Information Fusion 62 (2020) 110–120

Contents lists available at ScienceDirect

Information Fusion

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

Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion

Jiayi Ma[a,*], Wei Yu[a], Chen Chen[b], Pengwei Liang[a], Xiaojie Guo[c], Junjun Jiang[d]

[a] *Electronic Information School, Wuhan University, Wuhan 430072, China*
[b] *Department of Electrical and Computer Engineering, University of North Carolina at Charlotte, NC 28223, USA*
[c] *School of Computer Software, Tianjin University, Tianjin 300350, China*
[d] *School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China*

ARTICLE INFO

Keywords:
Pan-sharpening
Unsupervised learning
Image fusion
Generative adversarial network
Deep learning

ABSTRACT

Pan-sharpening in remote sensing image fusion refers to obtaining multi-spectral images of high-resolution by fusing panchromatic images and multi-spectral images of low-resolution. Recently, convolution neural network (CNN)-based pan-sharpening methods have achieved the state-of-the-art performance. Even though, two problems still remain. On the one hand, the existing CNN-based strategies require supervision, where the low-resolution multi-spectral image is obtained by simply blurring and down-sampling the high-resolution one. On the other hand, they typically ignore rich spatial information of panchromatic images. To address these issues, we propose a novel unsupervised framework for pan-sharpening based on a generative adversarial network, termed as Pan-GAN, which does not rely on the so-called ground-truth during network training. In our method, the generator separately establishes the adversarial games with the spectral discriminator and the spatial discriminator, so as to preserve the rich spectral information of multi-spectral images and the spatial information of panchromatic images. Extensive experiments are conducted to demonstrate the effectiveness of the proposed Pan-GAN compared with other state-of-the-art pan-sharpening approaches. Our Pan-GAN has shown promising performance in terms of qualitative visual effects and quantitative evaluation metrics.

1. Introduction

Nowadays, a large amount of earth observation satellites have been launched, such as WorldView, QuickBird, SPOT, Landsat, IKONOS, GaoFen-1 and GaoFen-2, which lead to numerous remote sensing images available for various fields including geography, agriculture, land survey and environmental monitoring [1,2]. In remote sensing systems, satellites can obtain two kinds of images in entirely different modalities, *i.e.*, multi-spectral image and panchromatic image. The multi-spectral images possess high spectral resolution while with low spatial resolution, which are limited by the onboard storage and bandwidth transmission [3]. On the contrary, the panchromatic images have low spectral resolution but high spatial resolution due to the large instantaneous field of view [4]. The task of pan-sharpening in remote sensing image fusion aims at obtaining fused images with both high spectral resolution and high spatial resolution.

Over the past few decades, different pan-sharpening methods have been investigated. Traditional attempts can be generally divided into four categories, *i.e.*, i) methods based on component substitution (CS) [5]: For example, GS adaptive (GSA) [6] is a general scheme which is capable if modeling any CS image fusion method by adopting multivariate regression to improve spectral quality without diminishing spatial quality. Adaptive component-substitution-based fusion using partial replacement (PRACS) [7] generates high-/low-resolution synthetic component images by partial replacement and uses statistical ratio-based high-frequency injection; ii) multi-scale decomposition-based methods [8]: For instance, coupled nonnegative matrix factorization unmixing (CNMF) [9] alternately unmixes hyperspectral and multispectral data into end member and abundance matrices. Sensor observation models that relate the two data are built into the initialization matrix of each NMF unmixing procedure. Modulation transfer functions-generalized Laplacian pyramid (MTF-GLP) [10] is a multiscale and oversampled structure which relies on the generalized Laplacian pyramid (GLP). It selectively performs injection of spatial frequencies from an image to another with the constraint of thoroughly retaining the spectral information of the coarser data; iii) hybrid methods [11]: For example, in hybrid color mapping (HCM) [12], two newly developed techniques are integrated. One is an HCM algorithm and the other one is a plug-and-

[*] Corresponding author.
E-mail addresses: jyma2010@gmail.com (J. Ma), yuwei998@whu.edu.cn (W. Yu), chenchen870713@gmail.com (C. Chen), erfect@whu.edu.cn (P. Liang), xj.max.guo@gmail.com (X. Guo), junjun0595@163.com (J. Jiang).

https://doi.org/10.1016/j.inffus.2020.04.006
Received 3 January 2020; Received in revised form 30 March 2020; Accepted 26 April 2020
Available online 1 May 2020
1566-2535/© 2020 Elsevier B.V. All rights reserved.

play algorithm for single image super-resolution; iv) model-based methods [13]. In band-dependent spatial-detail (BDSD) [14], the solution minimizes the squared error between the low-resolution multi-spectral image and the fused result obtained by spatially enhancing a degraded version of the multi-spectral image through a degraded version, by the same scale factor, of the panchromatic image.

However, these traditional methods suffer from severe spectral distortion owing to the strong assumptions that are not realistic from the viewpoint of remote sensing physics [3]. Recently, deep learning has achieved great success in various fields [15], such as computer vision [16], pattern recognition [17] and image processing [18]. Many researchers have also introduced the convolution neural network (CNN) based deep learning methods into the task of pan-sharpening, such as PNN [19], PanNet [20] and PSGAN [21]. Although they have achieved desirable fusion performance with few spectral distortion, two problems have not been solved yet. One is that the traditional deep learning-based pan-sharpening schemes [19–21] require additional supervision. They treat original high-resolution multi-spectral (HRMS) images as ground-truth. Since the relation between low-resolution multi-spectral (LRMS) images and HRMS images tend not to obey simple blur and interpolation operation, it is not reasonable to obtain the LRMS images by blurring and downsampling the original HRMS images. The other problem goes to that they mainly use the spectral information by regarding the CNN model as a black-box under the supervision of so-called ground-truth (*i.e.*, original HRMS images), which, however, overlook the rich spatial information of panchromatic images.

To overcome the above-mentioned problems, we propose a novel unsupervised framework for pan-sharpening based on a generative adversarial network (Pan-GAN). The pan-sharpening can be formulated as a multi-task problem, aiming to preserve the spectral information of LRMS image and maintain the spatial information of panchromatic image. More specifically, due to the lack of ground-truth, it is assumed that the spectral distribution of the fused image should be consistent with that of the LRMS image, and the spatial distribution of the fused image should be consistent with that of the panchromatic image with the same resolution. Thus, in the proposed unsupervised Pan-GAN framework, the generator attempts to generate an HRMS image containing major spectral information of the LRMS image together with additional image gradients of the panchromatic image. And then, by the adversarial system, the spectral discriminator tries to force the spectral information of the generated image to be consistent with that of the LRMS image, and the spatial discriminator tries to force the spatial distribution of the generated image similar to that of the panchromatic image. Thus, our proposed Pan-GAN can jointly preserve the rich spectral information of the LRMS image and the abundant spatial information of the panchromatic image.

To show the superiority of our method, we present a representative example in Fig. 1. The left two images are the interpolated multi-spectral image and panchromatic image, where the multi-spectral image has higher spectral resolution with four spectral bands, and the panchromatic image has higher spatial resolution. The third image is the pan-sharpening result of an existing CNN-based method PNN [19]. Clearly, it only preserves the spectral information of the original LRMS image, which is very similar to the interpolated multi-spectral image. But the spatial information in the panchromatic image is nearly lost, for instance, the texture information on the street. In contrast, the result of our Pan-GAN can preserve both the spectral information of original LRMS image and the spatial information of panchromatic image, where the texture information on the street can also be clearly observed.

The major contributions of this paper are summarized as follows. First, unlike other CNN-based pan-sharpening methods, our unsupervised pan-sharpening framework, say Pan-GAN, is based on the generative adversarial network, which does not rely on the so-called ground-truth, and the training process is based on the original source image by designing specific loss functions. Second, our proposed Pan-GAN adopts two discriminators to force the spectral and spatial information of the generated image to be consistent with the LRMS and panchromatic images, respectively. In this way, the rich spectral information of the LRMS image and the abundant spatial information of the panchromatic image can be preserved simultaneously. Third, we provide both qualitative and quantitative comparisons between Pan-GAN and other state-of-the-art methods to show the validity and superiority of the proposed method.

The rest of our paper is organized as follows. Section 2 introduces the related work and background material. In Section 3, we describe our method in detail. Section 4 verifies the validity of our model structures and combined loss through extensive experiments, and we also illustrate our method for pan-sharpening compared with seven state-of-the-art methods on public available datasets including WorldView II and GaoFen-2. In Section 5, we give some concluding remarks.

2. Related work

This section briefly reviews the background material that our work is based on, including the methods for pan-sharpening, GANs, and GANs with multiple discriminators.

