Information Fusion 66 (2021) 40–53

Contents lists available at ScienceDirect

Information Fusion

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

Full length article

MFF-GAN: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion

Hao Zhang [a,1], Zhuliang Le [a,1], Zhenfeng Shao [b,*], Han Xu [a], Jiayi Ma [a]

[a] Electronic Information School, Wuhan University, Wuhan, 430072, China
[b] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, 430079, China

ARTICLE INFO

Keywords:
Image fusion
Multi-focus
Unsupervised learning
Generative adversarial network

ABSTRACT

Multi-focus image fusion is an enhancement method to generate full-clear images, which can address the depth-of-field limitation in imaging of optical lenses. Most existing methods generate the decision map to realize multi-focus image fusion, which usually lead to detail loss due to misclassification, especially near the boundary line of the focused and defocused regions. To overcome this challenge, this paper presents a new generative adversarial network with adaptive and gradient joint constraints to fuse multi-focus images. In our model, an adaptive decision block is introduced to determine whether source pixels are focused or not based on the difference of repeated blur. Under its guidance, a specifically designed content loss can dynamically guide the optimization trend, that is, force the generator to produce a fused result of the same distribution as the focused source images. To further enhance the texture details, we establish an adversarial game so that the gradient map of the fused result approximates the joint gradient map constructed based on the source images. Our model is unsupervised without requiring ground-truth fused images for training. In addition, we release a new dataset containing 120 high-quality multi-focus image pairs for benchmark evaluation. Experimental results demonstrate the superiority of our method over the state-of-the-art in terms of both subjective visual effect and quantitative metrics. Moreover, our method is about one order of magnitude faster compared with the state-of-the-art.

1. Introduction

Due to the limitations of optical lenses, it is difficult to have all objects of different depth-of-field to be all-in-focus within one image [1]. In this context, multi-focus image fusion as an image enhancement method can fuse images with different focused regions to obtain a single full-clear image, which has good application prospects in various fields. For example, Chandana et al. [2] apply the multi-focus image fusion to medical images such as CT, MRI and mammograms, which is helpful to improve the diagnosis accurate. Nowadays, the multi-focus image fusion has become a research hotspot in the field of image fusion [3–6].

Over the past few decades, researchers have proposed a number of methods to solve the problem of multi-focus image fusion, which can be divided into two categories: spatial-domain and transform-domain methods. In the spatial domain fusion methods, the fusion is usually based on pixels, blocks or regions [7–9]. Differently, the idea of transform domain methods is to transform the image to other domains and make use of the characteristics of the domains to achieve the goal more effectively, including multi-scale transform [10,11], sparse representation [12,13], hybrid [14], subspace [15] and other methods [16–18].

Although existing methods have been able to produce promising results in most cases, there are still several aspects to be improved. First, the existing methods usually need to manually design the activity level measurement and fusion rules, which limits fusion results because it is impossible to consider all the factors in one manually designed way. Second, many existing methods perform multi-focus image fusion by generating the decision map, which is more like a classification problem based on sharpness detection in essence. These methods often fail to classify the focused and defocused regions well near the boundary lines. Third, almost all deep learning-based methods require post-processing, such as consistency checks, when generating the decision map, which significantly increases the complexity of methods. Moreover, these methods typically require manual construction of decision map as ground truth to train the network, which further limits the scope of application of such methods.

* Corresponding author.
E-mail addresses: zhpersonalbox@gmail.com (H. Zhang), lezhuliang@whu.edu.cn (Z. Le), shaozhenfeng@whu.edu.cn (Z. Shao), xu_han@whu.edu.cn (H. Xu), jyma2010@gmail.com (J. Ma).
[1] These authors contributed equally to the work.

https://doi.org/10.1016/j.inffus.2020.08.022
Received 2 June 2020; Received in revised form 22 August 2020; Accepted 29 August 2020
Available online 4 September 2020

