the registration accuracy because of the different sizes of wavelet transform images. [12] compares various multi-resolution decomposition algorithms, such as Curvelet and Contourlet, and studies the influence of the number of decomposition levels and the selection of filters on the fusion performance. [13] proposed a multi-sensor image fusion algorithm based on discrete wavelet packet transform. It can fully fuse the information in source image and improve the ability of information analysis and feature extraction. Burt P.J [14] proposed an image fusion algorithm based on Laplacian pyramid transformation. Pyramid transform fusion algorithm can represent important features and details of images at different scales.

1.3. The VO methods

In order to solve the problems of the above two methods, the theory of variational method is used in remote sensing image fusion method. The method assumes that the image is smooth, and puts forward some constraint conditions for image fidelity as the premise, constructs an energy generic function about the image processing problem, and obtains the final image by solving the minimization value of the universal function.

The earliest variational fusion method was P+XS proposed by C. Ballester in 2006 [15]. The basic idea is to extract the spatial information from Pan image and add it to MS image to improve the spatial resolution of MS image. M. Möller et al. [16] proposed variational wavelet pansharpening algorithm. They introduced a match item into the wavelet fusion image and enhanced the texture by combining geometric matches from the P+XS model. Performing minimization in the wavelet domain allows different parameters to be used for different levels of wavelet decomposition. In addition, two function terms are added to help improve the quality of spatial information and preserve the relationship between different bands. Chen et al. [17] proposed dynamic gradient sparse fusion (DGSF) based on local spectral consistency and dynamic gradient sparse fusion. The variational method can protect spectral information and spatial information well.

In addition, there is still a challenge for VO methods to achieve a good balance between spatial and spectral fidelity. Deng et al. [18] provided a good pansharpening method base on Reproducible kernel Hilbert space and Heaviside function. Then, Tian et al. [19] proposed an innovative solution using Cartoon-texture Similarities, which can preserve the global and local spatial details well and obtained high quality pansharpening images. However, the solving process of energy functional is too complicated and the time complexity of the algorithm is very high, so it is difficult to have real-time performance.

1.4. The GAN-free deep learning methods

With the development of deep learning technology, several well-recognized deep learning frameworks have emerged, such as convolutional neural networks, deep residual networks, recurrent neural networks, and auto-encoder, etc. These networks have been introduced into the field of pansharpening and achieved remarkable results.

The deep learning-based pansharpening methods are first inspired by super-resolution methods. The MS imagery pansharpening can be regarded as a special super-resolution technology, i.e., the super resolution of the MS imagery by the simultaneously acquired Pan imagery. From this viewpoint, [20] designed a novel pansharpening neural network (PNN) by modifying the convolutional neural network (CNN) based super-resolution method [21] to carry out the pansharpening work. Recently, the attention mechanism was involved to adjust the spectral and spatial fidelity [22–26]. For example, [22] designed a channel attention model to adaptively correct the characteristics of the channel; To make full use of the inherent similar information between the MS imagery and the Pan imagery, [25] proposed a non-local attention residual network. [26] built an encoding attention module and a fusion attention module to improve the contour information of the fused images.

In addition, several hybrid strategies have been introduced to pansharpening field. In 2017, [27] proposed a mixed model method called PanNet. They train network in high-pass filtering domain to preserve spatial structure, and for the spectral preservation, they adds upsampled MS imagery to network output. Later, [28] proposed a two stage pansharpening network, in which the generalized Laplacian pyramid was included by CNN to predicts the initial spatial information in the first stage. [29] incorporates the component substitution and multiresolution analysis fusion schemes into CNN to estimate the nonlinear information injection models. [30] designed a novel mapping CNN that maps the differential information between the Pan imagery and the MS imagery to the differential information between the Pan imagery and the fused imagery. [31] introduce a deep detail network architecture with grouped multiscale dilated convolutions to pansharpen multiband spectral information. Furthermore, a new method [32] introduced popular Transformer architecture to the field of pansharpening, which aims to build up a long-distance dependency, to make full use of more useful features. These approaches can yield competitive performance. However, they are also limited by the lack of ideal fused images, and rely on the down-sampled Pan and MS images as training samples.

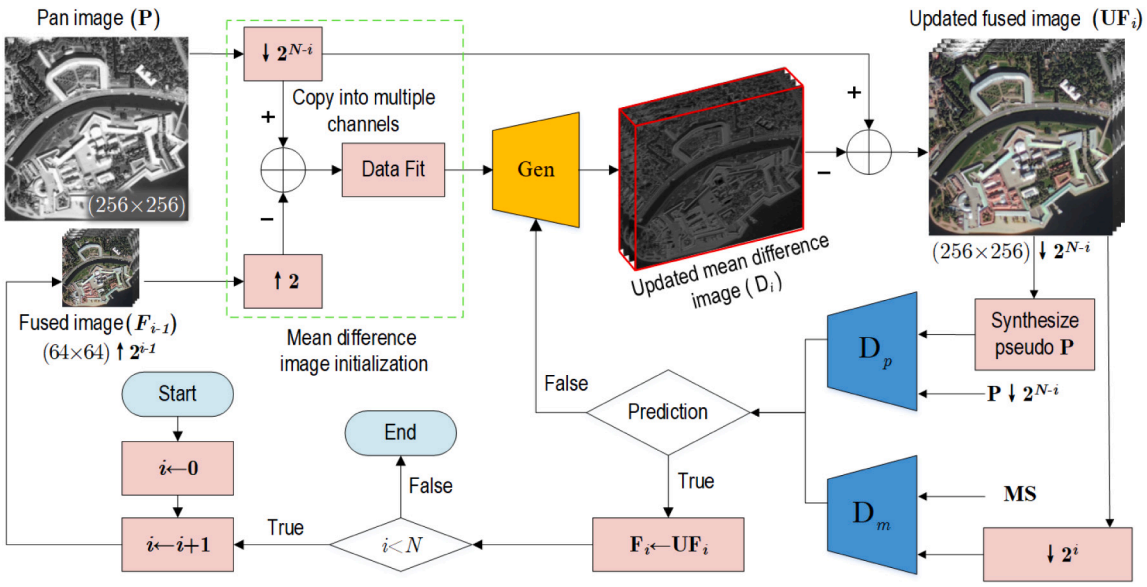


Fig. 1. The flowchart of the proposed method, where Gen is the mean difference image generator, and D_p and D_m are the two fidelity constraint discriminators. Here, the pseudo Pan image is synthesized by applying ratio transformation to the updated fusion image. The notation "↓" denotes the average-based decimation downsampling operation, while the notation "↑" denotes the bilinear interpolation upsampling operation.

1.5. The GAN-based deep learning methods

GAN [33] is a vital machine learning architecture, in which the generator and the discriminator compete with each other to gradually obtain a more accurate prediction. In 2020, [34] first built a pansharpening GAN (PSGAN) to fuse the Pan and MS image fusion. The PSGAN has achieved high-fidelity pansharpening results on the down sampled Pan and MS images. Subsequently, [35] introduced the residual network encoder–decoder model into the generator, and constructed a conditional discriminator network to retain more spatial information. Recently, [36] converted the pansharpening of the MS imagery into the colorization of the Pan imagery, and accordingly proposed a pan-colorization GAN framework (PCGAN). In addition, to improve the fusion performance on the full-resolution scenario, [37] proposed a GAN based pansharpening model that can be trained by the full-resolution Pan and MS imagery; more recently, [38] also designed a pansharpening GAN with full-resolution spectral and spatial discriminators. The two models have obtained a breakthrough in avoiding the lack of ideal training samples. In 2022, Ref. [39] constructed a GAN network with dual discriminators based on gradient and intensity to pansharpening the images effectively. It can ensure that the generated image contains the desired geometric structure and conspicuous information and achieve the best pansharpening effect. However, these models often still encounter the distortion problem when processing the images with rich spatial and spectral information.

Because of the fact that the spatial information of the MS imagery is of great difference from that of the Pan imagery, the spatial distortion is also caused by one-step injecting the spatial information into the MS imagery [21]. These methods have greatly improved the spatial fidelity of fusion performance. However, owing to the lack of ideal fused images, several methods are trained from the down-sampled Pan and MS images. Such trained models often cause spatial distortion when pansharpening the MS images at their original spatial resolution, especially for the regions with sharp edges or rich texture. In this paper, we proposed an unsupervised pansharpening method, named as UPanGAN, based on the spectral and spatial loss constrained GAN. The main purposes of this paper are as follows: (1) to develop an unsupervised pansharpening model that is free of the ideal training samples; (2) to improve the fusion performance on the full-resolution images with rich spatial and spectral information. In general, this work has made three contributions:

Table 1
The frequently used abbreviations.

MS	MS imagery (The size is 64 × 64)
P	Pan imagery (The size is 256 × 256)
$\mathbf{F}_0$	Initial fused image, i.e., MS imagery
$\mathbf{F}_i \uparrow 2$	Fused image F_i upsampled by the ratio of 2 × 2
$\mathbf{P} \downarrow 2^{N-i}$	Pan imagery downsampled by the ratio of $2^{N-i} \times 2^{N-i}$
$\mathbf{UF}_i$	Updated fused image after the generator (The size is 256 × 256 $\downarrow 2^{N-i}$, which increases as i increases.)

(1) A new unsupervised pansharpening GAN model directly trained by original Pan and MS imagery was proposed to solve the training sample problem. Compared with models trained on down-sampled images, the proposed model is more suitable for pan-sharpening the original full-resolution images with rich spatial and spectral information.

(2) A coarse-to-fine scheme with a good initial difference image was designed to improve the fusion fidelity. Network will work better if the initial input is better. The coarse-to-fine scheme can accurately extract the mean difference image than one-step scheme and obtain better fusion images.

(3) According to the spatial and spectral preservation requirements, the well-designed loss functions were employed to fine-tune the UPanGAN training. Experiments demonstrated that the proposed method can indeed achieve good fusion performance.

The paper is organized as follows: the next section describes the proposed methods; in Section 3, we describe the extensive experiments that were carried out to test and validate the proposed method. Finally, a brief summary is presented.

2. Methodology

The proposed model, i.e., UPanGAN, is composed by a mean difference image initialization module and a spectral and spatial loss constrained GAN module. The GAN module includes a mean difference image generator and two loss constrained discriminators. As shown in Fig. 1, the UPanGAN is a coarse-to-fine fusion scheme. Compared with the one-step fusion schemes, the coarse-to-fine fusion scheme can more accurately extract the mean difference image between the Pan imagery and the MS imagery, so that can make a better spatial and spectral

preservation for the full-resolution images with rich spatial information. In the experiments, we have validated this idea by the comparison between the coarse-to-fine scheme and the one-step scheme.