2.1. Methods for pan-sharpening

In recent years, a large number of pan-sharpening and image fusion methods have been proposed due to the fast-growing demands and progress of image representation methods [22–28]. According to the theories adopted, these fusion methods can be broadly divided into five categories. The first is component substitution-based methods [5,29,30]. These methods usually use a linear transformation and substitution for some components in their transformed domain, which are very fast and easy to be implemented, such as intensity-hue-saturation [31] and principle components transform [32]. The second is multi-scale decomposition-based methods [8,33], which involve three steps, *i.e.*, decomposition, fusion and transform. The decomposition-based methods provide both spatial and frequency domain localization and achieve better performance. The third is hybrid methods [11,33], which combine the advantages of both component substitution and multi-scale decomposition methods. A representative is the fusion based on curvelet and ICA [33]. The fourth is model-based methods [13,34,35]. An MRF model

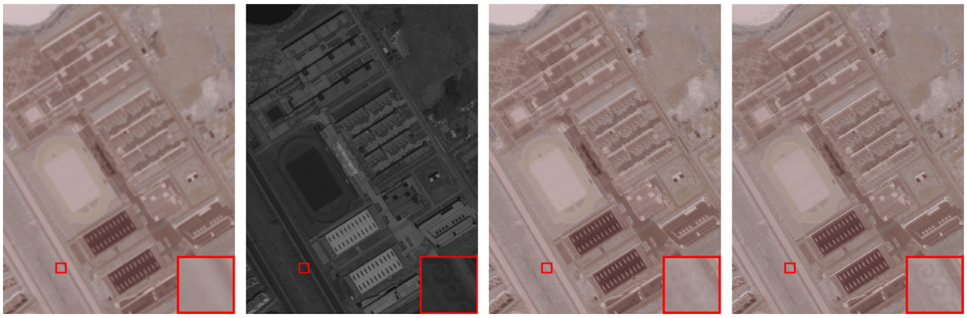

Fig. 1. Schematic illustration of pan-sharpening. From left to right: interpolated original multi-spectral image, original panchromatic image, pan-sharpening result of PNN [19], pan-sharpening result of our proposed Pan-GAN.

has been adopted to model the images for the fusion of edge information in [34]; while [35] adopts a superposition strategy to make full use of the information of LRMS and panchromatic images to decrease the spectral distortion and preserve the spatial information. The fifth is deep-learning-based methods [19–21,36,37]. These methods can typically achieve desirable fusion performance relying on powerful feature extracting ability of convolution network with few spectral distortions. However, there are still some drawbacks. In PNN [19], the three-layer architecture is modified by SRCNN with more specific knowledge in remote sensing and with HRMS images for supervised learning. The similarity of them is measured by the mean square error. In PanNet [20], the up-sampled LRMS image is directly propagated to the output of the network to preserve the spectral information and the network is trained in the high-pass domain to preserve the spatial structure. However, the operation of up-sampling and the selection of high-pass domain will introduce some blurred information into the result or lose some spatial structures. In PSGAN [21] and RED-cGAN [37], which are GAN-based models, the generator tries to generate the image similar to the ground truth and the discriminator tries to distinguish between the generated images and the HRMS images. In RED-cGAN, the results are further improved by introducing the residual encoder-decoder network and conditional GAN. So both the generator and the discriminator network in these two methods need the HRMS images for supervised learning.

In general, the above-mentioned methods either suffer from severe spectral distortion or require so-called ground-truth as supervision. In this paper, we introduce an unsupervised learning method based on GAN with specifically designed loss functions for pan-sharpening to overcome these problems.

2.2. Generative adversarial networks

The GAN was firstly proposed by Goodfellow et al. [38], which is originally designed to generate more realistic images in an unsupervised way. Since then, it has achieved great successes in many computer vision tasks [39–43]. The main idea of GANs is to build a minmax two-player game between learning a generator and a discriminator. The generator takes noise as input and tries to generate different samples to fool the discriminator, while the discriminator aims to determine whether a sample is from the model distribution or the data distribution. The adversarial relationship between the generator and the discriminator continues until the generated samples cannot be distinguished by the discriminator. The generator then can capture the data distribution with more realistic image samples produced. Mathematically, a generative model G aims to generate samples, whose distribution (P_G) tries to approximate the distribution (P_{data}) of real training data, G and D play the minimax two-player game as follows:

$$\min_G \max_D V_{\text{GAN}}(G,D) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))]. \tag{1}$$

The original GANs can work well for generating digital images on MNIST dataset. However, there still exist noise and incomprehensibility in the generated results, especially for high-resolution images. In order to improve the quality of generated images, there are many works proposed recently. Laplacian pyramid has been utilized in LAPGAN [44] to generate high-resolution image supervised by the low-resolution image, but it does not work well for the images containing wobbly objects. In [45,46], it has succeeded to generate nature images, but cannot leverage the generators for supervised learning. Also, to improve the GANs' stability during the training process, DCGAN [47] makes an achievement on applying deeper convolution neural networks to GANs, which drafts a rule about designing CNN architecture of generator and discriminator for steady training. WGAN [48] relaxes the GANs training requirement by modifying the objective function of GANs, which makes the model slower to converge compared with the original GANs. LSGAN [49] overcomes this question by using the least squares loss function. Minimizing the objective function of LSGAN yields minimizing the Pearson χ^2 divergence, and LSGAN has two advantages over the regular GANs. On the one hand, LSGAN can generate higher quality images than regular GANs. On the other hand, LSGAN performs more stably than regular GANs during the learning process. The objective functions of LSGANs are defined as follows:

$$\min_D V_{\text{LSGAN}}(D) = \frac{1}{2}\mathbb{E}_{x\sim p_{\text{data}}(x)}[(D(x)-b)^2] + \frac{1}{2}\mathbb{E}_{z\sim p_z(z)}[(D(G(z))-a)^2], \tag{2}$$

$$\min_G V_{\text{LSGAN}}(G) = \frac{1}{2}\mathbb{E}_{z\sim p_z(z)}[(D(G(z))-c)^2]. \tag{3}$$

Considering the speed of convergence and the difficulty of training, we choose LSGAN as the basic GAN in our paper.

2.3. GANs with multiple discriminators

Unlike standard GANs which have a single generator corresponding to a single discriminator, there are some GANs with multiple discriminators proposed according to multiple different tasks remained to be solved. GMAN [50] extends GANs to multiple discriminators, which can be trained reliably with the original and can produce higher quality samples in a fraction of iterations. PS^2-MAN [51] iteratively generates low-resolution to high-resolution images in an adversarial way using multi-adversarial networks. Moreover, FakeGAN [52] uses two discriminators, where the one helps generator to be close to the deceptive review distribution and the other makes generated results more realistic.

In our proposed method, the goals of multiple adversarial relationships are originally defined as the spatial and spectral information preservation. Therefore, we use two discriminators, *i.e.*, a spatial discriminator and a spectral discriminator to improve the quality of our fusion results.

3. Method

In this section, we describe the proposed unsupervised pan-sharpening framework, namely Pan-GAN, for multi-spectral image and panchromatic image fusion, also called pan-sharpening. First, we formulate the pan-sharpening problem based on the GANs. Then, we present the loss functions and architectures of both generator and discriminator. Finally, we give some implementation details in our training process.

3.1. Overview of the framework

Tracing the source, the primary target of the multi-spectral and panchromatic image fusion is to preserve the spatial and spectral information. However, existing methods based on CNN, such as PNN [19] and PSGAN [21], usually treat pan-sharpening as a black-box deep learning problem. Even though PanNet [20] focuses on preserving the spatial and spectral information, it gets the fused image by combining the interpolated multi-spectral image with the high frequency information obtained by CNN, which very likely leads to blurry results. Besides, the above-mentioned methods rely on the ground-truth image, *i.e.*, Wald's protocol [53], where all original images are blurred by a Gaussian kernel and then downsampled by a factor of 4. All these downsampled images are treated as training data, and the original images are regarded as ground-truth. However, this operation may not make sense. In fact, the relation between the LRMS and HRMS images always tends not to obey simple blur and down-sample operation, which is influenced by many different factors in real-world scenes. Therefore, we propose an unsupervised pan-sharpening framework Pan-GAN, which uses the original source images as training data, and gets HRMS fused image without supervision by ground-truth.