To solve above mentioned challenges, in this paper we design an unsupervised generative adversarial network with adaptive and gradient joint constraints, termed as *MFF-GAN*. We propose an adaptive decision block, which uses the repeated blur principle to determine whether the corresponding pixel is focused. Concretely, when the clear image is blurred, the change of pixel value before and after is larger than the change in which the originally blurred image is blurred again. The decision block generates a score map for each source image, which has the same size as the source image. In other words, the decision block makes a clear score for each pixel. According to the principle of maximum selection, the screening map used to guide the optimization can be obtained from the score maps. The screening maps act on a specifically designed loss function, which forces the generator to generate a fused image that is consistent with the clear source image. To further enhance the texture details of the fused image, we establishes an adversarial game between the generator and discriminator. After continuous adversarial learning, the gradient map of the fused image will approximate the joint gradient map we construct, thus containing richer texture details. Our method has the following advantages. First of all, our method does not need to design the activity level measurement and fusion rules, nor does it need any post-processing, which can implement simple and fast fusion of multi-focus images. Secondly, our network does not require ground truth for supervised learning, but unsupervised learning with weak constraints. As a result, the network can be easily trained on any pair of multi-focus images. Finally, our method is not based on the decision map, but realizes multi-focus image fusion by extracting and reconstructing information, and hence there is almost no blurring and detail loss near the boundary line. It is worth noting that due to the use of 1×1 convolution kernels and the control of the number of feature channels, the quantity of parameters in our network is limited within a certain range. As a result, our method can complete the fusion task at a high speed, which is about one order of magnitude faster compared with existing alternatives.

To intuitively demonstrate the characteristics of our method, we provide a typical example of our fused result with comparison to two state-of-the-art methods based on decision map, *i.e.*, guided filter-based method GFDF [19] and deep learning-based method SESF [20], as shown in Fig. 1. Clearly, our result maintains the detail in the far-focused source image, *e.g.*, the pipeline near the boundary line between the focused and defocused regions, while the other two methods cannot, resulting in the loss of information.

The major contributions of this paper are summarized as follows. First, we propose a new unsupervised GAN model with adaptive and gradient joint constraints for multi-focus image fusion, which realizes fusion by extracting and reconstructing information. As a result, there is almost no blurring and detail loss near the boundary line of the focused and defocused regions. Second, we design an adaptive decision block based on the repeated blur principle, which can effectively perform focus detection in pixel units, thereby guiding the generator to adaptively learn the distribution of clear source images, which avoids generating a fused result between clarity and blur. Third, we present a specific adversarial loss function based on the joint gradient constraint, which can further enhance the texture details of the fused result. Finally, we create a new multi-focus image fusion dataset named *MFI-WHU* based on publicly datasets, which contains 120 high-quality image pairs. As there are very few publicly available datasets for multi-focus image fusion and existing datasets typically only have dozens of image pairs, our MFI-WHU dataset provides a new choice for image fusion benchmark evaluation.

The remainder of this paper is organized as follows. Section 2 describes some related work, including an overview of existing deep learning-based multi-focus image fusion methods and generative adversarial network. In Section 3, we describe our method in detail, including overview of proposed method, loss functions and network architecture design. In Section 4, we give the detailed experimental settings and compare our method with several state-of-the-art methods on publicly available datasets by qualitative and quantitative comparisons. In addition, we also carry out the comparison of efficiency, the ablation experiments, sequence multi-focus image fusion experiment and generalization experiment in this section. Conclusions are given in Section 5.

To facilitate reading, the frequently used symbols and their definitions in this paper are summarized as follows. $\mathbf{G}$ is the generator, $\mathbf{D}$ represents the discriminator, and $\mathbf{S}$ indicates the screening map generated by the decision block. $\mathrm{Grad}_{\mathrm{joint}}$ is the joint gradient map, and $\mathrm{Grad}_{\mathrm{fused}}$ refers to the gradient map of the fused image. In addition, $\mathcal{L}$ represents the loss function.

2. Related work

This section describes the background and existing works that are most related to our approach, including the deep learning-based multi-focus image fusion and the development of generative adversarial network (GAN).

2.1. Deep learning-based multi-focus image fusion

In recent years, with the development of deep learning which has strong ability of extracting features, the multi-focus image fusion methods based on deep learning stands out among many traditional algorithms. Deep learning can learn fusion models with good generalization ability from a large amount of data, so as to make the fusion process more robust, showing strong development potential [21].