Assume that the spatial resolution ratio between the Pan imagery and the MS imagery is 2^N. For the convenience of discussion, some abbreviations used in the following description are listed in Table 1. The proposed method first generates the mean difference image $\mathbf{D_1}$ through $\mathbf{P} \downarrow 2^{N-1}$ and $\mathbf{F}_0 \uparrow 2$, and then iteratively generates the mean difference image $\mathbf{D}_i$ until $i \geq N$. Here, the notation "↓" denotes the downsampling operation, while the notation "↑" denotes the upsampling operation. In our method, all downsampling operation is carried out through the average-based decimation, and the upsampling operation is conducted by the bilinear interpolation. These two methods are common sampling methods, and they are simple to implement and can achieve good results [40]. Notably, as illustrated in Fig. 1, the generator of our method does not directly generate the fused image $\mathbf{F}_i$, but indirectly obtains the fused imagery by subtracting $\mathbf{D}_i$ from $\mathbf{P} \downarrow 2^{N-1}$, as follows:

$$\mathbf{UF}_i = \mathbf{P} - \mathbf{D}_i = \mathbf{P} - f\left(\mathbf{D}'; \Theta\right) \tag{1}$$

where, D' represents input data of generator, $f(\cdot)$ denotes the generator network, Θ denotes the trainable parameters of the network. We will give a detailed description about the mean difference image initialization and the adversarial learning approach, including the mean difference image generator and the two loss constrained discriminators.

2.1. Mean difference image initialization

The mean difference image initialization is inspired by our previous CS pansharpening method [4]. The mean difference image is the difference image between the mean information of PAN image and MS image. The less difference information, the better the fidelity of the fused image. In fact, the GAN can work much better if reducing the difference between its initial input and its acceptable output. We observe that there is a great spatial difference between the ideal fused imagery and the MS imagery, meanwhile there is also a great spectral difference between the ideal fused imagery and the Pan imagery. However, there is a small variation between the initial mean difference image and the acceptable mean difference image. Accordingly, different from the previous GAN-based fusion methods, the proposed model calculates the mean difference image between the Pan imagery and the fused imagery, rather than directly obtains the fused imagery.

To initialize a qualified mean difference image, the data fitting scheme [4] is employed to smooth the difference image between $\mathbf{F}_{i-1} \uparrow 2$ and $\mathbf{P} \downarrow 2^{N-i}$. It can be written as

$$\mathbf{D}_i(b) = \left[\left(\mathbf{P} \downarrow 2^{N-i}\right) - \left(\mathbf{F}_{i-1}(b) \uparrow 2\right)\right] * \mathbf{G} \tag{2}$$

where $\mathbf{D}_i$ is the initial mean difference image, and $\mathbf{G}$ is a Gaussian filter, and b is the band sequence value. According to our earlier work [4], the optimal standard deviation of the Gaussian filter is set to 2, and its optimal kernel size is set to 7. They have been verified to be the best parameters.

2.2. Spectral and spatial loss constrained GAN

The adversarial learning architecture consists of three modules, i.e., the mean difference image generator, the spatial discriminator and the spectral discriminator. As illustrated by Fig. 1, all these modules are implemented based on CNN. The mean difference image generator iteratively carries out the fine-tuning of the mean difference image ($\mathbf{D}_i$) according to a well-designed loss constraint. The spectral and spatial discriminators check whether the updated mean difference image meets the spatial and spectral preservation constraints.

2.2.1. The mean difference image generator

Network architecture. The network architecture of the mean difference image generator is given by Fig. 2(a). The optimization goal of this generator is to indirectly generate a fused image that is sufficient to deceive the spatial discriminator and spectral discriminator. The mean difference image generator has 7 convolutional layers. In this generator, the odd-numbered layers execute the standard convolution operations, while the even-numbered layers conduct the deformable convolution operations. Unlike standard convolution which can only extract rectangular features (N × N), deformable convolution can adaptively capture features of any shape. Therefore, the addition of deformable convolution enables the model generator to better generate mean difference images. Moreover, small convolution kernels are suitable to extract the tiny spatial difference between the initial and acceptable mean difference image. Therefore, for all the layers, the convolution kernel size is set to 3 × 3; From the first layer to the seventh layer, Leaky ReLU is selected as the activation function. Whereas, for the last layer, Tanh is taken as the activation function. In addition, to improve the reuse rate of feature maps, the skip connections developed by ResNet are adopted by the mean difference image generator. Consequently, the input of a layer is the sum of both the input and output of its previous layers.

The ultimate goal of the proposed generator is to obtain an acceptable mean difference image ($\mathbf{D}_i$) between the ideal fused image ($\mathbf{F}_i$) and the downsampled Pan image ($\mathbf{P} \downarrow 2^{N-i}$). In fact, in contrast with the fused image, the mean difference image is much poorer with spatial details. We have observed that a low-layer network is capable of fine-tuning the mean difference image. To make a balance between the fusion performance and computational cost, the proposed generator was designed to be a 7-layer network.

Loss function. In the proposed generator, the loss function (l_g) plays a key role in generating high-fidelity fused image, and directly affects the quality of the fused image. It consists of the spectral loss function (l_m^g) and the spatial loss function (l_p^g). Hence, it is defined as $l_g = l_p^g + l_m^g$. The l_m^g represents the difference between the updated fused image ($\mathbf{UF}_i$) and $\mathbf{MS}$. Here, $\mathbf{UF}_i$ is calculated by

$$\mathbf{UF}_i^k(b) = \left[\left(\mathbf{P} \downarrow 2^{N-i}\right) - \mathbf{D}_i^k(b)\right] \tag{3}$$

where k is the sequence of the training samples, b is the band sequence number, and $\mathbf{D}_i^k(b)$ is the updated mean difference image. Then, the spectral loss function can be written as follows:

$$l_m^g = \frac{1}{KB}\sum_{k=1}^{K}\sum_{b=1}^{B}\left\|\frac{\mathbf{UF}_i^k(b) \downarrow 2^i - \mathbf{MS}^k(b)}{\mathbf{MS}^k(b)} \times \mathbf{M}_b^k\right\|_F^2 + \alpha l_{a1} \tag{4}$$

where K is the total number of training samples, B is the total band number of $\mathbf{MS}$ imagery, $\mathbf{M}_b^k$ is the mean value of $\mathbf{MS}^k(b)$, and $\|\cdot\|_F^2$ stands for the matrix Frobenius norm. α is a set parameter utilized to adjust the weight of two terms. l_{a1} is the spectral adversarial loss and used to measures the spectral information diversity between the updated fusion image and original MS image, which can be defined as follows:

$$l_{a1} = \frac{1}{K}\sum_{k=1}^{K}\left(\mathrm{D}_m(\mathbf{UF}_i^k \downarrow 2^i) - u\right)^2 \tag{5}$$

where D_m represents the spectral discriminator and u denotes the confidence of the spectral discriminator on the fused image generated by the generator. In the design of the spectral loss function, we have considered the relative values of $\mathbf{MS}^k(b)$. The reason for such consideration is that, when the value of $\mathbf{MS}^k(b)$ is small, a slight change of $\mathbf{UF}_i^k(b)$ may cause obviously spectral distortion. The proposed spectral loss function can magnify the slight spectral variation to avoid spectral distortion.

Due to the lack of ideal fused imagery as reference imagery, it is hard to directly measure the spatial loss. To solve this problem, the

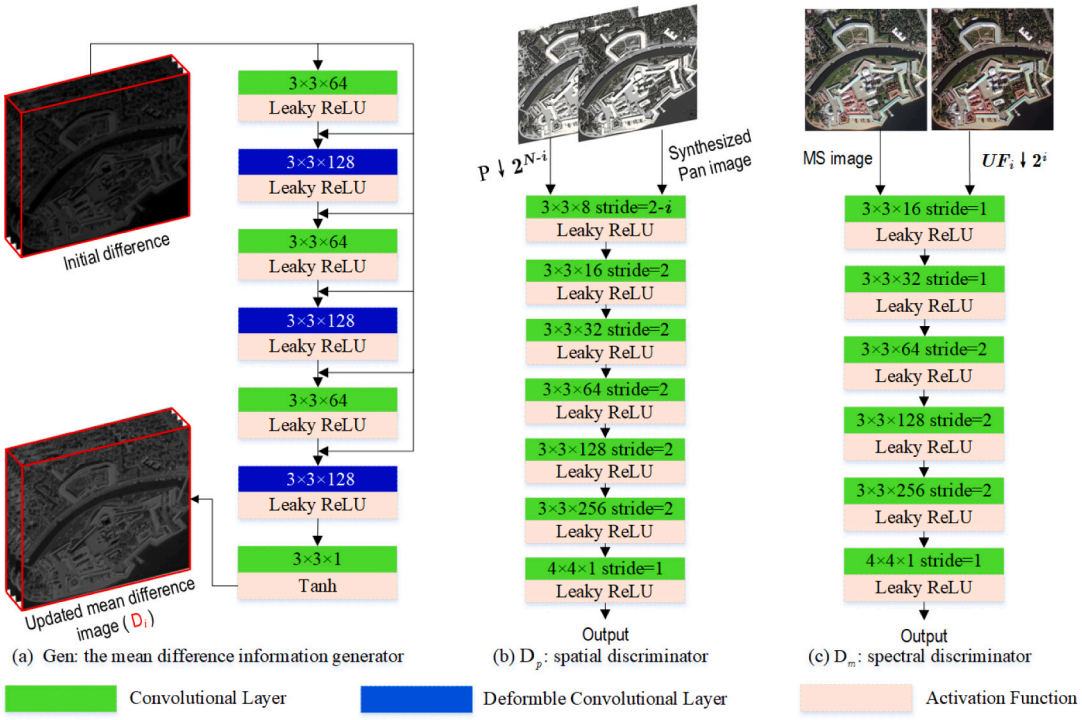


Fig. 2. The detailed structure of the proposed adversarial learning architecture.

downsampled Pan image $\mathbf{P} \downarrow 2^{N-i}$ is taken as the spatial reference, and then a pseudo Pan image $\mathbf{P}_i$ is synthesized to assess the spatial loss. To overcome the oversaturated distortion problem [41], the ratio transformation is applied to synthesize the pseudo Pan image by the candidate fused image. First, the image ratio $\mathbf{R}_i$ is computed by

$$\mathbf{R}_i^k = \sum_{b=1}^{B} \mathbf{UF}_i^k(b) \Big/ \sum_{b=1}^{B} \left[\left(\mathbf{UF}_i^k(b) \downarrow 2 \right) \uparrow 2 \right] \tag{6}$$

where k is the sequence of the training samples, b is the band sequence value, and $\mathbf{UF}_i$ is the updated fused image. The pseudo Pan image is then defined as

$$\hat{\mathbf{P}}_i^k = \mathbf{R}_i^k \times \left(\mathbf{P} \downarrow 2^{N-i} \right) \tag{7}$$

In this situation, the spatial loss can be accurately measured by the difference between $\mathbf{P} \downarrow 2^{N-i}$ and $\hat{\mathbf{P}}_i^k$. Therefore, the spatial loss function is finally defined as

$$l_p^g = \frac{1}{4^i K} \sum_{k=1}^{K} \left\| \left(\mathbf{P} \downarrow 2^{N-i} \right) - \hat{\mathbf{P}}_i^k \right\|_F^2 + \beta l_{a2} \tag{8}$$

where, β is a set parameter utilized to strike a balance between the first and second terms. It deserves mentioning that l_m^g and l_p^g are equally important to fusion performance. Therefore, no additional parameter is set to adjust the weights of l_m^g and l_p^g. Similar to the above spectral loss, we also add the spatial adversarial loss in second term as follows:

$$l_{a2} = \frac{1}{K} \sum_{k=1}^{K} \left(\mathrm{D}_p(\hat{\mathbf{P}}_i^k) - v \right)^2 \tag{9}$$

where D_p represents the spectral discriminator and v denotes the confidence of the spatial discriminator on the fused image generated by the generator.