In order to preserve the spectral information of LRMS image and spatial information of panchromatic image, we formulate the pan-sharpening problem as a multi-task problem, and utilize a generative

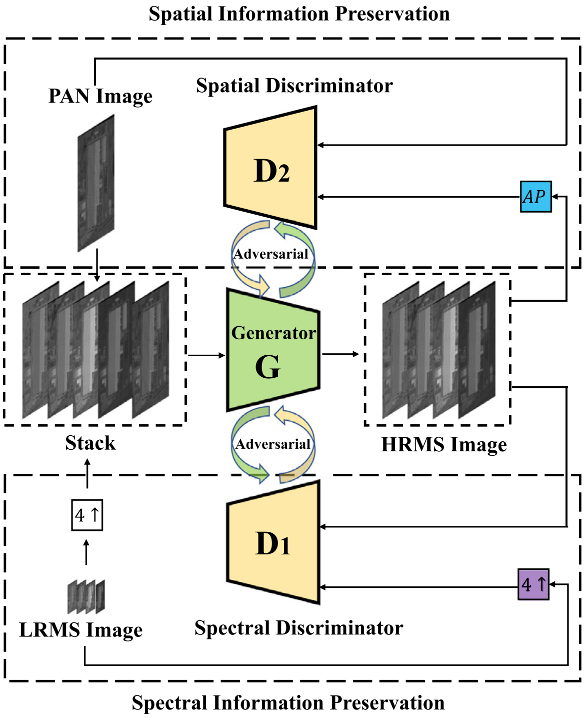

Fig. 2. Framework of our Pan-GAN for pan-sharpening. 4↑ represents upsampling the original multi-spectral image to the same resolution as the panchromatic image, and *AP* represents implementing the average pooling operation for the generated image to convert it into a single channel. The arrows between the generator and discriminators represent the adversarial relationships.

adversarial strategy to solve it, which is schematically illustrated in Fig. 2. Without loss of generality, all LRMS images used in this paper have 4 spectral bands. To begin with, we interpolate the LRMS image into the same resolution as the panchromatic image, and the next step is to stack them at the channel dimension. In particular, the first channel of the stacked image is the panchromatic image, and the rest channels correspond to the interpolated LRMS image. Then, the stacked images are fed into the generator G, and the output of G is the pan-sharpening image, *i.e.*, an HRMS image. However, the generated result only guided by the loss function (will be discussed later) without the two discriminators always tends to have severe spectral distortion or lack of spatial information, which cannot strike a balance between the spectral and spatial information.

To overcome this challenge, we regard preserving the spectral information of the LRMS image and maintaining the spatial information of the panchromatic image as two tasks, which we can utilize two discriminators to deal with, respectively. The First discriminator D_1 named spectral discriminator, aims to force the spectral information of the generated image to be consistent with that of the LRMS image. We first interpolate the generated LRMS image to be of the same resolution as the generated HRMS image, then input the HRMS image and upsampled LRMS image into D_1, which can also make the spectral distribution of the generated HRMS image consistent with that of the original LRMS image. The second discriminator D_2, named spatial discriminator, aims to force the spatial information of the generated image to be consistent with that of the panchromatic image. We implement the average pooling for the HRMS image generated by the generator along the channel dimension to get the image in a single channel. Then we input this single-channel image and the panchromatic image into D_2, which will make the spatial distribution of the generated HRMS image consistent with that of the original panchromatic image. During the training process, once these two discriminators cannot distinguish their inputs, we can then obtain the desirable HRMS image. In addition, in order to maintain the spectral information better, we perform a histogram matching between the results and the interpolated LRMS. More concretely, the histogram of the fused images should be as similar as that of the interpolated LRMS.

3.2. Loss functions

Our Pan-GAN consists of three parts, *i.e.*, a generator and two discriminators (including the spatial discriminator and the spectral discriminator). Next, we introduce them separately.

3.2.1. Loss function of generator

The loss function of our generator G can be defined as follows:

$$\mathcal{L}_G = \mathcal{L}_{\text{spectral}} + \mathcal{L}_{\text{spatial}}, \tag{4}$$

where $\mathcal{L}_G$ denotes the total loss of G. The first term $\mathcal{L}_{\text{spectral}}$ on the right hand represents the spectral loss between the spectral information of the generated HRMS image and that of the original LRMS image, and the definition is as follows:

$$\mathcal{L}_{\text{spectral}} = \frac{1}{N}\sum_{n=1}^{N}\left\| \downarrow I_f^{(n)} - I_{\text{ms}}^{(n)} \right\|_F^2 + \alpha\mathcal{L}_{\text{adv1}}, \tag{5}$$

where $I_f^{(n)}$ denotes the generated HRMS image of our generator with $n \in \mathbb{N}_N$ and N being the number of training data, $\downarrow I_f^{(n)}$ stands for downsampling the generated image to be the same resolution as the LRMS image, $I_{\text{ms}}^{(n)}$ denotes the original LRMS image, $\|\cdot\|_F$ stands for the matrix Frobenius norm, and α is a regularization parameter used to strick a balance between the two terms. The first term on the right hand in $\mathcal{L}_{\text{spectral}}$ which we call basic loss aims to keep the spectral information in the LRMS image. However, only relying on interpolation cannot represent the relation between LRMS and HRMS images. Therefore, we introduce the second term in $\mathcal{L}_{\text{spectral}}$, which can be defined as follows:

$$\mathcal{L}_{\text{adv1}} = \frac{1}{N}\sum_{n=1}^{N}(D_1(I_f^{(n)}) - c)^2, \tag{6}$$

where D_1 represents the spectral discriminator, and c denotes the value that the generator wants the spectral discriminator to believe for fake data. This term actually measures the spectral information diversity between the generated HRMS image and the LRMS image, which is also called spectral adversarial loss. It will be discussed in detail in the loss function of discriminator.

The second term $\mathcal{L}_{\text{spatial}}$ in $\mathcal{L}_G$ denotes the spatial loss between the spatial information of the generated HRMS image and that of the original panchromatic image, which is defined as follows:

$$\mathcal{L}_{\text{spatial}} = \mu\frac{1}{N}\sum_{n=1}^{N}\left\| \nabla AP(I_f^{(n)}) - \nabla I_{\text{pan}}^{(n)} \right\|_F^2 + \beta\mathcal{L}_{\text{adv2}}, \tag{7}$$

where $I_{\text{pan}}^{(n)}$ denotes the original panchromatic image, ∇ denotes the gradient operator to extract high frequency spatial information, μ is a regularization parameter which is set to balance loss between spectral and spatial information (this term we also call basic loss), β is a regularization parameter used to strick a balance between the two terms, and $AP(\cdot)$ represents the average pooling function along the channel dimension. However, the spatial information cannot be totally represented by gradients. Therefore, we also similarly add the second term to fill this gap, which can be written in the following form:

$$\mathcal{L}_{\text{adv2}} = \frac{1}{N}\sum_{n=1}^{N}(D_2(AP(I_f^{(n)})) - d)^2, \tag{8}$$

where D_2 denotes the spatial discriminator, and d represents the value that the generator wants the spatial discriminator to believe for fake data. This term is also called spatial adversarial loss, which will be explained in detail later.

3.2.2. Loss function of discriminator

Actually, there are two discriminators in our Pan-GAN: one is used for spectral preservation, and the other is for spatial preservation. Their loss functions can be uniformly defined as follows:

$$\mathcal{L}_{\mathrm{D}} = \frac{1}{N}\sum_{n=1}^{N}(D(I^{(n)}) - b)^2 + \frac{1}{N}\sum_{n=1}^{N}(D(I_f^{(n)}) - a)^2, \tag{9}$$

where $I^{(n)}$ denotes the target image whose distribution we want to fit, a and b respectively denote the labels of the target image $I^{(n)}$ and the generated HRMS image $I_f^{(n)}$, $D(I^{(n)})$ and $D(I_f^{(n)})$ respectively denote the classification results of the target image and the generated HRMS image. We adopt the least square loss [49] as the loss function in this paper. It obeys minimizing the Pearson χ^2 divergence, making the training process more steady and converge quickly.

To preserve the spectral information, we set $D = D_1$ and $I^{(n)} = \uparrow I_{\mathrm{ms}}^{(n)}$. In other words, our spectral discriminator is designed to distinguish the generated HRMS images from the interpolated LRMS images. Based on the assumption that the spectral information distribution will not change with scales, we also enforce the spectral information of the generated HRMS image to have similar distribution to the LRMS image. Once the spectral discriminator cannot distinguish $I_f^{(n)}$ and $\uparrow I_{\mathrm{ms}}^{(n)}$ during the adversarial procedure, we reach our goal.

To preserve the spatial information, we set $D = D_2$, $I_f^{(n)} = AP(I_f^{(n)})$ and $I^{(n)} = I_{\mathrm{pan}}^{(n)}$, and our spatial discriminator is designed to distinguish the generated average pooling HRMS images from the original panchromatic images. The spatial information cannot be only represented by gradients, and they can obey a specific distribution. Once the spatial discriminator cannot distinguish $AP(I_f^{(n)})$ and $I_{\mathrm{pan}}^{(n)}$, our generated HRMS image can then well preserve the spatial information of the original panchromatic image.

3.3. Network architectures

Our network architecture includes the generator, spectral discriminator and spatial discriminator. Architecture of them are all designed based on the CNN, which are shown in Fig. 3.