On the whole, the development of these methods is a transition from supervised to unsupervised, but most of them still focus on the generation of accurate decision maps. Liu et al. [22] used the convolutional neural network (CNN) to classify focused and defocused regions, thus generating the decision map for fusion. It is worth noting that they used the manually constructed decision map as the ground truth for supervised training to improve the accuracy of classification. Further, Du et al. [23] proposed a novel image segmentation-based multi-focus image fusion algorithm, in which the task of detecting the decision map is treated as image segmentation between the focused and defocused regions in the source images. Although this segmentation-based method increases the accuracy of the boundary line in the decision map to a certain extent, it would still result in the loss of details. Innovatively, Guo et al. [24] proposed to use the conditional GAN for multi-focus image fusion. However, labeled images are still needed in this method to conduct supervised training the network. To solve this problem, Ma et al. [20] proposed an unsupervised network to generate the decision map for fusion. Relying only on the learning ability of the neural network, even if there is ground truth for reference, these methods cannot generate an ideal decision map. Therefore, they require post-processing to further optimize the decision map, such as consistency verification or guided-filtering, which does not seem to make full use of the performance of neural networks. In contrast, our proposed MFF-GAN is not only unsupervised, but also does not require any post-processing.

2.2. GAN and its variants

Our method is based on GAN. Here we introduce some basic concepts of GAN and its variants such as deep convolutional GANs (DCGANs) and least squares GAN (LSGAN), which are the closest models to our algorithm.

Fig. 1. Illustration of multi-focus image fusion. From left to right: the far-focused image, the near-focused image, fused results of GFDF [19], SESF [20] and our proposed MFF-GAN. Our result maintains the details in the near-focused image, while the other two methods cannot.

2.2.1. GAN

GAN is an innovative generative model proposed by Goodfellow et al. [25], which is dedicated to estimate the target distribution and generate data which match the target distribution without relying on any a priori assumptions. In recent years, GAN has been widely used in various visual tasks including image fusion [26–28], and has achieved good performance.

From the perspective of frame structure, GAN is mainly composed of two components: the generator G and the discriminator D. The min–max game between the generator and the discriminator can gradually improve their abilities, and finally the expected generator that can estimate the target distribution is obtained. Here, we give a more formal description for this process. Assuming that the training data are $X = \{x_1, x_2, \ldots, x_n\}$, the generator can estimate the distribution characteristics from these data and then try to generate data $G(X)$ that match this distribution. For the discriminator, the task is to identify as much as possible which are the real training data X and which are the fake data $G(X)$ generated by the generator. In other words, the purpose of the whole GAN is to make the divergence between the estimated data distribution P_G and the real data distribution P_{data} as small as possible. Therefore, the objective function of GAN is defined as follows:

$$\min_G \max_D E_{x\sim P_{\text{data}}}[\log D(x)] + E_{x\sim P_G}[\log(1 - D(G(x)))]. \tag{1}$$

With continuous adversarial learning of the generator and the discriminator, the data distribution generated by the generator is closer and closer to the real data distribution, until the discriminator cannot distinguish whether they are real data or fake data. At this time, the trained generator is the expected generative which can generate the real-like data.

2.2.2. LSGAN

Our method is based on a variant of GAN, *i.e.*, the least squares GAN (LSGAN). LSGAN is an improvement of GAN, proposed by Mao et al. [29] in 2017. The traditional GAN uses a cross entropy loss function, which is prone to gradient disappearance during training. LSGAN can solve this problem better. Instead, it uses least square loss as loss function, and introduces labels as the goal of network optimization:

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

$$\min_G V_{\text{LSGAN}}(G) = \frac{1}{2}\mathbb{E}_{x\sim P_G}[(D(G(x)) - a)^2], \tag{3}$$

where $D(\cdot)$ is the discriminator function and $G(\cdot)$ is the generator function. In addition, c is the fake label that the discriminator determines fake data generated by the generator, b is the true label that the discriminator determines real data, and a is the label that the generator expects the discriminator to determine fake data. Obviously, c is as close to 0 as possible. On the contrary, a and b are as close to 1 as possible. That is, the discriminator wants to be able to accurately discriminate between true data and fake data, while the generator is trained to generate data which can be discriminated by the discriminator as the real data.