2.2.2. The spatial and spectral discriminators

Network architecture. As illustrated by Fig. 2(b) and (c), the network structure of the spatial discriminator is similar as that of the spectral discriminator. The identical components of the two networks are as follows: (1) the D_p structure have 7 convolution layers, while the D_m has 6 layers; (2) the kernel size of the first few layers is 3×3, while the kernel size of the last layer is 4×4; (3) for all convolution layers, their number of extracted feature maps are set to [8/none, 16, 32, 64, 128, 256, 1]; (4) the Leaky ReLU is taken as the activation function. Due to the fact that the input image size of the spatial discriminator is 4 times as that of the spectral discriminator, the step size of the spatial discriminator is different from that of the spectral discriminator. Accordingly, the step of the spatial discriminator is set to [2-i, 2, 2, 2, 2, 2, 1], whereas the step size of the spectral discriminator is set to [1, 1, 2, 2, 2, 1]. The two discriminators respectively check the spatial and spectral fidelity of the updated fused image according to the following loss functions. Ultimately, the generator can provide a qualified fused image when passing the fidelity check.

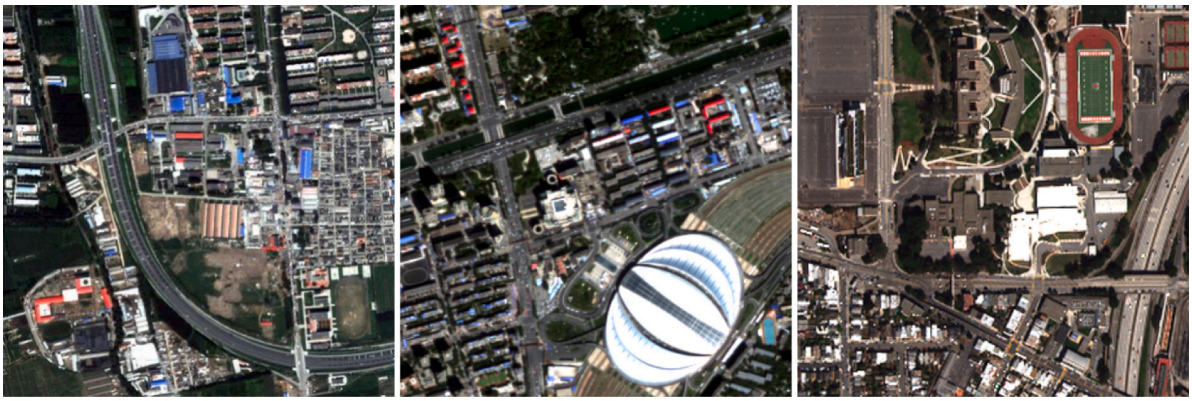
Loss function. The spatial loss function (l_p^d) and the spectral loss function (l_m^d) have been designed to respectively evaluate the spatial fidelity and the spectral fidelity of the updated fused image. The evaluation is a probability whose value ranges from 0 to 1. To assess the spatial fidelity of the updated fused image, similarly as the generator's loss function l_p^g, a pseudo Pan image is synthesized by ratio transformation according to the following formulas:

$$\hat{\mathbf{P}}_i^k = \frac{\left(\mathbf{P} \downarrow 2^{N-i} \right) \times \sum_{b=1}^{B} \mathbf{UF}_i^k(b)}{\sum_{b=1}^{B} \left[\left(\mathbf{UF}_i^k(b) \downarrow 2 \right) \uparrow 2 \right]} \tag{10}$$

Likewise, in the spatial discriminator, the spatial loss function is also calculated by evaluating the difference between $\hat{\mathbf{P}}_i^k$ and $\mathbf{P} \downarrow 2^{N-i}$. Here, $\mathbf{P} \downarrow 2^{N-i}$ is regarded as the spatial target image. Accordingly, the spatial loss function is defined as follows:

$$l_p^d = \frac{1}{K} \sum_{k=1}^{K} \left(\left[\mathrm{D}_p \left(\hat{\mathbf{P}}_i^k \right) - a \right]^2 + \left[\mathrm{D}_p \left(\mathbf{P}^k \downarrow 2^{N-i} \right) - b \right]^2 \right) \tag{11}$$

where a and b denote the labels of the synthesized Pan image ($\hat{\mathbf{P}}_i^k$) and the target image ($\mathbf{P}^k \downarrow 2^{N-i}$), respectively; $\mathrm{D}_p(\hat{\mathbf{P}}_i^k)$ and $\mathrm{D}_p(\mathbf{P}^k \downarrow 2^{N-i})$ stand for the classification results of the pseudo Pan image and the target image. To distinguish the fused image as a fake image and the



(a) QuickBird (b) GaoFen-2 (c) WorldView-2

Fig. 3. The examples of the experimental MS images displayed by the combination of Red, Green and Blue bands.

Table 2
The characteristics of the preprocessed experimental datasets.

Satellite name	Location	Spatial resolution	Size of image (pixel)	Spectral bands and their spectral range	Image number (pair)	
					Train images	Test images
QuickBird	Beijing	MS:2.44 m Pan:0.61 m	MS: 3072 × 3072 Pan: 12288 × 12288	Blue:450–520 nm Green:520–600 nm Red:630–690 nm NIR:760–900 nm Pan:450–900 nm	3	3
GaoFen-2	Beijing	MS:3.24 m Pan:0.81 m	MS: 3072 × 3072 Pan: 12288 × 12288	Blue:450–520 nm Green:520–590 nm Red:630–690 nm NIR:770–890 nm Pan:450–890 nm	3	3
WorldView-2	San Francisco	MS:1.64 m Pan:0.41 m	MS: 3072 × 3072 Pan: 12288 × 12288	Coastal:400–450 nm Blue:450–510 nm Green:510–580 nm Yellow:585–625nm Red:630–690 nm Red Edge:705–745nm NIR1:770–895nm NIR2:860–1040 nm Pan:450–800 nm	3	3

pan image as a real image as much as possible, a is set to 0 and b is set to 1 in experiments.

In the spectral discriminator, the MS image is selected as the spectral target image. Likewise, the spectral loss function is written as follows:

$$l_m^d = \frac{1}{K}\sum_{k=1}^{K}\left(\left[\mathrm{D}_m\left(\mathbf{UF}_i^k \downarrow 2^i\right)-c\right]^2+\left[\mathrm{D}_m\left(\mathbf{MS}^k\right)-d\right]^2\right) \tag{12}$$

where c and d respectively denote the labels of the downsampled fused image ($\mathbf{UF}_i^k \downarrow 2^i$) and the target image ($\mathbf{MS}^k$), $\mathrm{D}_m\left(\mathbf{UF}_i^k \downarrow 2^i\right)$ and $\mathrm{D}_m\left(\mathbf{MS}^k\right)$ are the classification results of the downsampled fused image and the target image. Here, c and d are also set to 0 and 1 respectively in experiments.

3. Experiments and analysis

The proposed method was implemented by TensorFlow and was trained on a workstation with a NVIDIA GeForce RTX 3090 GPU and 64 GB memory. The RMSProp optimizer was employed and the initial learning rate was set to 0.0002, with decay rate is 0.99. The decay step is set to 10000, and the epoch is set to 100. The size of batch images is set to 32. It is observed that the trained pansharpening model is sensitive to the variation of imaging resolution. Therefore, the training samples were divided into different groups according to their satellite sensors. Moreover, the proposed method was respectively trained by different sample groups. The experiments for all methods are completed based on same training and testing samples. We set the best parameters of these methods according to the original papers. In addition, all deep learning-based methods are performed on GPU, while other methods are performed on CPU.

3.1. Datasets and assessment metrics

A variety of large coverage remote sensing images are necessary to validate the fusion performance of different pansharpening methods. Therefore, the Pan and MS images acquired by QuickBird, GaoFen-2 and WorldView-2 satellites were chosen as experimental datasets. The WorldView-2 data set includes a total of 8 bands. Except for R-G-B and NIR bands, other bands provide more information. Experiments on 8 bands of MS images fully verify the effectiveness of the method. Most importantly, the images that have rich texture and colorful land cover were included in our experiments. Fig. 3(a)–(c) exhibits the examples of the experimental images. Such images are highly recommended to test the performance of fusion methods [42].

For the convenience of experimental analysis, the Pan and MS images were registered in advance; Subsequently, the Pan and MS

Table 3
The evaluation scores of different methods tested by the QuickBird dataset.

Method	SAM	ERGAS	Q4	D_λ	D_s	QNR	Time	
							I_t (s)	T_t (h)
Ideal value	0	0	1	0	0	1	–	–
EXP	0.878 ± 0.105	4.525 ± 0.423	0.638 ± 0.052	**0.007 ± 0.001**	0.246 ± 0.035	0.749 ± 0.032	–	–
PRACS	2.801 ± 0.166	2.946 ± 0.330	0.843 ± 0.034	0.152 ± 0.027	0.168 ± 0.029	0.706 ± 0.025	0.983	–
BDSD-PC	2.347 ± 0.144	2.822 ± 0.316	0.854 ± 0.028	0.146 ± 0.020	0.166 ± 0.032	0.715 ± 0.023	0.117	–
MTF-GLP-HPM-H	1.974 ± 0.157	2.613 ± 0.204	0.878 ± 0.021	0.138 ± 0.023	0.150 ± 0.024	0.729 ± 0.019	1.354	–
LGC	1.660 ± 0.152	2.429 ± 0.258	0.862 ± 0.033	0.140 ± 0.016	0.141 ± 0.024	0.739 ± 0.021	0.098	–
A-PNN	1.397 ± 0.141	2.212 ± 0.216	0.886 ± 0.031	0.137 ± 0.014	0.133 ± 0.022	0.748 ± 0.020	0.125	3.23
SDPNet	1.215 ± 0.122	2.351 ± 0.233	0.878 ± 0.027	0.123 ± 0.009	0.127 ± 0.019	0.766 ± 0.018	0.083	2.50
DIRCNN	1.153 ± 0.125	2.436 ± 0.201	0.870 ± 0.029	0.107 ± 0.011	0.122 ± 0.018	0.784 ± 0.017	0.152	3.68
PSGAN	1.236 ± 0.139	2.281 ± 0.189	0.905 ± 0.025	0.134 ± 0.010	0.094 ± 0.007	0.785 ± 0.015	**0.076**	**1.96**
PanGAN	1.192 ± 0.126	1.924 ± 0.177	0.914 ± 0.023	0.115 ± 0.008	0.119 ± 0.015	0.780 ± 0.016	0.092	2.58
UPanGAN	**0.864 ± 0.117**	**1.673 ± 0.183**	**0.942 ± 0.022**	0.102 ± 0.008	**0.096 ± 0.009**	**0.812 ± 0.014**	0.107	2.75

Bold indicates the best result. I_t and T_t denote inference time and training time, respectively.

images were cut to 12288 × 12288 pixels and 3072 × 3072 pixels, respectively. However, for the purpose of clear visualization, only small subset of the experimental images are displayed for visual assessment. Table 2 gives a summary of the experimental images. For each satellite, 3 pairs of Pan and MS images were chosen as training images, while another 3 pairs of Pan and MS images were taken as test images. Specifically, in the training processing, the mean difference images of were divided into 256 × 256 pixels patches. Accordingly, the Pan and MS imagery were divided into 256 × 256 pixels and 64 × 64 pixels patches, respectively. Therefore, the 3 pairs of training images are partitioned into 6912 training samples.