3.3.1. Network architecture of generator

There are several CNN architectures that can be chosen to implement our generator, for example, residual network [16], which has been used in PanNet. In this paper, we adopt the architecture of PNN, which is originally used for image super-resolution [54]. Compared with residual network, the architecture of PNN is more simple and easy to train. Accordingly, our generator architecture has three convolution layers with filter sizes of 9 × 9, 5 × 5, and 5 × 5, respectively. The stride is set to 1 with padding, and the numbers of extracted feature maps in each layer are set to 64, 32 and 4, respectively. In order to speed up the convergence of model training and make it more steady, we obey the rule designed by DCGAN [47], *i.e.*, all activation functions are ReLU, except that the last layer is tanh. We employ the batch normalization except the last layer, which can overcome the sensitivity to data initialization and avoid the problem of vanishing gradient. In addition, we also update the PNN architecture by adding some skip connections inspired by the DenseNet [55]. These skip connections can transfer more details to the later layers to make full use of the valid information and make our training process efficient. Experimental results also verify the validity of these skip connections.

It should be noted that the image scale in our generator remains the same in different layers, which is different from the original GAN. Existing works usually use the architecture of encoder and decoder for the generator. However, the encoder needs to downsample the image, which may lose some important information of the original image. Therefore, we avoid this operation in our generator.

3.3.2. Network architecture of discriminator

Although our Pan-GAN consists of two discriminators, *i.e.*, the spectral discriminator and the spatial discriminator, they have the same structure with different inputs. We use fully convolution neural networks for our discriminators, and each of them consists of six layers. The filter size of the first five layers is 3 × 3, and that of the last layer is 4 × 4. The stride of the first five layers is set to 2, and the last one is set to 1. The number of extracted feature maps in different layers are set to 16, 32, 64, 128, 256 and 1, respectively. In addition, we also obey the rules proposed by DCGAN, except for the first layer, *i.e.*, the batch normalization and leaky ReLU are used as activation function in the other

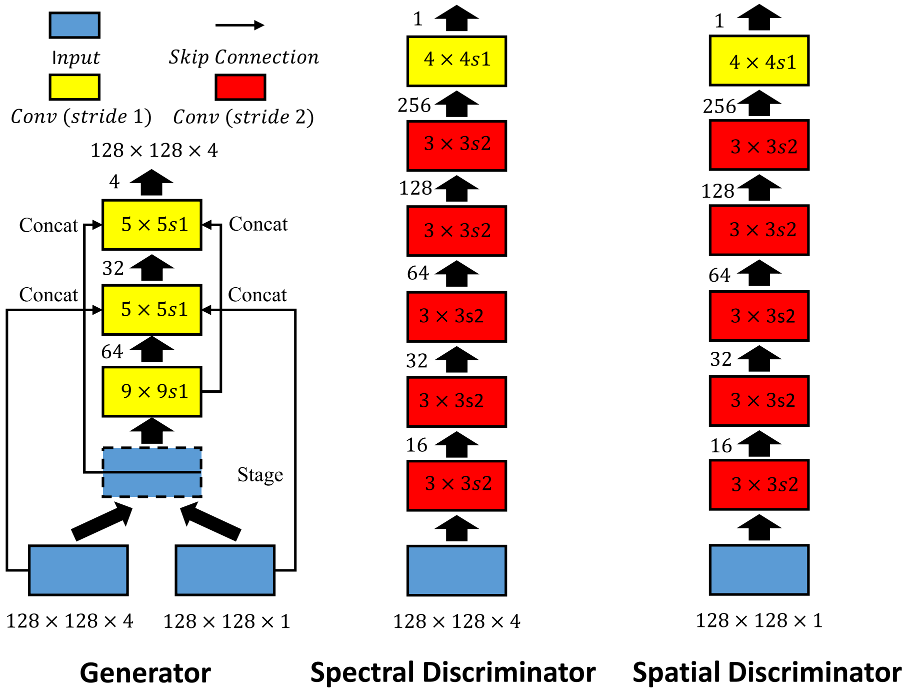

Fig. 3. Network architecture of the generator, spectral discriminator and spatial discriminator. The numbers of feature maps are shown on the left, and the sizes of filters are given on the right.

five layers. For the spectral discriminator, the input is the generated HRMS image or the interpolated LRMS image. For the spatial discriminator, the input is the original panchromatic image or the single channel image generated by the generated HRMS image after average pooling along the channel dimension. The output of discriminators is both the classification results.

3.4. Training details

Our training data consists of the original LRMS image with resolution of 32 × 32 and the original panchromatic image with resolution of 128 × 128. The size of batch images is set to 32. The number of training steps is set to 100,000 (the epoch is set to 53). μ is set to 5, α is set to 0.002, and β is set to 0.001. The initialized learning rate is set to 0.0001 with decay rate setting to 0.99, and the decay step is set to 10,000. Our optimizer solver is the RMSProp optimizer [56]. To speed up the training process and make our model steady, on the one hand, different from traditional training strategy, we first train the generator and the spectral discriminator. When the spectral discriminator gets convergency, the spatial discriminator is added into our model, and we train them simultaneously. On the other hand, the labels *a, b, c* and *d* are not fixed, which are also called soft labels. For the discriminator, the label *a* of the generated HRMS image is a random number ranging from 0 to 0.3, while the label *b* of the target image is a random number ranging from 0.7 to 1.2. For the generator, labels *c* and *d* are random numbers ranging from 0.7 to 1.2.

4. Experiments and evaluations

Note that our model consists of a generator and two discriminators, *i.e.*, a spatial discriminator and a spectral discriminator, and the loss function of our Pan-GAN is constructed by using a bunch of losses including two basic losses and two adversarial losses, we first conduct ablation study which we also call analysis of different network structures on different loss combinations to verify the validity of spatial discriminator and spectral discriminator. Then, several comparative experiments are conducted respectively on datasets from WorldView II and GaoFen-2 (GF-2) satellites to demonstrate the superiority of our Pan-GAN. We use seven state-of-the-art pan-sharpening methods for qualitative and quantitative comparison, including P + XS [13], MTF-GLP [10], BDSD [14], SIRF [57], PNN [19], PSGAN [21] and PanNet [20]. The codes of all these compared methods are publicly available, and we set the parameters of these methods according to the original papers. The experiments are conducted on a desktop with 2.4 GHz Intel Xeon CPU E5-2673 v3, GeForce GTX 1080Ti, and 64 GB memory.

4.1. Datasets and metrics

There are two datasets used in our experiments, including WorldView II and GF-2. The spatial resolutions of panchromatic images in these two satellites are 0.5m and 0.8m, respectively, while the spatial resolutions of their corresponding LRMS images are 1.8 m and 3.2 m with four bands including red, green, blue and near-infrared. We crop the panchromatic and LRMS images into 60,000 image patch pairs of sizes 128 × 128 and 32 × 32, respectively, and then randomly split them into 90% and 10% as our training data and validation data, respectively.

During the testing phase, we can directly feed the entire source images into the generator to produce the results without need for cropping them into small patches with the same size as the training data because the generator of Pan-GAN is fully convolutional networks (FCN). FCN is different from classic CNN in that classic CNN uses fully connected layers to get fixed length eigenvectors after convolution layers while FCN does not. Thus, FCN can accept input images of any size. In order to fully illustrate the effectiveness of our method, we conduct two types of testing: testing under Wald's protocol and full-resolution testing. It is difficult to get a comprehensive evaluation of different methods only by subjective judgement. Thus we introduce six widely used quantitative metrics to characterize the fusion performance, *i.e.*, relative dimensionless global error in synthesis (ERGAS) [58], root mean squared error (RMSE), relative average spectral error (RASE) [59], the filtered correlation coefficients (FCC) [60], quality with no reference (QNR) [61] and generalized QNR (GQNR) [62].

The first three metrics evaluate the similarity or the error between the fused result and the ground truth. More concretely, ERGAS is a global quality metric to measure mean shifting and dynamic range change [63]. RMSE is used to measure the changes of the pixel values between the ground truth and the fused image. RASE is used to measure the spectral quality by computing the relative error between the fused image and the multi-spectral image. As for the rest three metrics FCC, QNR and GQNR, they are three approaches to assess the pan-sharpening performance without reference by checking main properties of the fused results. FCC uses the PAN image to evaluate the spatial quality of the fused image. QNR uses the original LRMS and PAN images to measure the spectral distortion between the resampled LRMS images and the fused images, and the spatial distortion caused by discrepancies in spatial details originated by fusion. Based on QNR, GQNR is applicable to scenarios especially when the panchromatic band does not overlap the short-wave infrared bands. Generally speaking, larger values of FCC and QNR indicate better performance, and smaller values of ERGAS, RMSE, RASE and GQNR indicate better performance.

In order to fully illustrate the effectiveness of our method, when performing comparative experiments, we perform two types of testing, *i.e.*, testing under the Wald's protocol where original HRMS images are downsampled to LRMS images and then HRMS images are used as references and testing on full-resolution images. Under the Wald's protocol, the metrics we used are ERGAS, RMSE and RASE. When testing on full-resolution images, we use the panchromatic images and the original LRMS images as the test data. Since there is no ground truth, we use metrics FCC, QNR and GQNR mentioned above to evaluate the results.