2.2.3. DCGANs

Inspired by DCGANs, our method combines CNN and GAN to better realize the multi-focus image fusion task. Radrord et al. [30] firstly proposed the DCGANs, who discussed the combination of CNN and GAN and gave a series of suggestions. Specifically, DCGANs provides five improvements to realize a better combination of CNN and GAN, so as to reduce the instability during the training process as much as possible. First, pooling layers are removed in both the generator and the discriminator. Instead, the stride convolution is used in discriminator to realize down-sampling, and the transposed convolution is applied in generator to achieve up-sampling. Second, batch normalization layers are applied to both the generator and the discriminator. Third, the rectified linear unit (ReLU) activation function is used on all layers of the generator except the output layer, while the output layer uses the tanh as the activation function. Fourth, leaky ReLU is used as the activation function on all layers of the discriminator. Finally, fully connected layers are removed in deeper models. On the whole, DCGANs introduces the strong feature extraction ability of CNN into GAN. On the one hand, the judgment accuracy of the discriminator can be improved. On the other hand, the feature extraction and reconstruction performance of the generator can be promoted.

3. Method

In this section, we give a detailed introduction of our method. We first introduce the overview of the proposed method, and then give the definition of loss function. Finally, the detailed structure of the proposed MFF-GAN is provided.

3.1. Overview of the proposed method

The idea of image fusion is to extract and combine the most meaningful information from source images. For multi-focus image fusion, the most meaningful information is the sharp regions in source images, which are reflected in the intensity distribution and texture details. Naturally, in the process of information extraction, these information in the sharp regions should be retained and these information in the fuzzy regions should be discarded. Therefore, it is necessary to introduce a mechanism to adjust the loss function in the optimization process, so as to constrain the network to selectively extract and reconstruct information. In addition, the details of the results should be strengthened to reduce the smoothing effect commonly seen in image generation tasks by neural networks. Based on these considerations, we design a generative adversarial network with adaptive and gradient joint constraints, and the overall fusion framework is shown in Fig. 2(a).

First, we design an adaptive decision block, which can evaluate the sharpness of each pixel based on the repeated blur principle, as illustrated in Fig. 2(b). That is, the image with higher sharpness, after the blur, the pixel value changes more. Based on this observation, the screening maps are generated to characterize the location of valid information. The screening maps act on a specific content loss function we construct, so as to adjust the optimization target at the pixel scale. In