Due to the lack of ideal ground truth images, two different strategies were used to make quantitative assessment. (1) The assessments with reference were employed to assess the quality of the downsampled fused images. Here, such assessments included relative dimensionless global error in synthesis (ERGAS) [43], Spectral Angle Mapper (SAM) [44] and $Q2^n$ [45]. For 4-band datasets QuickBird and GaoFen-2, the $Q2^n$ is Q4; and for 8-band dataset WorldView-2, it is Q8. For all the experiments, the Pan and MS images were fused under the full-resolution state. To make reference-based evaluation, the fused images were downsampled to compare with original MS images used as reference. In general, the assessments drawn on the downsampled images cannot match the user's expectation at full scale [46]. (2) The assessments without reference were used to evaluate the full-resolution fused images. Three reference-free indicators, i.e., the spectral distortion index (D_λ), spatial distortion index (D_s) and quality with no reference (QNR) [47], were introduced to assess the full-resolution fused images. Moreover, human subjective visual observation was also employed to evaluate the fused images. A good fusion result can be intuitively recognized its advantages in color and texture details.

To make experimental comparison, nine state-of-the-art fusion methods were selected. Since the performance of classical algorithms does not depend on the correctness of the training phase, four deep learning-free methods such as PRACS [48], BDSD-PC [49], MTF-GLP-HPM-H [50] and LGC [51] were compared with the proposed methods. Moreover, five deep learning methods, i.e., target-adaptive CNN-based pansharpening (A-PNN) [52], surface- and deep-level constraint-based pansharpening network (SDPNet) [53], differential information residual convolutional neural network (DIRCNN) [54], generative adversarial network for pansharpening (PSGAN) [34], and unsupervised GAN for pansharpening (PanGAN) [37], were also utilized for comparison. These methods were reported a good fusion performance. Likewise, each compared deep leaning-based method was also respectively trained by training samples of different datasets. In addition, the resampled low resolution MS image by the bilinear interpolation without pansharpening processing is also included in the comparisons, referred as EXP.

3.2. Comparison between different methods

3.2.1. Results on QuickBird dataset

Fig. 4(a)–(x) give a subset of experimental results from QuickBird dataset. The true color MS images in Fig. 4(a) and (m) are taken as the visual reference for evaluating spectral quality. Meanwhile, the full-resolution Pan images are exhibited in Fig. 4(b) and (n) to play the role of spatial reference. In general, all the experimental methods obtained acceptable fusion performance on QuickBird dataset. However, from the detail windows, i.e., Fig. 4(m)–(x), small spatial and spectral difference can be distinguished. In the detailed images, the building is blurred in MS image but visible in PAN image. Among these methods, only PSGAN and the proposed method can clearly retain this detail, while that of other methods are weak. However, the PRACS, A-PNN and PSGAN has some spectral distortion. In general, deep learning-based methods have relatively better results than most classic methods. The proposed method obtained a good balance between spatial and spectral preservation. The visual evaluation of the proposed method is better than those of compared methods.

Furthermore, quantitative assessments were also provided by calculating the quality indicators on the 3 pairs of QuickBird test images. The results were the averages and the corresponding variances of the quantities obtained from the 3 pairs of test images. The assessment scores are reported in Table 3. The numerical values evidence that spectral distortion (D_λ) of EXP is almost zero. Because the characteristics of spectral bands are preserved without the pansharpening processing. Compared with EXP, we can found that all these methods have a certain degree of spectral distortion, but incorporates rich spectral feature details. According to [43], a good fusion quality can be achieved when the ERGAS score is less than 3. From this point, the assessment scores are consistent with our visual observation. From Table 3, we can see that the proposed methods have obtained better scores than the compared methods both in terms of spectral fidelity and spatial fidelity. Moreover, we also report the time consumption of different methods in the last two columns of the table. Among them, the interference time-consuming values are derived from the fusion of 512 × 512 size PAN image and 128 × 128 size MS image. The training time-consuming values are the times they take to train the resulting models. Although the time consumption of our method is not the least compared with other methods, it is acceptable.

3.2.2. Results on GaoFen-2 dataset

To evaluate the fusion algorithms for different sensors, GaoFen-2 dataset has been involved in the experiments. Fig. 5(a)–(x) shows a subset of fusion results from GaoFen-2 dataset. The true-color MS images are exhibited in Fig. 5(a) and (m), respectively. The fusion results obtained from GaoFen-2 images are slightly better than those from the QuickBird images. From Fig. 5(a)–(l), it was observed that there was no obvious spatial and spectral distortion in the fused images.

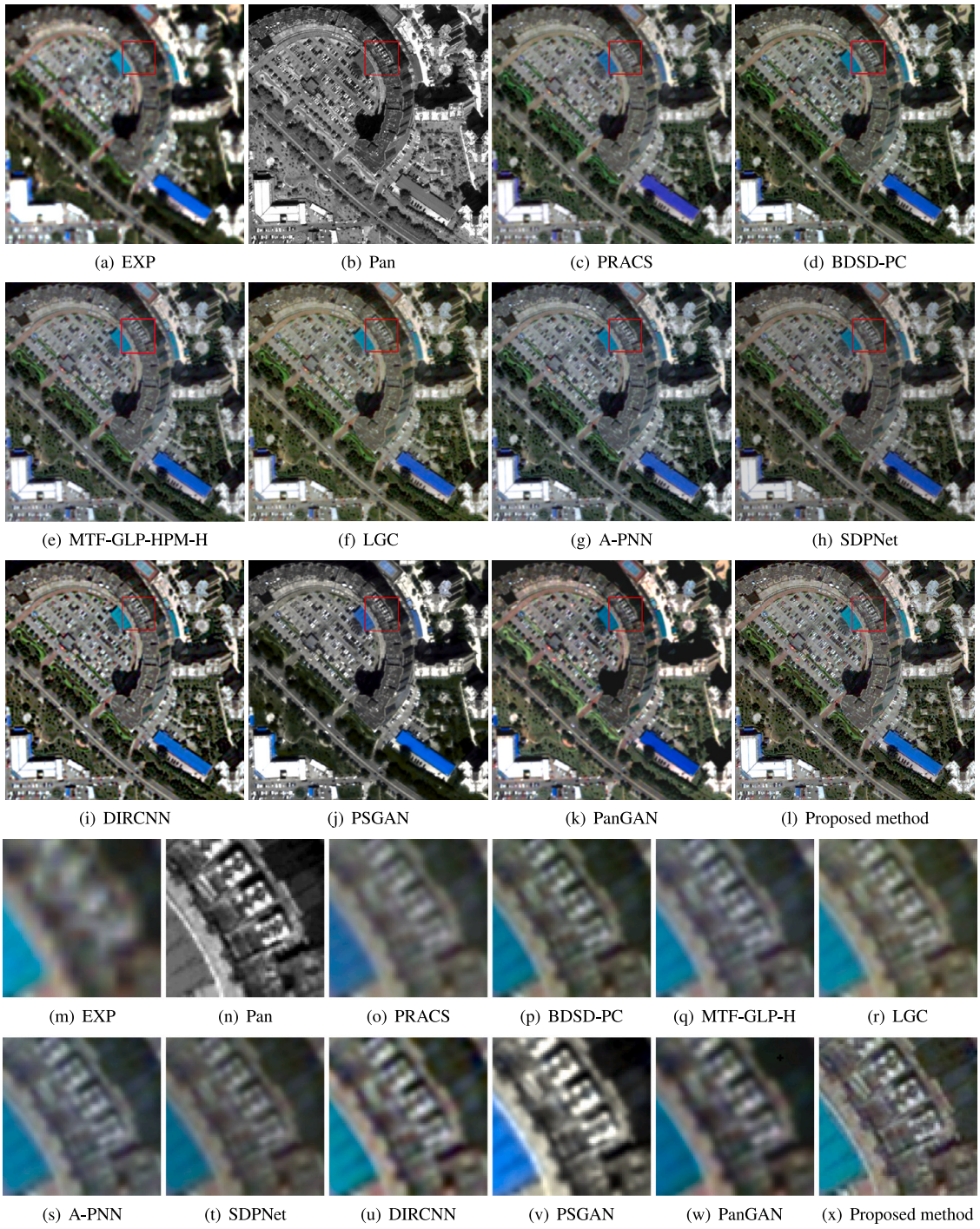


Fig. 4. The experimental result subsets (512 × 512 pixels) of QuickBird satellite images generated by different methods. The full name of subfigure (q) is MTF-GLP-HPM-H.

Furthermore, from Fig. 5(m)–(x), we can obviously see that the fused images of SDPNet, PanGAN and the proposed method are clearer than those of other methods, and also have good spectral preservation. Compared the Fig. 5(t) and (w) with Fig. 5(x), we also found visual performance of the proposed method was better than that of PanGAN and SDPNet.

As shown in Table 4, the results are averages of the quantities obtained from the 3 pairs of GaoFen-2 test images. The quantitative assessments of the fused results from GaoFen-2 images are also slightly better than those from QuickBird images. The proposed method has obtained better scores than the compared methods. Particularly, compared with PRACS, the spatial distortion (D_s) value of the proposed method has improved more than 45% without losing the spectral fidelity. Our approach also obtains the better spatial distortion value (D_s) and spectral distortion value (D_λ) than state-of-the-art deep learning method. At the same time, the interference and training time-consuming values in the table show that our method is also competitive, especially for the deep learning-free methods.

3.2.3. Results on WorldView-2 dataset

Fig. 6(a)–(x) illustrate the fusion results of WorldView-2 dataset. The true color MS and Pan images in Fig. 6(a) and (b) are taken as the visual reference for subjective evaluation. The results of PRACS and A-PNN have relatively large spectral distortion, while the results

Table 4
The evaluation scores of different methods tested by the GaoFen-2 dataset.