4.2. Ablation study

Ablation study on different loss combinations is conducted in this section. Considering whether spectral adversarial loss ($\mathcal{L}_{\text{spectral}}$) or spatial adversarial loss ($\mathcal{L}_{\text{spatial}}$) being optimized, we divide our model structures into 4 types, *i.e.* Model-Generator with only generator preserved and only basic loss optimized, Model-Spatial with spatial discriminator preserved and basic loss and $\mathcal{L}_{\text{spatial}}$ optimized, Model-Spectral with spectral discriminator preserved and basic loss and $\mathcal{L}_{\text{spectral}}$ optimized, and our Pan-GAN with $\mathcal{L}_{G}$ optimized. The four structures are described in Fig. 4.

We train these four models on WorldView II and test them under the Wald's protocol. We tune the parameters so that all the models can obtain their best performance. In particular, μ is set to 5, α is set to 0.002, and β is set to 0.001 with the other training setting being the same. Fusion results of these four structures are presented in Fig. 5, where the HRMS image is the ground truth, and the LRMS image is interpolated into the same resolution as the panchromatic image using bicubic interpolation. From the results, we observe that the highlighted region in the LRMS is rather blurry, where the pipelines cannot be distinguished which are visible in the panchromatic image and all the fusion results. This indicates that the most of spatial information can be preserved by only optimizing the basic loss. However, it is clear that the fusion result of Model-Generator show a different color distribution with the original LRMS, whereas the other three model structures do not suffer from this problem. This demonstrates that the adversarial loss in $\mathcal{L}_{\text{spectral}}$ and $\mathcal{L}_{\text{spatial}}$ can improve the spectral information preservation.

To get a more accurate assessment for spectral distortion and spatial information loss, we compute the residual images between HRMS (ground truth) and down-sampled fusion results for analyzing the spectral distortion (as shown in Fig. 6), and compute the gradient residual images between gradient of panchromatic image and gradients of

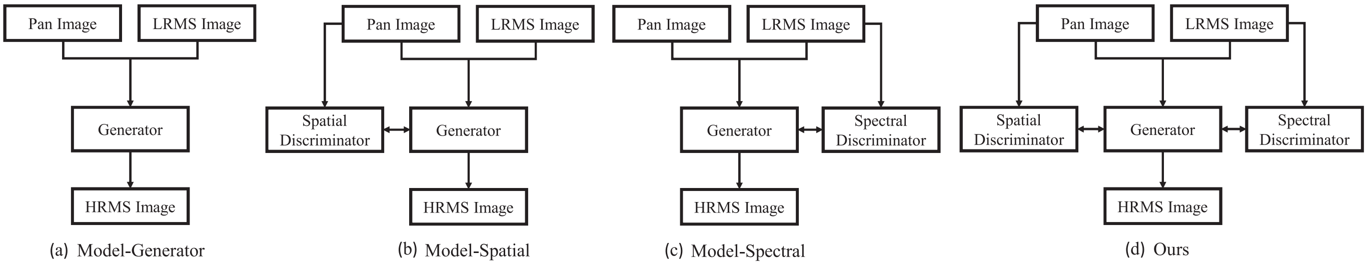

Fig. 4. The four different network structures of our proposed method for ablation study. (a) Model-Generator. (b) Model-Spatial. (c) Model-Spectral. (d) Our Pan-GAN.

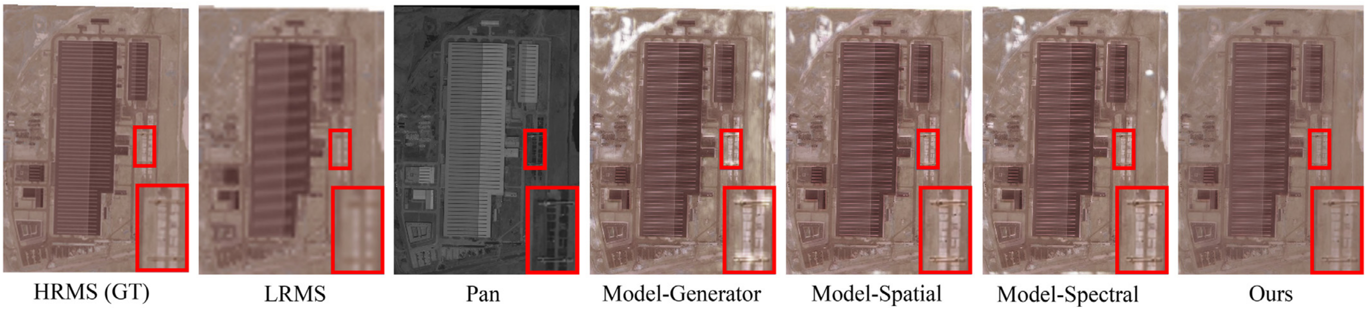

Fig. 5. Fusion result of four different structures. From left to right, interpolated LRMS image, original panchromatic image, fusion result of Model-Generator, Model-Spatial, Model-Spectral, and our Pan-GAN.

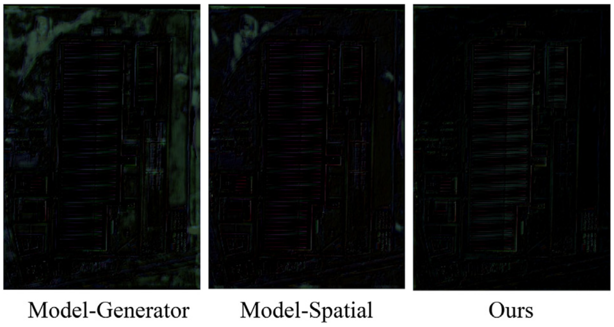

Fig. 6. Spectral distortion analysis. Residual images between HRMS and fusion results of three model structures including Model-Generator, Model-Spatial and our Pan-GAN. We multiply the pixel values in each error image by 4 for clear comparison of visual effects. Please zoom in to see the details.

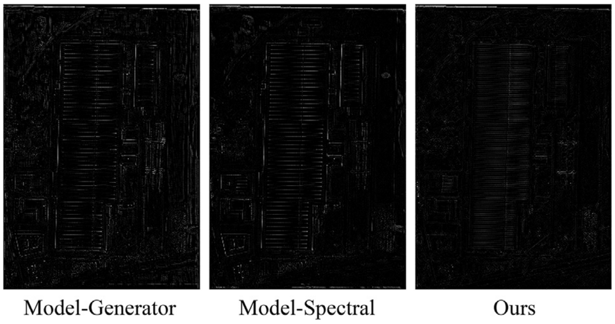

Fig. 7. Spatial loss analysis. Residual images between gradient of panchromatic image and gradients of fusion results of three model structures including Model-Generator, Model-Spectral and our Pan-GAN. We multiply the pixel values in each error image by 4 for clear comparison of visual effects. Please zoom in to see the details.

fusion results for analyzing the spatial information loss (as shown in Fig. 7). From Fig. 6, we see that the Model-Generator clearly has the largest spectral distortion, where the smooth regions in the residual image tend to have lots of shadow. For the fusion result of Model-Spatial without spectral discriminator, there exist some regions (*e.g.*, top left and top right corners) still suffering from large spectral distortion although most regions can well preserve the spectral information. But compared with the Model-Generator, it exhibits much less spectral distortion. This demonstrates that the panchromatic image can not only provide the spatial information, but also improve the spectral quality. In contrast, the fused image of our Pan-GAN has much smaller residuals and we can hardly find regions suffering from obvious spectral distortion. This demonstrates that the spectral discriminator plays an important role in preserving the spectral information during the adversarial learning. Note that there are some traces in the residual images of Model-Spatial and our Pan-GAN. From Fig. 6, we see that with the spectral discriminator and spatial discriminator gradually being added, the gradient residual image gets better and better.

Further, we provide a quantitative comparison of these four structures on WorldView II. The statistical results of three full-reference metrics are summarized in Table 1. The average metric values and the corresponding variances are all listed in the table, while the ideal value of each metric is also given in the last row. From the results, we see that our Pan-GAN is able to achieve the best average values in all cases.

In conclusion, only using generator will lead to several spectral distortion; the spatial and spectral discriminators can respectively improve the spatial information preservation and avoid spectral distortion; by applying these two discriminators simultaneously, we can get the best fusion result.

4.3. Comparative experiments

In this section, we demonstrate the efficiency of the proposed method on WorldView II and GF-2 and compare it with seven state-of-the-art pan-sharpening methods such as P + XS [13], MTF-GLP [10], BDSD [14], SIRF [57], PNN [19], PSGAN [21] and PanNet [20].

Table 1
Qualitative comparison of four different structures on 120 satellite images from the WorldView II dataset. The runtime is GPU time. **Bold** indicates the best result.