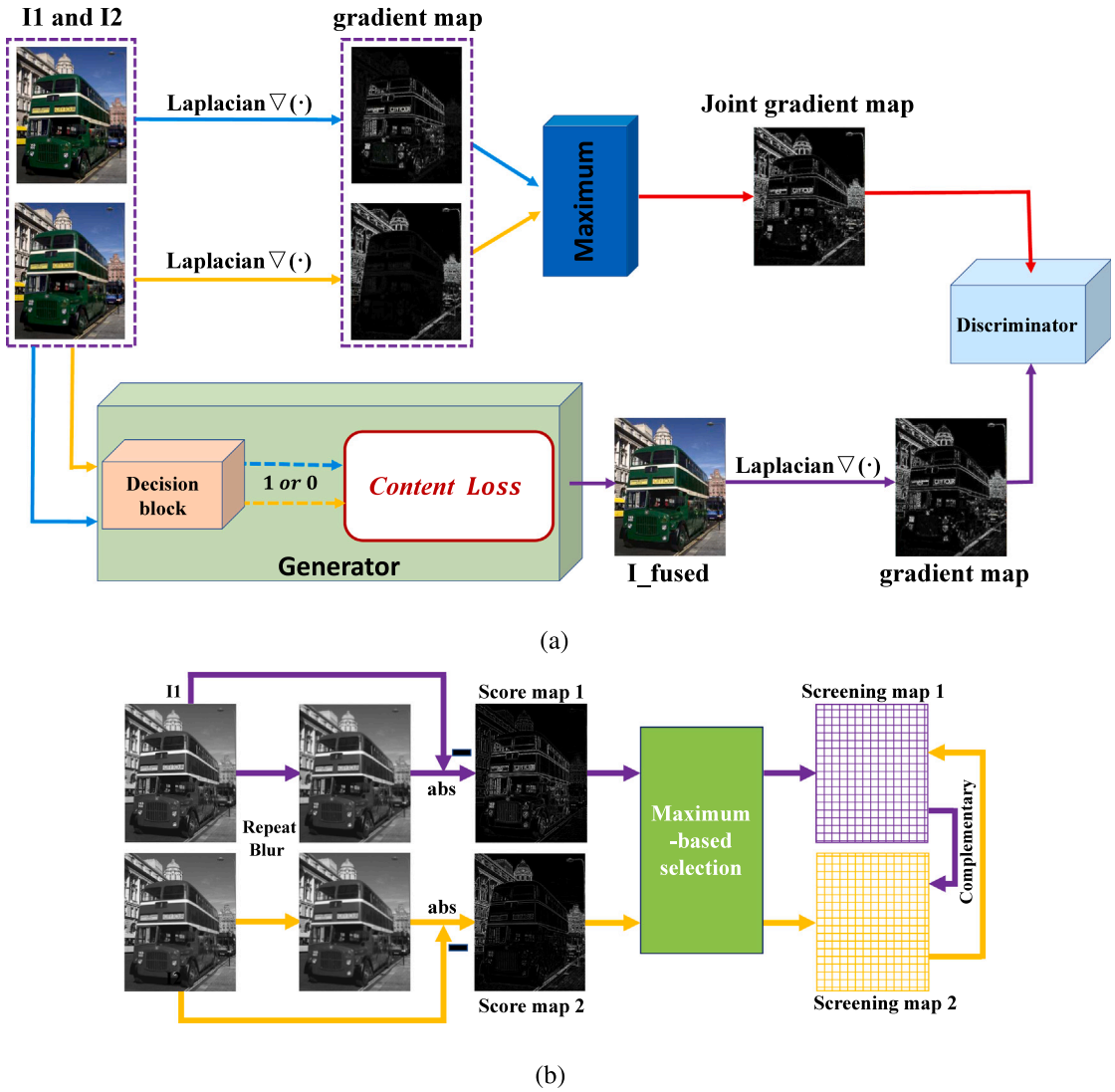

Fig. 2. Our MFF-GAN. (a) Overall fusion framework; (b) illustration of decision block.

other words, the decision block can adaptively guide the fused result to approximate the intensity distribution and gradient distribution of the clear source image on the pixel scale. Our specific approach is to select pixels with larger scores (abandoning smaller ones) as the optimization target at the corresponding pixel positions of the two source images. Under the combined action of the decision block and the content loss, the generator can obtain relatively clear and natural results. Different from the fixed loss function of traditional neural networks, the loss function in our model is constructed dynamically. Specifically, only those pixels judged as sharp by the decision block can participate in the loss function calculation in the optimization process.

To further enhance the texture details and improve the quality of the fused image, we establish an adversarial game between the generator and discriminator. Specifically, we use the Laplacian operator $\nabla(\cdot)$ to find the gradient maps of the two source images, and then construct the stronger joint gradient map based on the principle of maximum selection. We define the joint gradient map as real data and the gradient map of the fused image as fake data. Continuous adversarial learning can guide the generator to focus more on the preservation of textures. As a result, we can get the fused result with higher quality, which contains richer texture details.

3.2. Loss functions

The loss functions consist of generator loss $\mathcal{L}_G$ and discriminator loss $\mathcal{L}_D$.

3.2.1. Loss function of generator

The loss of generator has two parts, *i.e.*, the content loss $\mathcal{L}_{G_{\text{con}}}$ for extracting and reconstructing information, and the adversarial loss $\mathcal{L}_{G_{\text{adv}}}$ for enhancing texture details. We formalize it as:

$$\mathcal{L}_G = \mathcal{L}_{G_{\text{adv}}} + \alpha L_{G_{\text{con}}}, \tag{4}$$

where α is used to adjust these two loss terms to the same level of importance.

The adversarial loss of the generator is to further enhance the texture details of the fused image, which is defined as:

$$\mathcal{L}_{G_{\text{adv}}} = \frac{1}{N}\sum_{n=1}^{N}\left(D(\nabla(I^n_{\text{fused}})) - a\right)^2, \tag{5}$$

where N is the number of fused images in a batch during training, a is the probability label that the generator expects the discriminator to determine the fused image, and $\nabla(\cdot)$ represents the operation of finding the gradient map using the Laplacian operator. This adversarial loss can force the generator to pay more attention to the preservation of texture

details. In other words, the adversarial game makes the fused image tend to have stronger textures on the premise that the fusion rules of generator are satisfied. Here a should be set to 1.

The content loss consists of two items: intensity loss and gradient loss. We formalize it as:

$$L_{G_{\text{con}}} = \beta_1 \mathcal{L}_{\text{int}} + \beta_2 \mathcal{L}_{\text{grad}}, \tag{6}$$

where $\beta_{(\cdot)}$ is used to adjust these two loss terms to the same level of importance.