Method	SAM	ERGAS	Q4	D_λ	D_s	QNR	Time	
							I_t (s)	T_t (h)
Ideal value	0	0	1	0	0	1	–	–
EXP	0.890 ± 0.118	4.386 ± 0.383	0.646 ± 0.048	**0.004 ± 0.001**	0.309 ± 0.034	0.688 ± 0.030	–	–
PRACS	2.795 ± 0.174	2.823 ± 0.311	0.845 ± 0.028	0.140 ± 0.026	0.204 ± 0.025	0.685 ± 0.023	1.201	–
BDSD-PC	2.402 ± 0.152	2.773 ± 0.297	0.853 ± 0.031	0.133 ± 0.023	0.184 ± 0.022	0.708 ± 0.022	0.136	–
MTF-GLP-HPM-H	2.244 ± 0.162	2.511 ± 0.241	0.870 ± 0.032	0.120 ± 0.015	0.165 ± 0.023	0.731 ± 0.020	1.437	–
LGC	1.884 ± 0.157	2.410 ± 0.224	0.875 ± 0.028	0.119 ± 0.017	0.160 ± 0.018	0.740 ± 0.020	0.117	–
A-PNN	1.691 ± 0.136	2.373 ± 0.208	0.887 ± 0.025	0.115 ± 0.013	0.153 ± 0.017	0.750 ± 0.020	0.122	3.90
SDPNet	1.232 ± 0.118	2.254 ± 0.189	0.893 ± 0.024	0.107 ± 0.011	0.132 ± 0.013	0.775 ± 0.019	0.095	3.04
DIRCNN	1.146 ± 0.133	2.112 ± 0.161	0.905 ± 0.022	0.085 ± 0.007	0.143 ± 0.014	0.784 ± 0.017	0.147	4.71
PSGAN	1.217 ± 0.142	2.004 ± 0.174	0.924 ± 0.025	0.098 ± 0.010	0.125 ± 0.012	0.789 ± 0.014	**0.074**	**2.35**
PanGAN	1.129 ± 0.128	1.725 ± 0.162	0.932 ± 0.019	0.082 ± 0.008	0.090 ± 0.008	0.835 ± 0.015	0.099	3.16
UPanGAN	**0.882 ± 0.130**	**1.610 ± 0.154**	**0.952 ± 0.014**	0.078 ± 0.006	**0.075 ± 0.006**	**0.863 ± 0.012**	0.104	3.33

Bold indicates the best result. I_t and T_t denote inference time and training time, respectively.

Table 5
The evaluation scores of different methods tested by the WorldView-2 dataset.

Method	SAM	ERGAS	Q8	D_λ	D_s	QNR	Time	
							I_t (s)	T_t (h)
Ideal value	0	0	1	0	0	1	–	–
EXP	0.940 ± 0.095	4.512 ± 0.362	0.673 ± 0.042	**0.005 ± 0.001**	0.263 ± 0.029	0.733 ± 0.030	–	–
PRACS	3.220 ± 0.162	3.199 ± 0.352	0.846 ± 0.031	0.149 ± 0.029	0.174 ± 0.022	0.670 ± 0.026	1.375	–
BDSD-PC	2.641 ± 0.142	2.988 ± 0.273	0.881 ± 0.026	0.121 ± 0.028	0.161 ± 0.020	0.731 ± 0.027	0.143	–
MTF-GLP-HPM-H	2.577 ± 0.151	2.960 ± 0.278	0.874 ± 0.024	0.120 ± 0.022	0.160 ± 0.023	0.731 ± 0.022	1.897	–
LGC	2.141 ± 0.136	2.776 ± 0.247	0.896 ± 0.026	0.098 ± 0.014	0.117 ± 0.020	0.787 ± 0.023	0.126	–
A-PNN	2.056 ± 0.129	2.733 ± 0.238	0.912 ± 0.028	0.116 ± 0.019	0.152 ± 0.019	0.741 ± 0.022	0.152	5.38
SDPNet	1.177 ± 0.105	2.484 ± 0.255	0.918 ± 0.022	0.092 ± 0.013	0.140 ± 0.015	0.759 ± 0.021	0.127	4.22
DIRCNN	1.239 ± 0.118	2.680 ± 0.196	0.909 ± 0.020	0.096 ± 0.012	0.133 ± 0.010	0.772 ± 0.019	0.233	6.32
PSGAN	1.132 ± 0.113	2.546 ± 0.177	0.916 ± 0.017	0.088 ± 0.015	0.112 ± 0.013	0.794 ± 0.016	**0.108**	**3.30**
PanGAN	1.298 ± 0.124	2.172 ± 0.184	0.921 ± 0.017	0.102 ± 0.010	0.080 ± 0.009	0.830 ± 0.018	0.139	4.42
UPanGAN	**0.903 ± 0.097**	**1.865 ± 0.168**	**0.936 ± 0.015**	0.079 ± 0.008	**0.072 ± 0.009**	**0.860 ± 0.016**	0.182	4.64

Bold indicates the best result. I_t and T_t denote inference time and training time, respectively.

Table 6
The evaluation scores of different datasets for reduced resolution experiments with assessment strategy 1.

Dataset	SAM	ERGAS	Q4/Q8	D_λ	D_s	QNR	Time	
							I_t (s)	T_t (h)
Ideal value	0	0	1	0	0	1	–	–
QuickBird	0.871 ± 0.119	1.498 ± 0.169	0.932 ± 0.023	0.109 ± 0.008	0.098 ± 0.005	0.833 ± 0.010	0.105	1.54
GaoFen-2	0.894 ± 0.126	1.681 ± 0.160	0.948 ± 0.024	0.083 ± 0.011	0.081 ± 0.006	0.847 ± 0.011	0.107	1.89
WorldView-2	0.909 ± 0.095	1.774 ± 0.121	0.941 ± 0.031	0.069 ± 0.007	0.074 ± 0.008	0.869 ± 0.013	0.166	2.76

Table 7
The evaluation scores of different datasets for reduced resolution experiments with assessment strategy 2.

Dataset	SAM	ERGAS	Q4/Q8	D_λ	D_s	QNR	Time	
							I_t (s)	T_t (h)
Ideal value	0	0	1	0	0	1	–	–
QuickBird	0.877 ± 0.123	1.579 ± 0.154	0.925 ± 0.026	0.115 ± 0.010	0.107 ± 0.006	0.819 ± 0.007	0.105	1.54
GaoFen-2	0.892 ± 0.110	1.673 ± 0.155	0.932 ± 0.022	0.078 ± 0.005	0.080 ± 0.009	0.842 ± 0.009	0.107	1.89
WorldView-2	0.912 ± 0.103	1.735 ± 0.143	0.934 ± 0.025	0.078 ± 0.011	0.072 ± 0.008	0.870 ± 0.012	0.166	2.76

of PRACS and BDSD have spatial distortion. Moreover, by comparing Fig. 6(o)–(r) with Fig. 6(s)–(x), we can observe that the spectral performance of deep learning methods is better than that of deep learning-free methods. Not only for the R-G-B true color band, but other bands also have the same results. It can also be observed that the fused images of the proposed method are clearer than those of the compared methods, especially for spatial texture.

The quantitative assessments, which are the averages of the scores from 3 pairs of WorldView-2 images, are given in Table 5. Overall, the proposed method obtained the best scores. Especially, compared with the deep learning-free methods, the proposed method greatly reduces the spectral distortion value (D_s) and improves the spatial distortion value (D_λ). UPanGAN can also get clearer details than deep learning-based methods. In addition, in contrast with the results on GaoFen-2 datasets, the fusion performance obtained from WorldView-2 dataset slightly decreased. The reason may be that the MS images of WorldView-2 dataset have 8 bands and the Pan images have more high-resolution spatial details, especially the sharp edges. In this situation, it is more difficult to achieve good balance between spatial and spectral preservation. In addition, we can see that the interference and training time values in the last two columns of the table is larger than that of QuickBird dataset and GaoFen-2 dataset. This may be due to the increase in the number of data bands. But compared with other methods, our approach can achieve comparable efficiency.

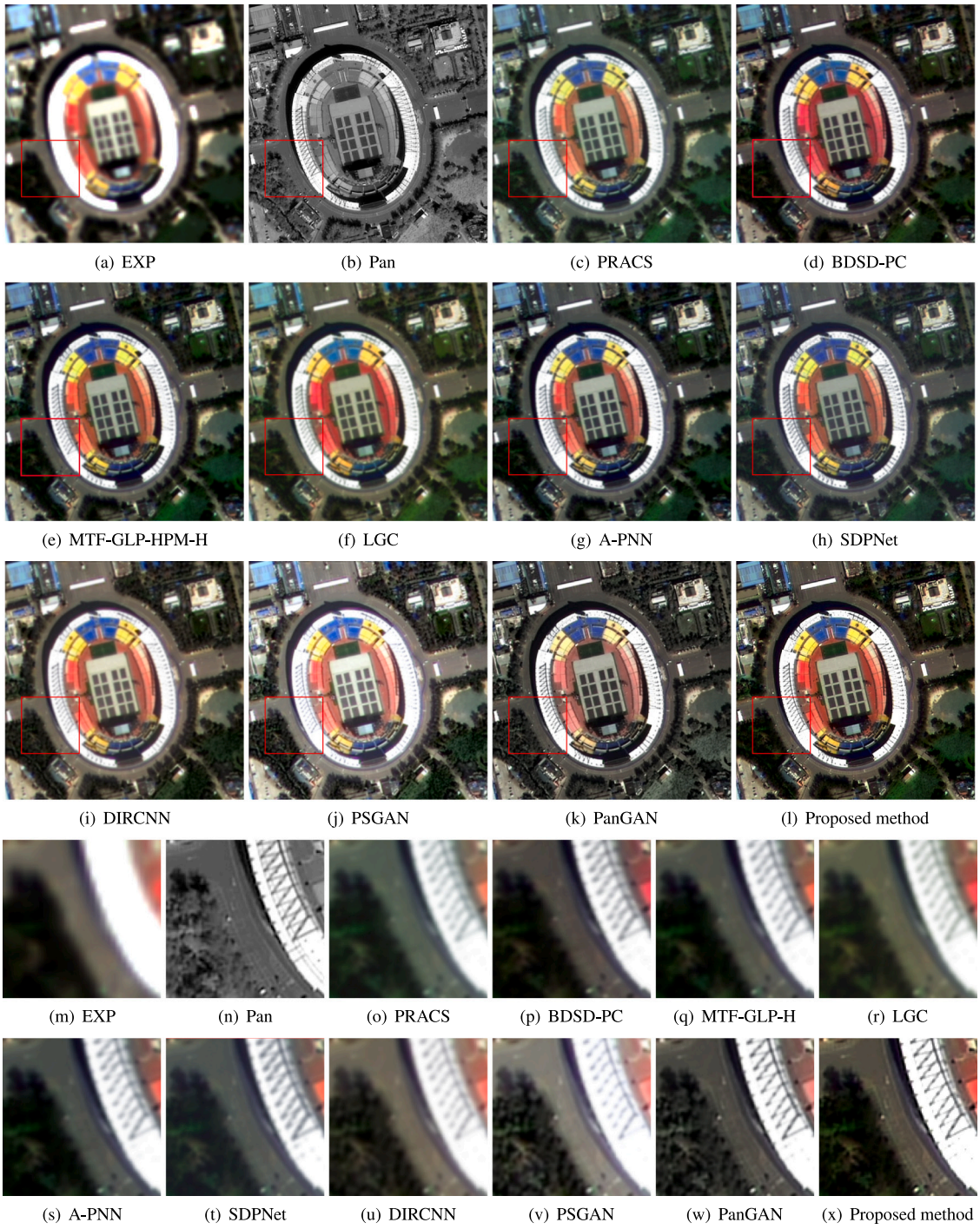


Fig. 5. The experimental result subsets (512 × 512 pixels) of GaoFen-2 satellite images generated by different methods. The full name of subfigure (q) is MTF-GLP-HPM-H.