Method	ERGAS	RMSE	RASE	Runtime (s)
Model-Generator	4.899 ± 1.230	13.617 ± 4.258	18.448 ± 4.202	**0.025**
Model-Spatial	4.227 ± 0.979	11.749 ± 3.130	16.069 ± 3.656	**0.025**
Model-Spectral	4.577 ± 1.019	12.744 ± 3.239	17.433 ± 3.708	**0.025**
Ours	**2.766 ± 0.466**	**8.000 ± 1.086**	**11.078 ± 1.896**	**0.025**
Desired value	0	0	0	-

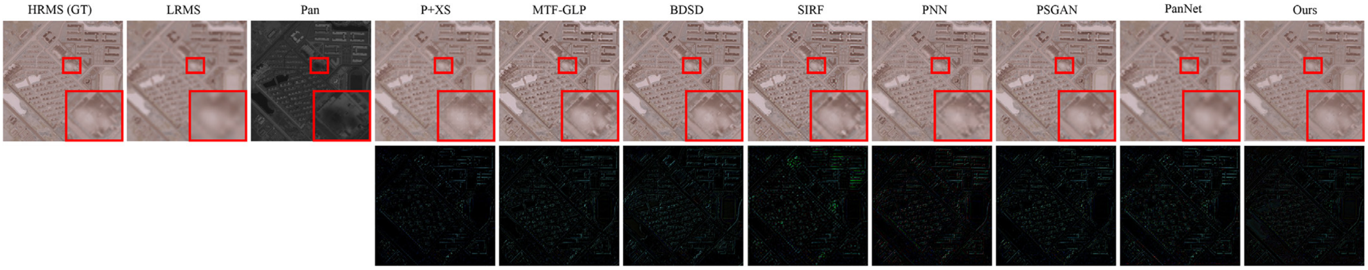

Fig. 8. Qualitative comparison of different methods for pan-sharpening under the reduced-resolution test on the data from WorldView II. The size of the PAN image is 304 × 304. The first row is the fusion results and the second row is the corresponding error images of different methods. We multiply the pixel values in each error image by 4 for clear comparison of visual effects. Please zoom in to see the details.

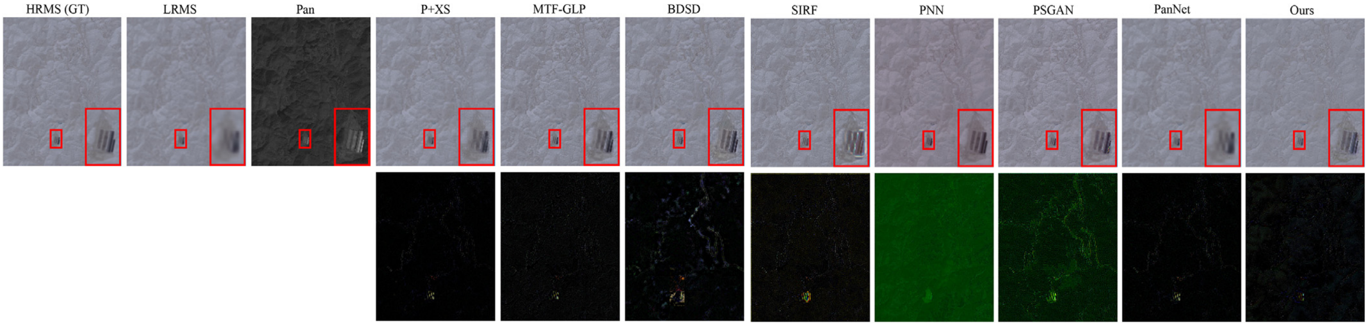

Fig. 9. Qualitative comparison of different methods for pan-sharpening under the reduced-resolution test on the data from GF-2. The size of the PAN image is 500 × 400. The first row is the fusion results and the second row is the corresponding error images of different methods. We multiply the pixel values in each error image by 4 for clear comparison of visual effects. Please zoom in to see the details.

4.3.1. *Qualitative comparison*

As mentioned earlier, we conduct two types of comparison experiments, *i.e.*, testing under the Wald's protocol and testing on full-resolution images.

First, we give the comparative results under the Wald's protocol, where original HRMS images are downsampled to LRMS images and then HRMS images are used as references, as shown in Figs. 8 and 9. The first row shows the original HRMS (GT), LRMS and panchromatic images, as well as the fusion results of the eight methods. The second row shows the corresponding error images between HRMS and fusion results of different methods. In Fig. 8, cars under the shadows are blurred in LRMS but visible in panchromatic images as highlighted. Among these methods, only the results of MTF-GLP, BDSD and our method can clearly retain this detail, while that of other methods are missing or weak. However, there are some color distortions in the results of MTF-GLP and BDSD compared to Pan-GAN. As a result, our method not only maintains the spectral distribution, but also better preserves the spatial details, which is consistent with the error images on the second row. The results in Fig. 9 are similar to Fig. 8. For the houses in the highlighted part, our method is the best in both the spectrum distribution and the spatial details.

Second, We present the qualitative comparison of different methods, which is tested on full-resolution images. The results are shown in Figs. 10 and 11. The first row shows the original LRMS and panchromatic images, as well as the fusion results of the eight methods. The second row shows the corresponding gradient error images between the panchromatic image and fusion results. In Fig. 10, in terms of visual perception, there is some spectral distortion in the results of PNN and PSGAN, while other methods can well maintain spectral color distribution. However, it shows great diversity for the spatial information of different methods. For example, the highlighted part contains a strip-shaped object, which is visible in the panchromatic image but cannot be found in the LRMS image due to the limited spatial resolution. For P + XS, SIRF, PNN, PSGAN and PanNet, this information is lost or weakened in their results. Only MTF-GLP, BDSD and our method can better maintain the details well, which are clear as that in the panchromatic image. For better visualization, we compute the error images to compare the spatial information difference (represented by gradient) between the fused image and panchromatic image. From the results, we see that the spatial loss of our Pan-GAN is much smaller. In Fig. 11, we can draw a similar conclusion as that in Fig. 10. In particular, the pool highlighted in the red box region is textured, which is very blurry in the results of the competitors. Nevertheless, such information in the result of our Pan-GAN is shown clearly with high contrast and clear edges. The spatial error images also demonstrate that our fusion result can well preserve the spatial information. In addition, it is also worth noting that our Pan-GAN can also perform very well in maintaining the spectral information, and there is no distortion of color distribution in the intuitive sense.

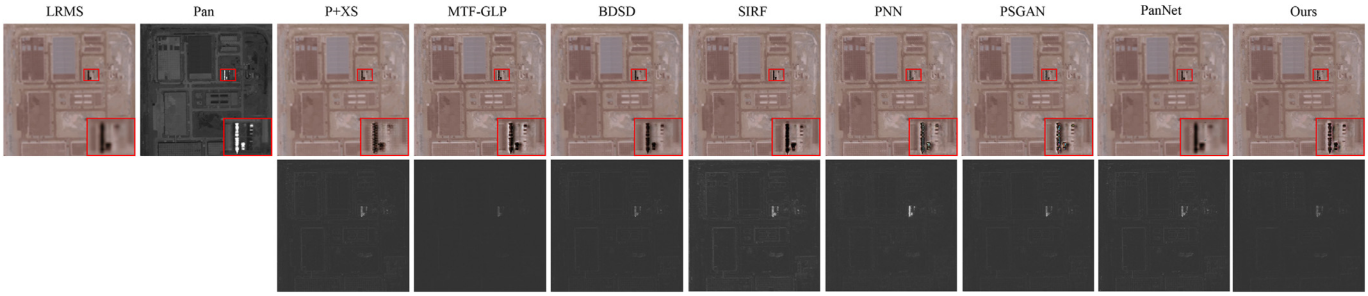

Fig. 10. Qualitative comparison of different methods for pan-sharpening under the full-resolution test on the data from WorldView II. The size of the PAN image is 624 × 624. The first row is the fusion results and the second row is corresponding gradient error images of different methods. We multiply the pixel values in each error image by 4 for clear comparison of visual effects. Please zoom in to see the details.

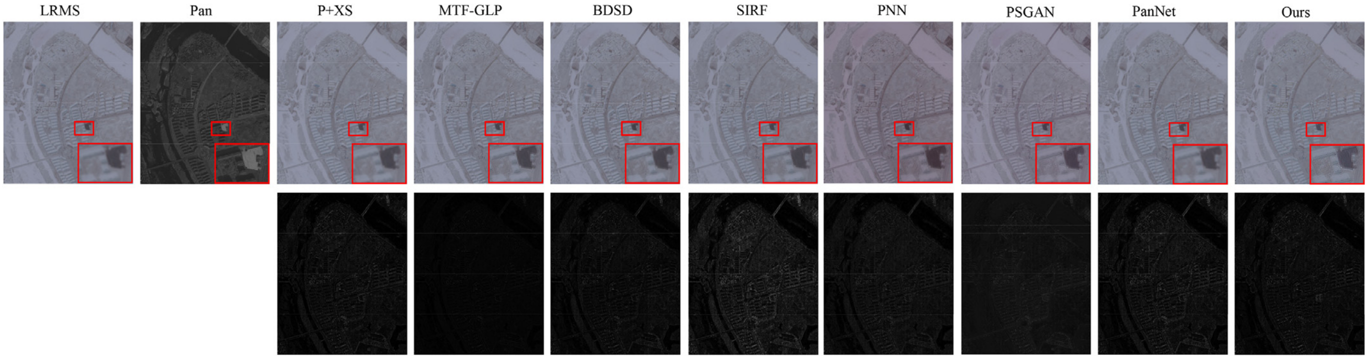

Fig. 11. Qualitative comparison of different methods for pan-sharpening under the full-resolution test on the data from GF-2. The size of the PAN image is 500 × 400. The first row is the fusion results and the second row is the corresponding gradient error images of different methods. We multiply the pixel values in each error image by 4 for clear comparison of visual effects. Please zoom in to see the details.