The intensity loss $\mathcal{L}_{\text{int}}$ can constrain the fused image to have the same intensity distribution as the clear regions of source images, which is defined as:

$$\mathcal{L}_{\text{int}} = \frac{1}{HW}\sum_i\sum_j S_{1_{i,j}} \cdot (I_{\text{fused}_{i,j}} - I_{1_{i,j}})^2 + S_{2_{i,j}} \cdot (I_{\text{fused}_{i,j}} - I_{2_{i,j}})^2, \tag{7}$$

where i and j represent the pixel in the ith row and the jth column in the screening map or source images, H and W represent the height and width of the image, I_1 and I_2 are the source images, and I_{fused} is the fused image generated by the generator, which is defined by $G(I_1, I_2)$. In addition, $S_{(\cdot)}$ is the screening map generated by the decision block based on the sharpness of the source images. Specifically, the generation process of the screening map can be formalized:

$$S_{1_{i,j}} = \text{sign}(RB(I_{1_{i,j}}) - \min(RB(I_{1_{i,j}}), RB(I_{2_{i,j}}))), \tag{8}$$

$$S_{2_{i,j}} = 1 - S_{1_{i,j}}, \tag{9}$$

where $\min(\cdot)$ denotes a minimum function, and $\text{sign}(\cdot)$ is the sign function. Repeated blur function $RB(\cdot) = abs(I_{i,j} - LP(I_{i,j}))$, $LP(\cdot)$ denotes the low pass filter function. It is worth noting that the size of $S_{(\cdot)}$ is also $H \times W$.

The gradient loss term can constrain the generator so that the fused image has the same texture detail as the sharp source images, which is essential to improve the clarity of the fussed image. The gradient loss is still related to the screening maps generated by the decision block, and formalized as:

$$\begin{aligned}\mathcal{L}_{\text{grad}} = \frac{1}{HW}\sum_i\sum_j & S_{1_{i,j}} \cdot (\nabla I_{\text{fused}_{i,j}} - \nabla I_{1_{i,j}})^2 \\ & + S_{2_{i,j}} \cdot (\nabla I_{\text{fused}_{i,j}} - \nabla I_{2_{i,j}})^2.\end{aligned} \tag{10}$$

3.2.2. Loss function of discriminator

The loss function of discriminator enables the discriminator to accurately identify real and fake data. In our method, the fake data is the gradient map of the fused image. The real data is the joint gradient map we construct. It is obtained based on the principle of maximum selection, and has a stronger gradient distribution. The gradient map of the fused image and joint gradient map can be formalized as:

$$Grad_{\text{fused}} = \text{abs}(\nabla I_{\text{fused}}), \tag{11}$$

$$Grad_{\text{joint}} = \max(\text{abs}(\nabla I_1), \text{abs}(\nabla I_2)), \tag{12}$$

where $\text{abs}(\cdot)$ is the absolute value function and $\max(\cdot)$ is the maximum function. Then, the loss function of the discriminator is defined as:

$$\mathcal{L}_{D_{\text{adv}}} = \frac{1}{N}\sum_{n=1}^{N}[D(Grad^n_{\text{fused}}) - b]^2 + [D(Grad^n_{\text{joint}}) - c]^2, \tag{13}$$

where b is the label of the gradient map of fused image, which should be set as 0. c is the label of the joint gradient map, which should be set as 1. That is to say, the discriminator expects to accurately recognize the joint gradient map as real data, and the gradient map of the fused image as fake data. Under this constraint, the discriminator can guide the generator's tendency in information maintaining, that is, in favor of strongly textured preservation.

3.3. Network architecture

The main structures of our generator and discriminator are CNNs, which are given in detail as follows.

3.3.1. Generator architecture

The network structure of the generator is shown in Fig. 3. We split the generator into two paths to extract information, corresponding to two source images. The design of the generator network is inspired by the pseudo-Siamese network, and it is skilled in dealing with two relatively different inputs. Because multi-focus image pairs are sharp or blurred at the corresponding pixel locations, the pseudo-Siamese network is suitable for such images.