3.2.4. Results on reduced resolution images

To answer for the synthesis property of the Wald's protocol, we present the supplementary experiment in the reduced resolution assessment. We use the modulation transfer function (MTF) tools in Ref. [55] to reduce the resolution of the original multispectral and panchromatic images, called MS_LR and PAN_LR. The size of them is reduced four times, then the PAN_LR has the same size as the original MS image and MS_LR is one-quarter the size of the original MS. After that, we trained the UPanGAN model based on the reduced resolution dataset. At the same time, to be consistent with the full-resolution training model, we do not use the original MS as the ground truth but use the low-resolution images (PAN_LR and MS_LR) as the training reference.

In addition, we still use low-resolution images (PAN_LR and MS_LR) to test the model and the fused image with the same size as the original MS image is obtained. We use two different assessment strategies to calculate the performance metrics of the model. The first is the traditional assessment method, which calculates the spatial and spectral assessment index using the original MS as a reference. The second is an assessment strategy that is set to have the same situation as the actual functioning of the algorithms. That is, the fused image is further degraded and is compared with the input MS_LR to acquire the results of the spectral assessment; the findings of the spatial assessment are also obtained by comparing the fused image with the input PAN_LR. The specific results can be seen in Tables 6 and 7. We can find that the

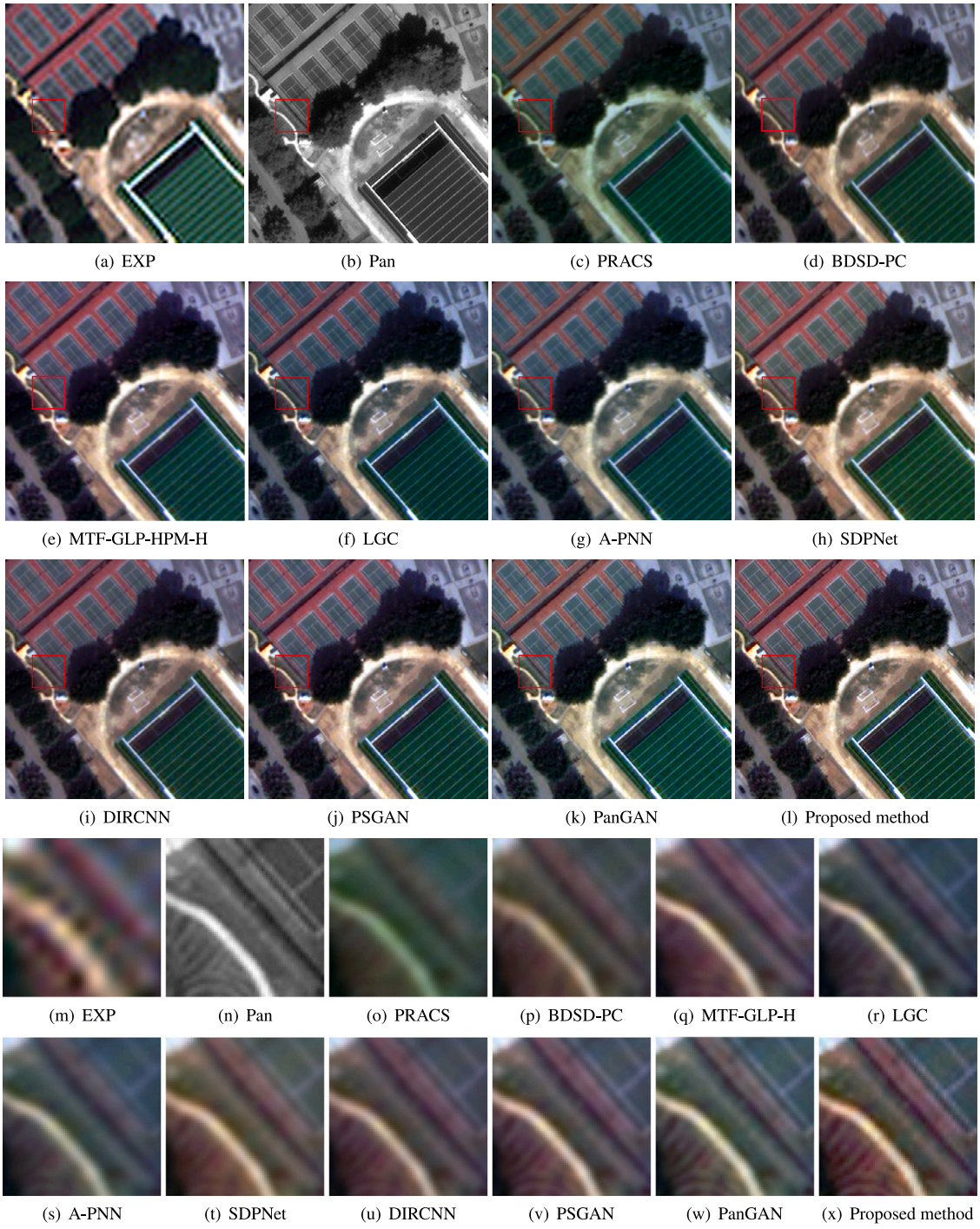


Fig. 6. The experimental result subsets (512 × 512 pixels) of WorldView-2 satellite images generated by different methods. The full name of subfigure (q) is MTF-GLP-HPM-H.

proposed model can still obtain good fusion results for images under these two reduced resolution assessment methods.

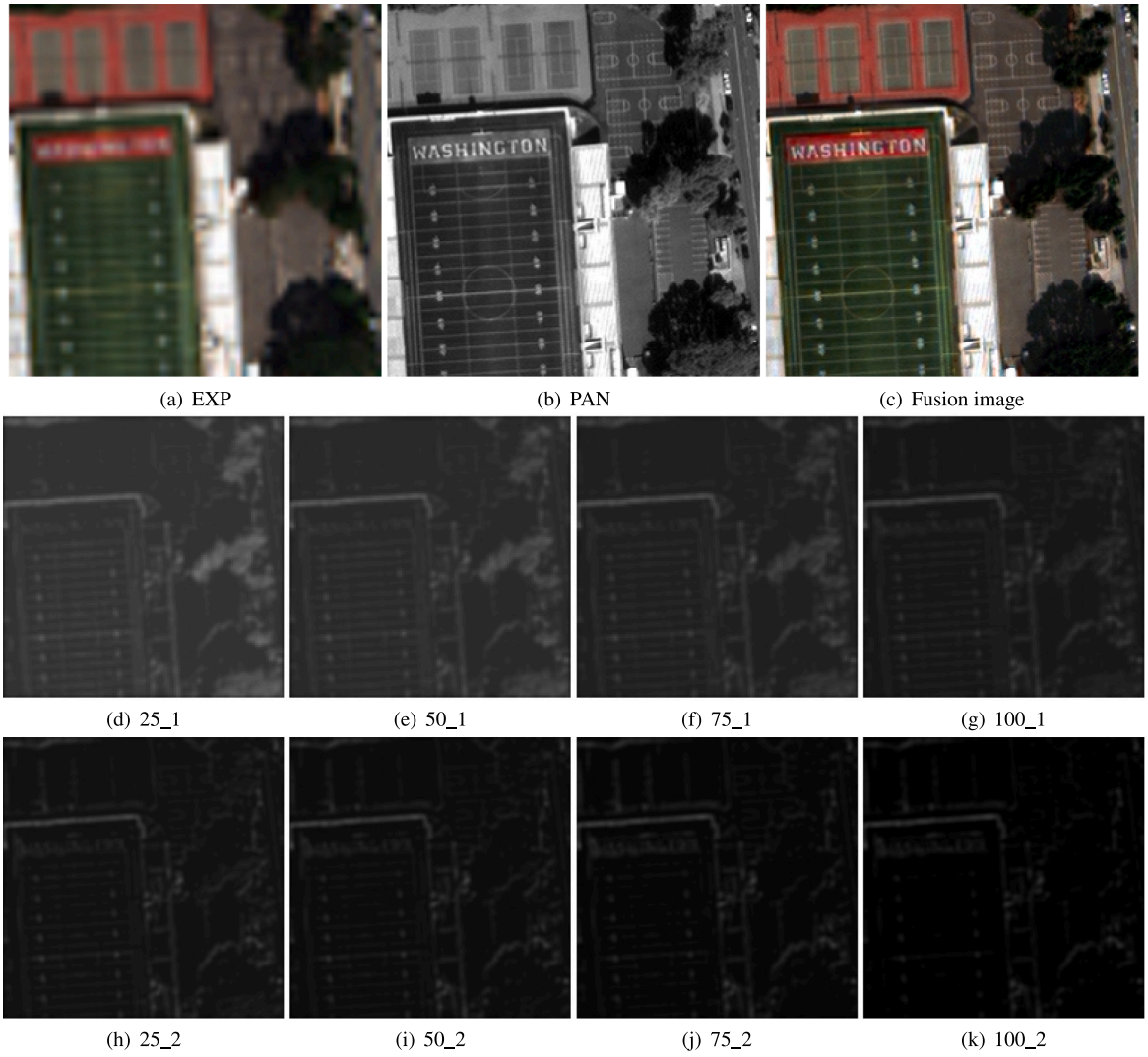
3.3. Analysis of different scheme settings

3.3.1. Analysis of coarse-to-fine scheme

In our approach, the coarse-to-fine scheme has played an important role. The main purpose of this scheme is to reduce the difference between initial input and output of model at each iteration, and make the model optimally optimized. Fig. 7 shows the mean difference images during iteration. The images (128 × 128) in first row are the mean difference images generated in first iteration, and images (256 × 256) in second row are generated in second iteration. It is clearly observed that the difference between PAN and fused images becomes smaller as the number of iterations increases. Therefore, we can say that the proposed scheme can optimize the model from coarse to fine, and finally a good fusion image can be obtained. In contrast with the one-step scheme, it further reduces the spatial difference between Pan and MS imagery. Consequently, the generator can more accurately obtain the mean difference image under the constraints of the loss functions. All else being equal, we replace the coarse-to-fine strategy of the proposed by the one-step strategy. That is, the original MS image

Table 8
The evaluation scores of one step scheme and coarse-to-fine scheme tested by three datasets.