Table 2
Quantitative comparison of eight methods on 120 images from the WorldView II dataset. All CNN-based methods are performed on GPU, while other methods are performed on CPU. **Bold** indicates the best result.

Method	ERGAS	RMSE	RASE	FCC	QNR	GQNR	Runtime (s)
P+XS [13]	2.910 ± 0.504	6.553 ± 1.202	12.039 ± 1.413	0.815 ± 0.034	0.899 ± 0.037	0.047 ± 0.018	172.568
MTF-GLP [10]	3.122 ± 0.606	7.150 ± 1.358	13.183 ± 1.912	0.937 ± 0.028	0.950 ± 0.032	**0.014 ± 0.013**	0.543
BDSD[14]	3.283 ± 0.691	7.438 ± 1.662	13.685 ± 2.382	0.921 ± 0.046	0.948 ± 0.035	0.062 ± 0.033	0.018
SIRF [57]	4.366 ± 0.852	10.114 ± 2.638	18.756 ± 4.746	0.275 ± 0.186	0.781 ± 0.049	0.041 ± 0.056	71.246
PNN [20]	2.915 ± 0.555	6.349 ± 1.339	11.668 ± 1.768	0.676 ± 0.064	0.839 ± 0.04	0.103 ± 0.036	0.028
PSGAN [21]	2.836 ± 0.561	6.318 ± 1.404	11.595 ± 1.952	0.768 ± 0.037	**0.952 ± 0.033**	0.031 ± 0.014	**0.011**
PanNet [19]	2.841 ± 0.564	6.078 ± 1.083	11.254 ± 1.651	0.502 ± 0.115	**0.952 ± 0.034**	0.017 ± 0.012	**0.011**
Pan-GAN	**2.631 ± 0.490**	**5.661 ± 1.208**	**10.406 ± 1.656**	**0.960 ± 0.012**	**0.952 ± 0.036**	0.020 ± 0.012	0.025
Desired value	0	0	0	1	1	0	-

Table 3
Quantitative comparison of eight methods on 120 images from the GF-2 dataset. All CNN-based methods are performed on GPU, while other methods are performed on CPU. **Bold** indicates the best result.

Method	ERGAS	RMSE	RASE	FCC	QNR	GQNR	Runtime (s)
P+XS [13]	1.647 ± 0.249	3.177 ± 0.213	4.742 ± 0.338	0.835 ± 0.018	0.958 ± 0.015	0.041 ± 0.021	191.694
MTF-GLP [10]	1.627 ± 0.333	3.161 ± 0.412	4.719 ± 0.626	0.896 ± 0.014	**0.973 ± 0.018**	**0.038 ± 0.020**	1.943
BDSD[14]	1.925 ± 0.322	3.804 ± 0.396	5.674 ± 0.573	0.867 ± 0.025	0.965 ± 0.018	0.072 ± 0.056	0.018
SIRF [57]	4.139 ± 0.943	12.233 ± 3.392	18.263 ± 5.036	0.725 ± 0.085	0.844 ± 0.044	0.066 ± 0.051	82.867
PNN [19]	4.274 ± 0.883	11.648 ± 1.880	17.379 ± 2.782	0.666 ± 0.034	0.947 ± 0.01	0.113 ± 0.015	**0.011**
PSGAN [21]	2.165 ± 0.237	5.164 ± 0.555	7.701 ± 0.787	0.796 ± 0.021	0.958 ± 0.014	0.080 ± 0.050	**0.011**
PanNet [20]	1.786 ± 0.244	3.816 ± 0.399	5.699 ± 0.652	0.520 ± 0.046	0.825 ± 0.045	0.168 ± 0.149	0.028
Pan-GAN	**1.543 ± 0.220**	**3.047 ± 0.365**	**4.549 ± 0.547**	**0.906 ± 0.010**	0.963 ± 0.016	0.102 ± 0.066	0.025
Desired value	0	0	0	1	1	0	-

4.3.2. *Quantitative comparison*

We further provide quantitative comparisons of the seven methods on the two datasets from WorldView II and GF-2. On the one hand, for the metrics which need the ground-truth data, *i.e.*, ERGAS, RMSE and RASE, we downsample the source images into images with a lower resolution and use the original HRMS images as the ground-truth data for calculation. On the other hand, for the no-reference metrics, *i.e.*, FCC, QNR and GQNR, considering that these metrics do not need reference images and to keep the major advantage of unsupervised methods, we calculate these metrics on full-resolution images. The statistical results of the six metrics and runtime are shown in Tables 2 and 3.

From the results, we see that our Pan-GAN is able to achieve much better average values than all the other competitors on all the three full-reference metrics on both datasets. These three metrics can show that our Pan-GAN can generate the closest fusion result to HRMS (the ground truth), no matter from the spectral information or the spatial information. For the no-reference metrics, our method achieves the best performance in terms of FCC. This shows that our method can well maintain the spatial information of panchromatic images. Relatively speaking, on QNR, our Pan-GAN also performs well, which ranks the first on the WorldView II dataset and third on the GF-2 dataset. On GQNR, our method can also achieve comparable performance. Moreover, we also report the time consumption of different methods on two datasets in the last column of the two tables. Our methods can achieve comparable efficiency compared with the other competitors.

In conclusion, our Pan-GAN can well preserve the basic spectral information of original LRMS image and spatial information of original panchromatic image without any ground truth as supervision. Compared with other state-of-the-art methods, our methods keep more clearly detail information (*i.e.,* spatial information) with few spatial information loss. No matter qualitative comparison or quantitative comparison, our proposed method can always generate satisfying performance.

5. Conclusion

In this paper, we propose an unsupervised pan-sharpening framework called Pan-GAN for multi-spectral image and panchromatic image fusion based on the generative adversarial network. Our method formulates pan-sharpening as a multi-task problem, that is, the preservation of spectral and spatial information. We use the same generator to establish adversarial games with the spectral discriminator and the spatial discriminator, respectively, so as to preserve the spectral information of the multi-spectral image and the spatial information of the panchromatic image simultaneously. Experiments on the WorldView II and GF-2 datasets demonstrate that our Pan-GAN can not only obtain the best evaluation metrics but also preserve the basic spectral information distribution well with few loss of spatial information.

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.

CRediT authorship contribution statement

Jiayi Ma: Writing - review & editing, Visualization, Methodology, Formal analysis. **Wei Yu:** Writing - review & editing, Visualization, Investigation, Methodology, Formal analysis. **Chen Chen:** Validation, Resources. **Pengwei Liang:** Investigation. **Xiaojie Guo:** Validation, Resources. **Junjun Jiang:** Methodology, Formal analysis.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant nos. 61773295, 61772512 and 61971165, and in part by the Natural Science Fund of Hubei Province under Grant no. 2019CFA037.

References

[1] J. Ma, H. Zhou, J. Zhao, Y. Gao, J. Jiang, J. Tian, Robust feature matching for remote sensing image registration via locally linear transforming, IEEE Trans. Geosci. Remote Sens. 53 (12) (2015) 6469–6481.

[2] Z. Shao, J. Cai, P. Fu, L. Hu, T. Liu, Deep learning-based fusion of landsat-8 and sentinel-2 images for a harmonized surface reflectance product, Remote Sens. Environ. 235 (2019) 111425.

[3] C. Thomas, T. Ranchin, L. Wald, J. Chanussot, Synthesis of multispectral images to high spatial resolution: a critical review of fusion methods based on remote sensing physics, IEEE Trans. Geosci. Remote Sens. 46 (5) (2008) 1301–1312.

[4] H. Ghassemian, A review of remote sensing image fusion methods, Inf. Fusion 32 (2016) 75–89.

[5] R. Haydn, G.W. Dalke, J. Henkel, J. Bare, Application of the IHS color transform to the processing of multisensor data and image enhancement, in: Proceedings of the International Symposium on Remote Sensing of Environment, 1982.