Method	SAM	ERGAS	Q4	D_λ	D_s	QNR
Ideal value	0	0	1	0	0	1
UPanGAN with one-step scheme	0.892	1.914	0.922	0.091	0.099	0.819
UPanGAN with coarse-to-fine scheme	0.883	1.716	0.942	0.086	0.081	0.840



(a) EXP (b) PAN (c) Fusion image

(d) 25_1 (e) 50_1 (f) 75_1 (g) 100_1

(h) 25_2 (i) 50_2 (j) 75_2 (k) 100_2

Fig. 7. Mean difference images during iteration. The number before the underscore is the epoch number, and the number after underscore is the stage of iterations. .

is directly upsampled to the size of the PAN image, and then fed into the model. After training, the final fusion model can be obtained. We made experimental comparison between the two schemes.

Fig. 8 illustrates the fused subsets of the coarse-to-fine scheme and the one-step scheme, respectively. By comparing the fused images of two schemes, we can observe that the two schemes have obtained almost the same spectral fidelity; however, the fused images of the coarse-to-fine scheme are clearer than those of the one-step scheme. Table 8 lists the quantitative evaluation of the fused images. In contrast with the one-step scheme, the coarse-to-fine scheme reduced the spatial distortion value (D_s) by 21.6%. The experimental results demonstrated that the coarse-to-fine scheme is an effective way to improve the spatial performance of deep learning based fusion methods.

3.3.2. Analysis of the loss functions

The loss functions of the generator can directly affect the quality of the fused images. In UPanGAN, the generator loss function (l_g) consist of the spectral loss function (l_m^g) and the spatial loss function (l_p^g), that is $l_g = l_m^g + l_p^g$. To obtain a good fusion performance, the spatial loss function should have the capability to check blurring spatial details. Meanwhile, the spectral loss function should consider the spectral preservation in both strong and weak spectral reflection areas. To verify the effectiveness of the proposed loss function, two loss functions for comparison, i.e., l_λ^g and l_s^g, have been designed as follows:

$$l_\lambda^g = \frac{1}{KB}\sum_{k=1}^{K}\sum_{b=1}^{B}\left\|\mathbf{UF}_i^k(b)\downarrow 2^i - \mathbf{MS}^k(b)\right\|_F^2 + \alpha l_{a1} \tag{13}$$

(a) Coarse-to-fine scheme

(b) One-step scheme

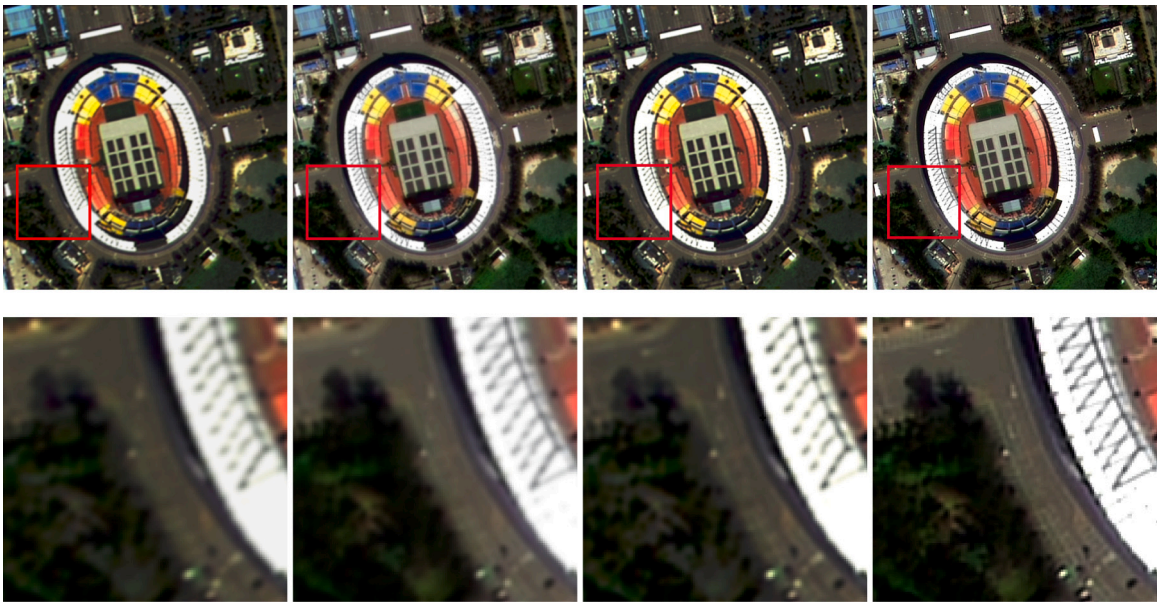
(c) Coarse-to-fine scheme

(d) One-step scheme

Fig. 8. The experimental comparison between the coarse-to-fine scheme and the one-step scheme.

Table 9
The evaluation scores of different loss function schemes tested by three datasets.

Method	SAM	ERGAS	Q	D_λ	D_s	QNR
Ideal value	0	0	1	0	0	1
UPanGAN with loss function l_0	0.926	1.909	0.887	0.113	0.107	0.792
UPanGAN with loss function l_1	0.894	1.775	0.903	0.087	0.099	0.827
UPanGAN with loss function l_2	0.905	1.812	0.912	0.100	0.090	0.819
UPanGAN with loss function l_g	0.883	1.716	0.942	0.086	0.081	0.840



(a) The loss function l_0 (b) The loss function l_1 (c) The loss function l_2 (d) The loss function l_g

Fig. 9. The experimental comparison between the different loss function schemes.

$$l_s^g = \frac{1}{4^i K}\sum_{k=1}^{K}\left\|\nabla\left(\sum_{b=1}^{B}\mathbf{UF}_i^k(b)\right)-\nabla\left(\mathbf{P}\downarrow 2^{N-i}\right)\right\|_F^2+\beta l_{a2} \tag{14}$$

where $\mathbf{UF}_i^k$ is the candidate fused image, ∇ denotes the gradient operator. αl_{a1} and βl_{a2} are the spectral and spatial adversarial losses, which are same with that of the proposed loss function. l_s^g and l_λ^g are used to computed spatial loss and spectral loss respectively. These loss functions are usually used in other GAN methods, such as PanGAN. As a consequence, the spectral loss in UPanGAN is replaced with l_λ^g to prove the good performance. Similarly, l_s^g can be used to verify the proposed spatial loss. We then defined three loss functions for comparison, i.e., $l_0 = l_s^g + l_\lambda^g$, $l_1 = l_s^g + l_m^g$ and $l_2 = l_p^g + l_\lambda^g$. All else being equal, l_0, l_1 and l_2 were used to train the proposed model, respectively.

Compared with the above formulas, we can see the optimization details of the proposed loss functions. Because for the weak spectral reflection areas, such as shadow and water, the spectral reflectance is small. If the absolute difference between the fused image and MS image is considered, the influence of these areas on the loss will be smaller than that of the strong reflection area. Therefore, these areas are prone to spectral distortion. Our improvement is to introduce the relative value of the spectral variation into the loss function. Meanwhile, the spatial loss function should have the capability to find the spatial difference details. Because in the original loss function, the single-channel image ($\sum_{b=1}^{B}\mathbf{UF}_i^k(b)$) synthesized from the fusion image has grayscale difference with the PAN image, it cannot accurately describe the spatial information difference between them. On the contrary, our approach is to inject texture features of the fusion image into the original pan to obtain a synthesized single-channel image ($\widehat{\mathbf{P}}_i^k$), the gray level of which is the same as the PAN image. Consequently, the optimized loss function can accurately capture the difference information.

The fused subsets generated by different loss function schemes are illustrated by Fig. 9. By comparing Fig. 9(a) with Fig. 9(b), we can see that the fused image of the l_1 scheme is not as clear as that of the l_g scheme. This demonstrates that the proposed spatial loss function is an effective constraint for spatial fidelity preservation. In Fig. 9(c), it can observe that the weak spectral refection areas, such as shadow areas and vegetation areas, have obvious spectral distortion. The root cause is that even small change in pixel values may result in obvious spectral distortion in weak spectral reflection areas. The loss function l_2 does not consider the difference between strong and weak spectral reflection areas. To better keep spectral fidelity, the proposed spectral loss function adjusts the spectral fidelity according to the variation ratio. Table 9 gives the quantitative assessment of the fusion results obtained by different schemes. From this table, it was observed that the proposed loss function scheme reduced the spectral distortion value (D_λ) of the loss function L_2 by 14%, and also reduced the spatial distortion value (D_s) of the loss function L_1 by 18%. According to the experimental results, we can see that the proposed loss function is effective to improve both spatial and spectral performance.

3.4. Discussion

In order to obtain good fusion performance, the proposed method has the limitations as follows: (1) The trained model of the proposed method is only suitable for fusing the images with same resolution as training samples. When testing images with other resolution, the fusion performance of model will be reduced. (2) The coarse-to-fine scheme of the proposed method requires the training samples with a ratio of 2^n for PAN and MS images. (3) The proposed method performs n iterations according to the image ratio (2^n) of PAN and MS images. Therefore, it cost more inference time than one-step method. Henceforth, we will further study the lightweight model.

4. Conclusion

In this paper, an unsupervised pansharpening method based on spectral and spatial loss constrained GAN is proposed to pansharpen the full-resolution MS images with rich spatial and spectral information. To obtain better spectral and spatial fidelity, we have employed a coarse-to-fine fusion framework to extract the mean difference image, and have designed a GAN-based deep learning method to tune the mean difference image according to the well designed loss constraints. Finally, the fused images can be generated by subtracting the mean difference image from the Pan imagery. Extensive experiments conducted on Quickbird, GaoFen-2 and Worldview-2 datasets demonstrated that our method had good fusion performance, and was superior to many state-of-the-art fusion methods.

In the future, we will improve the mean difference image generator, so that the trained model can fuse the imagery from different satellites with similar imaging resolution. In addition, we will extend the proposed framework to fuse the Pan and hyperspectral imagery.

CRediT authorship contribution statement

Qizhi Xu: Methodology, Formal analysis, Funding acquisition, Writing – original draft, Review. **Yuan Li:** Conceptualization, Investigation, Data curation, Writing – original draft, Review. **Jinyan Nie:** Visualization, Investigation. **Qingjie Liu:** Resources, Supervision. **Mengyao Guo:** Software, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 61972021 and Grant 61672076.

References

[1] P. Aplin, P.M. Atkinson, P. Curran, Fine spatial resolution satellite sensors for the next decade, Int. J. Remote Sens. 18 (18) (1997) 3873–3881.
[2] C.A. Laben, B.V. Brower, Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening, 2000, Google Patents, US Patent 6, 011, 875.
[3] S. Yang, M. Wang, L. Jiao, R. Wu, Z. Wang, Image fusion based on a new contourlet packet, Inf. Fusion 11 (2) (2010) 78–84.
[4] Q. Xu, B. Li, Y. Zhang, L. Ding, High-fidelity component substitution pansharpening by the fitting of substitution data, IEEE Trans. Geosci. Remote Sens. 52 (11) (2014) 7380–7392.
[5] T.M. Tu, P.S. Huang, C.L. Hung, C.P. Chang, A fast intensity–hue–saturation fusion technique with spectral adjustment for IKONOS imagery, IEEE Geosci. Remote Sens. Lett. 1 (4) (2004) 309–312.
[6] S. Rahmani, M. Strait, D. Merkurjev, M. Moeller, T. Wittman, An adaptive IHS pan-sharpening method, IEEE Geosci. Remote Sens. Lett. 7 (4) (2010) 746–750.
[7] Y. Leung, J. Liu, J. Zhang, An improved adaptive intensity–hue–saturation method for the fusion of remote sensing images, IEEE Geosci. Remote Sens. Lett. 11 (5) (2013) 985–989.
[8] V.P. Shah, N.H. Younan, R.L. King, An adaptive PCA-based approach to pan-sharpening, Proc. SPIE - Int. Soc. Opt. Eng. 6748 (2007) 674802–674802–9.
[9] G.D. Handayani, Pansharpening citra Landsat-8 metode brovey modif pada software Er mapper, 2014, Universitas Gadjah Mada.
[10] Q. Liu, Sharpening the pan-multispectral GF-1 camera imagery using the gram-schmidt approach: the different select methods for low resolution pan in comparison, in: Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery, 2020.
[11] T. Ranchin, L. Wald, The wavelet transform for the analysis of remotely sensed images, Int. J. Remote Sens. 14 (3) (1993) 615–619.
[12] S. Li, B. Yang, J. Hu, Performance comparison of different multi-resolution transforms for image fusion, in: Asia-Pacific Computer Systems Architecture Conference, 2008.
[13] H.H. Wang, J.X. Peng, W. Wu, A fusion algorithm of remote sensing image based on discrete wavelet packet, in: Machine Learning and Cybernetics, 2003 International Conference on, 2003.
[14] P.J. Burt, The Pyramid as a Structure for Efficient Computation, Springer, Berlin Heidelberg, 1984.
[15] C. Ballester, V. Caselles, L. Igual, J. Verdera, B. Rougé, A variational model for P+XS image fusion, Int. J. Comput. Vis. 69 (1) (2006) 43–58.
[16] M. Möller, T. Wittman, A.L. Bertozzi, M. Burger, A variational approach for sharpening high dimensional images, Siam J. Imag. Sci. 5 (1) (2013) 150–178.
[17] C. Chen, Y. Li, L. Wei, J. Huang, Image fusion with local spectral consistency and dynamic gradient sparsity, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2014.
[18] L.-J. Deng, G. Vivone, W. Guo, M. Dalla Mura, J. Chanussot, A variational pansharpening approach based on reproducible kernel Hilbert space and heaviside function, IEEE Trans. Image Process. 27 (9) (2018) 4330–4344.
[19] X. Tian, Y. Chen, C. Yang, J. Ma, Variational pansharpening by exploiting cartoon-texture similarities, IEEE Trans. Geosci. Remote Sens. 60 (2021) 1–16.
[20] G. Masi, D. Cozzolino, L. Verdoliva, G. Scarpa, CNN-based pansharpening of multi-resolution remote-sensing images, in: 2017 Joint Urban Remote Sensing Event, JURSE, 2017, pp. 1–4, http://dx.doi.org/10.1109/JURSE.2017.7924534.
[21] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2) (2016) 295–307, http://dx.doi.org/10.1109/TPAMI.2015.2439281.
[22] X. Li, F. Xu, X. Lyu, Y. Tong, Z. Chen, S. Li, D. Liu, A remote-sensing image pan-sharpening method based on multi-scale channel attention residual network, IEEE Access 8 (2020) 27163–27177, http://dx.doi.org/10.1109/ACCESS.2020.2971502.
[23] S. Luo, S. Zhou, Y. Qi, CSAFNet: Channel similarity attention fusion network for multispectral pansharpening, IEEE Geosci. Remote Sens. Lett. (2020) 1–5, http://dx.doi.org/10.1109/LGRS.2020.3040893.
[24] L. Zhang, J. Zhang, J. Ma, X. Jia, SC-PNN: Saliency cascade convolutional neural network for pansharpening, IEEE Trans. Geosci. Remote Sens. (2021) 1–19, http://dx.doi.org/10.1109/TGRS.2021.3054641.
[25] D. Lei, H. Chen, L. Zhang, W. Li, NLRNet: An efficient nonlocal attention ResNet for pansharpening, IEEE Trans. Geosci. Remote Sens. (2021).
[26] Q. Liu, L. Han, R. Tan, H. Fan, W. Li, H. Zhu, B. Du, S. Liu, Hybrid attention based residual network for pansharpening, Remote Sens. 13 (10) (2021) 1962.
[27] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, J. Paisley, PanNet: A deep network architecture for pan-sharpening, in: 2017 IEEE International Conference on Computer Vision, ICCV, 2017, pp. 1753–1761, http://dx.doi.org/10.1109/ICCV.2017.193.
[28] T. Benzenati, A. Kallel, Y. Kessentini, Two stages pan-sharpening details injection approach based on very deep residual networks, IEEE Trans. Geosci. Remote Sens. (2020).
[29] L.-J. Deng, G. Vivone, C. Jin, J. Chanussot, Detail injection-based deep convolutional neural networks for pansharpening, IEEE Trans. Geosci. Remote Sens. (2020) 1–16, http://dx.doi.org/10.1109/TGRS.2020.3031366.
[30] M. Jiang, H. Shen, J. Li, Q. Yuan, L. Zhang, A differential information residual convolutional neural network for pansharpening, ISPRS J. Photogramm. Remote Sens. 163 (2020) 257–271.
[31] X. Fu, W. Wang, Y. Huang, X. Ding, J. Paisley, Deep multiscale detail networks for multiband spectral image sharpening, IEEE Trans. Neural Netw. Learn. Syst. 32 (5) (2020) 2090–2104.
[32] X. Meng, N. Wang, F. Shao, S. Li, Vision transformer for pansharpening, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–11.
[33] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, 2014, arXiv preprint arXiv:1406.2661.
[34] Q. Liu, H. Zhou, Q. Xu, X. Liu, Y. Wang, PSGAN: A generative adversarial network for remote sensing image pan-sharpening, IEEE Trans. Geosci. Remote Sens. (2020).
[35] Z. Shao, Z. Lu, M. Ran, L. Fang, J. Zhou, Y. Zhang, Residual encoder–decoder conditional generative adversarial network for pansharpening, IEEE Geosci. Remote Sens. Lett. 17 (9) (2020) 1573–1577, http://dx.doi.org/10.1109/LGRS.2019.2949745.
[36] F. Ozcelik, U. Alganci, E. Sertel, G. Unal, Rethinking CNN-based pansharpening: Guided colorization of panchromatic images via GANs, IEEE Trans. Geosci. Remote Sens. 59 (4) (2021) 3486–3501, http://dx.doi.org/10.1109/TGRS.2020.3010441.
[37] J. Ma, W. Yu, C. Chen, P. Liang, X. Guo, J. Jiang, Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion, Inf. Fusion 62 (2020) 110–120.
[38] A. Gastineau, J.-F. Aujol, Y. Berthoumieu, C. Germain, Generative adversarial network for pansharpening with spectral and spatial discriminators, IEEE Trans. Geosci. Remote Sens. (2021) 1–11, http://dx.doi.org/10.1109/TGRS.2021.3060958.
[39] H. Zhou, J. Hou, Y. Zhang, J. Ma, H. Ling, Unified gradient-and intensity-discriminator generative adversarial network for image fusion, Inf. Fusion (2022).
[40] H. Kim, S. Park, J. Wang, Y. Kim, J. Jeong, Advanced bilinear image interpolation based on edge features, in: 2009 First International Conference on Advances in Multimedia, IEEE, 2009, pp. 33–36.
[41] Q. Xu, Y. Zhang, B. Li, Recent advances in pansharpening and key problems in applications, Int. J. Image Data Fusion 5 (3) (2014) 175–195.
[42] X.X. Zhu, R. Bamler, A sparse image fusion algorithm with application to pan-sharpening, IEEE Trans. Geosci. Remote Sens. 51 (5) (2013) 2827–2836.
[43] L. Wald, Quality of high resolution synthesised images: Is there a simple criterion? in: Proc. 3rd Conf. Fusion Earth Data: Merging Point Measurements, Raster Maps and Remotely Sensed Images, SEE/URISCA, 2000, pp. 99–103.
[44] R.H. Yuhas, A.F. Goetz, J.W. Boardman, Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm, in: Proc. Summaries 3rd Annu. JPL Airborne Geosci. Workshop, Vol. 1, 1992, pp. 147–149.
[45] Z. Wang, A.C. Bovik, A universal image quality index, IEEE Signal Process. Lett. 9 (3) (2002) 81–84.
[46] Y. Zhang, Understanding image fusion, Photogramm. Eng. Remote Sens. 70 (6) (2004) 657–661.
[47] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, M. Selva, Multispectral and panchromatic data fusion assessment without reference, Photogramm. Eng. Remote Sens. 74 (2) (2008) 193–200.

[48] J. Choi, K. Yu, Y. Kim, A new adaptive component-substitution-based satellite image fusion by using partial replacement, IEEE Trans. Geosci. Remote Sens. 49 (1) (2010) 295–309.
[49] G. Vivone, Robust band-dependent spatial-detail approaches for panchromatic sharpening, IEEE Trans. Geosci. Remote Sens. 57 (9) (2019) 6421–6433.
[50] S. Lolli, L. Alparone, A. Garzelli, G. Vivone, Haze correction for contrast-based multispectral pansharpening, IEEE Geosci. Remote Sens. Lett. 14 (12) (2017) 2255–2259.
[51] X. Fu, Z. Lin, Y. Huang, X. Ding, A variational pan-sharpening with local gradient constraints, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10265–10274.
[52] G. Scarpa, S. Vitale, D. Cozzolino, Target-adaptive CNN-based pansharpening, IEEE Trans. Geosci. Remote Sens. 56 (9) (2018) 5443–5457.
[53] H. Xu, J. Ma, Z. Shao, H. Zhang, J. Jiang, X. Guo, SDPNet: A deep network for pan-sharpening with enhanced information representation, IEEE Trans. Geosci. Remote Sens. 59 (5) (2020) 4120–4134.
[54] M. Jiang, H. Shen, J. Li, Q. Yuan, L. Zhang, A differential information residual convolutional neural network for pansharpening, ISPRS J. Photogramm. Remote Sens. 163 (2020) 257–271.
[55] G. Vivone, M. Dalla Mura, A. Garzelli, R. Restaino, G. Scarpa, M.O. Ulfarsson, L. Alparone, J. Chanussot, A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods, IEEE Geosci. Remote Sens. Mag. 9 (1) (2020) 53–81.