[6] B. Aiazzi, S. Baronti, M. Selva, Improving component substitution pansharpening through multivariate regression of ms + pan data, IEEE Trans. Geosci. Remote Sens. 45 (10) (2007) 3230–3239.

[7] 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.

[8] P. Burt, E. Adelson, The laplacian pyramid as a compact image code, IEEE Trans. Commun. 31 (4) (1983) 532–540.

[9] N. Yokoya, T. Yairi, A. Iwasaki, Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion, IEEE Trans. Geosci. Remote Sens. 50 (2) (2011) 528–537.

[10] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, M. Selva, Mtf-tailored multiscale fusion of high-resolution ms and pan imagery, Photogramm. Eng. Remote Sens. 72 (5) (2006) 591–596.

[11] S.A. Valizadeh, H. Ghassemian, Remote sensing image fusion using combining IHS and curvelet transform, in: Proceedings of the International Symposium on Telecommunications, 2012, pp. 1184–1189.

[12] C. Kwan, J. Choi, S. Chan, J. Zhou, B. Budavari, A super-resolution and fusion approach to enhancing hyperspectral images, Remote Sens. 10 (9) (2018) 1416.

[13] 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.

[14] A. Garzelli, F. Nencini, L. Capobianco, Optimal MMSE pan sharpening of very high resolution multispectral images, IEEE Trans. Geosci. Remote Sens. 46 (1) (2007) 228–236.

[15] Y. Liu, X. Chen, Z. Wang, Z.J. Wang, R.K. Ward, X. Wang, Deep learning for pixel-level image fusion: recent advances and future prospects, Inf. Fusion 42 (2018) 158–173.

[16] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: Single shot multibox detector, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 21–37.

[18] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[19] G. Masi, D. Cozzolino, L. Verdoliva, G. Scarpa, Pansharpening by convolutional neural networks, Remote Sens. 8 (7) (2016) 594.

[20] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, J. Paisley, Pannet: a deep network architecture for pan-sharpening, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1753–1761.

[21] X. Liu, Y. Wang, Q. Liu, PSGAN: a generative adversarial network for remote sensing image pan-sharpening, in: Proceedings of the IEEE International Conference on Image Processing, 2018, pp. 873–877.

[22] J. Ma, Y. Ma, C. Li, Infrared and visible image fusion methods and applications: a survey, Inf. Fusion 45 (2019) 153–178.

[23] X. Fu, Z. Lin, Y. Huang, X. Ding, A variational pan-sharpening with local gradient constraints, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10265–10274.

[24] Y. Zhang, C. Liu, M. Sun, Y. Ou, Pan-sharpening using an efficient bidirectional pyramid network, IEEE Trans. Geosci. Remote Sens. 57 (8) (2019) 5549–5563.

[25] X. Liu, Q. Liu, Y. Wang, Remote sensing image fusion based on two-stream fusion network, Inf. Fusion 55 (2020) 1–15.

[26] Y. Liu, X. Chen, R.K. Ward, Z.J. Wang, Image fusion with convolutional sparse representation, IEEE Signal Process. Lett. 23 (12) (2016) 1882–1886.

[27] Y. Liu, X. Chen, H. Peng, Z. Wang, Multi-focus image fusion with a deep convolutional neural network, Inf. Fusion 36 (2017) 191–207.

[28] Y. Liu, S. Liu, Z. Wang, A general framework for image fusion based on multi-scale transform and sparse representation, Inf. Fusion 24 (2015) 147–164.

[29] V.K. Shettigara, A generalized component substitution technique for spatial enhancement of multispectral images using a higher resolution data set, Photogramm. Eng. Remote Sens. 58 (1992) 561–567.

[30] W. Dou, Y. Chen, X. Li, D.Z. Sui, A general framework for component substitution image fusion: an implementation using the fast image fusion method, Computers & Geosciences 33 (2) (2007) 219–228.

[31] T.-M. Tu, S.-C. Su, H.-C. Shyu, P.S. Huang, A new look at IHS-like image fusion methods, Inf. Fusion 2 (3) (2001) 177–186.

[32] P. Chavez, S.C. Sides, J.A. Anderson, et al., Comparison of three different methods to merge multiresolution and multispectral data- landsat TM and spot panchromatic, Photogramm. Eng. Remote Sens. 57 (3) (1991) 295–303.

[33] M. Ghahremani, H. Ghassemian, Remote-sensing image fusion based on curvelets and ICA, Int. J. Remote Sens. 36 (16) (2015) 4131–4143.

[34] W. Wright, Fast image fusion with a Markov random field, in: Proceedings of the International Conference on Image Processing and its Applications, 1999, pp. 557–561.

[35] M. Guo, H. Zhang, J. Li, L. Zhang, H. Shen, An online coupled dictionary learning approach for remote sensing image fusion, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7 (4) (2014) 1284–1294.

[36] J. Ma, W. Yu, P. Liang, C. Li, J. Jiang, Fusiongan: a generative adversarial network for infrared and visible image fusion, Inf. Fusion 48 (2019) 11–26.

[37] 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. (2019).

[38] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[39] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5967–5976.

[40] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2242–2251.

[41] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4681–4690.

[42] J. Ma, P. Liang, W. Yu, C. Chen, X. Guo, J. Wu, J. Jiang, Infrared and visible image fusion via detail preserving adversarial learning, Inf. Fusion 54 (2020) 85–98.

[43] J. Ma, H. Xu, J. Jiang, X. Mei, X.-P. Zhang, DDcGAN: A Dual-discriminator Conditional Generative Adversarial Network for Multi-resolution Image Fusion, IEEE Trans. Image Process 29 (2020) 4980–4995.

[44] E.L. Denton, S. Chintala, R. Fergus, et al., Deep generative image models using a laplacian pyramid of adversarial networks, in: Advances in Neural Information Processing Systems, 2015, pp. 1486–1494.

[45] K. Gregor, I. Danihelka, A. Graves, D. Rezende, D. Wierstra, Draw: a recurrent neural network for image generation, in: Proceedings of the International Conference on Machine Learning, 2015, pp. 1462–1471.

[46] A. Dosovitskiy, T. Brox, Generating images with perceptual similarity metrics based on deep networks, in: Advances in Neural Information Processing Systems, 2016, pp. 658–666.

[47] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, in: Proceedings of the International Conference on Learning Representations, 2016.

[48] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein gan, arXiv:1701.07875(2017).

[49] X. Mao, Q. Li, H. Xie, R.Y. Lau, Z. Wang, S.P. Smolley, Least squares generative adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2813–2821.

[50] I. Durugkar, I. Gemp, S. Mahadevan, Generative multi-adversarial networks, in: Proceedings of the International Conference on Learning Representations, 2017.

[51] L. Wang, V. Sindagi, V. Patel, High-quality facial photo-sketch synthesis using multi-adversarial networks, in: Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition, 2018, pp. 83–90.

[52] H. Aghakhani, A. Machiry, S. Nilizadeh, C. Kruegel, G. Vigna, Detecting deceptive reviews using generative adversarial networks, in: Proceedings of the IEEE Security and Privacy Workshops, 2018, pp. 89–95.

[53] L. Wald, T. Ranchin, M. Mangolini, Fusion of satellite images of different spatial resolutions: assessing the quality of resulting images, Photogramm. Eng. Remote Sens. 63 (6) (1997) 691–699.

[54] 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.

[55] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, K. Keutzer, Densenet: implementing efficient convnet descriptor pyramids, arXiv:1404.1869(2014).

[56] T. Tieleman, G. Hinton, Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude, COURSERA 4 (2) (2012) 26–31.

[57] C. Chen, Y. Li, W. Liu, J. Huang, Sirf: simultaneous satellite image registration and fusion in a unified framework, IEEE Trans. Image Process. 24 (11) (2015) 4213–4224.

[58] L. Alparone, L. Wald, J. Chanussot, C. Thomas, P. Gamba, L.-M. Bruce, Comparison of pansharpening algorithms: outcome of the 2006 GRS-S data fusion contest, IEEE Trans. Geosci. Remote Sens. 45 (10) (2007) 3012–3021.

[59] M. Choi, A new intensity-hue-saturation fusion approach to image fusion with a tradeoff parameter, IEEE Trans. Geosci. Remote Sens. 44 (6) (2006) 1672–1682.

[60] J. Zhou, D. Civco, J. Silander, A wavelet transform method to merge landsat tm and spot panchromatic data, Int. J. Remote Sens. 19 (4) (1998) 743–757.

[61] 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.

[62] C. Kwan, B. Budavari, A.C. Bovik, G. Marchisio, Blind quality assessment of fused worldview-3 images by using the combinations of pansharpening and hypersharpening paradigms, IEEE Geosci. Remote Sens. Lett. 14 (10) (2017) 1835–1839.

[63] Q. Du, N.H. Younan, R. King, V.P. Shah, On the performance evaluation of pan-sharpening techniques, IEEE Geosci. Remote Sens. Lett. 4 (4) (2007) 518–522.