In both paths, there are four convolutional layers to extract the features. The first convolutional layer uses the 5×5 convolution kernel and the rest three use the 3×3 convolution kernel. They all use the Leaky ReLU as the activation function. In order to prevent the loss of information during the convolution process, we reuse the features based on the idea of the DenseNet [31]. That is, the input of each convolution layer is concatenated by the output of all previous convolutional layers. At the same time, in order to extract more sufficient information, we exchange information between the two paths. To be specific, the exchanged information is generated by the method of concatenating and convolution. Then, the exchanged information is concatenated together with the output of all previous convolutional layers as the input of the next convolutional layer.

Finally, we concatenate the output of all convolutional layers in the two paths, and then generate the fused image through a convolutional layer. The kernel size of the convolutional layer is 1×1, and the activation function is tanh. It is worth noting that in all convolution layers, the padding mode is set as "SAME", that is, the size of feature maps does not change in the whole convolution process, which is the same as the size of source images.

3.3.2. Discriminator architecture

The structure of the discriminator is shown in Fig. 4. There are two types of input in the discriminator, which are the joint gradient map based on the source images and the gradient map of the fused image. The discriminator consists of four convolution layers and one linear layer. The convolution kernel size of the four convolutional layers is 3×3, and they all use the leaky ReLU activation function. The stride of these convolution layers is set to 2. The last layer is the linear layer used to find the classification probability.

4. Experiments

In this section, we evaluate our MFF-GAN on publicly available datasets with comparison to five state-of-the-art methods including dctVar [32], DSIFT [9], S-A [33], CNN [22] and SESF [20]. Among them, dctVar, DSIFT, S-A are traditional methods, while CNN and SESF are methods based on deep learning. In addition, DSIFT, CNN and SESF are also methods based on the decision map. First, we introduce the experimental settings including datasets, training details and evaluation metrics. Then, we provide qualitative and quantitative results on two datasets. Subsequently, we give the comparison of efficiency to demonstrate the superiority of proposed method in terms of running time. In addition, we conduct the ablation experiments. Finally, we apply our method to sequence multi-focus image fusion and perform the generalization experiment.

4.1. Experimental settings

4.1.1. Datasets

Our experiments are conducted on two datasets, say the Lytro dataset [34] and the MFI-WHU dataset[2] we construct based on publicly database. The Lytro dataset contains multi-focus image pairs of various types of scenes, which is a commonly used datasets for multi-focus image fusion. The dataset contains 20 pairs of images together

[2] We will release the download link after the paper getting accepted.

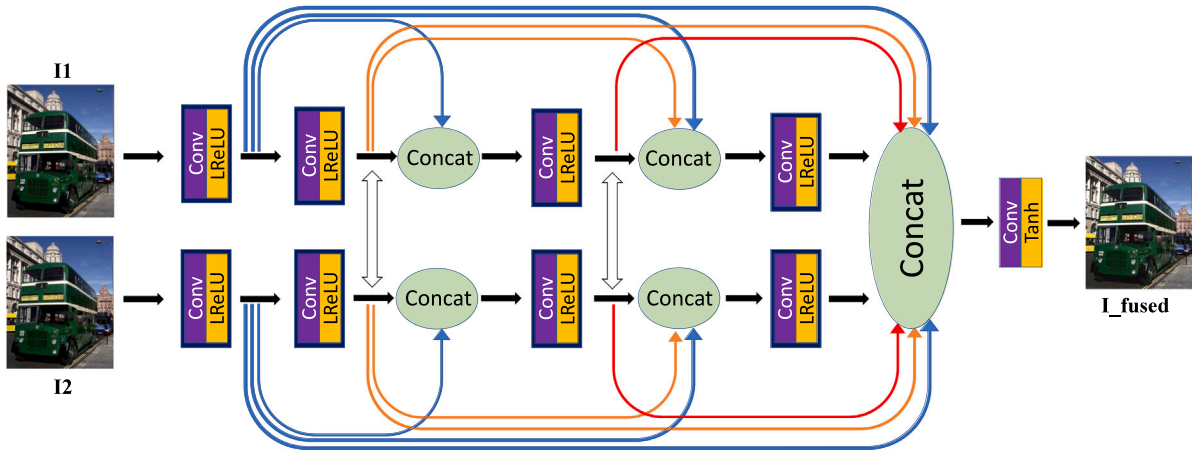

Fig. 3. Network architecture of the generator.

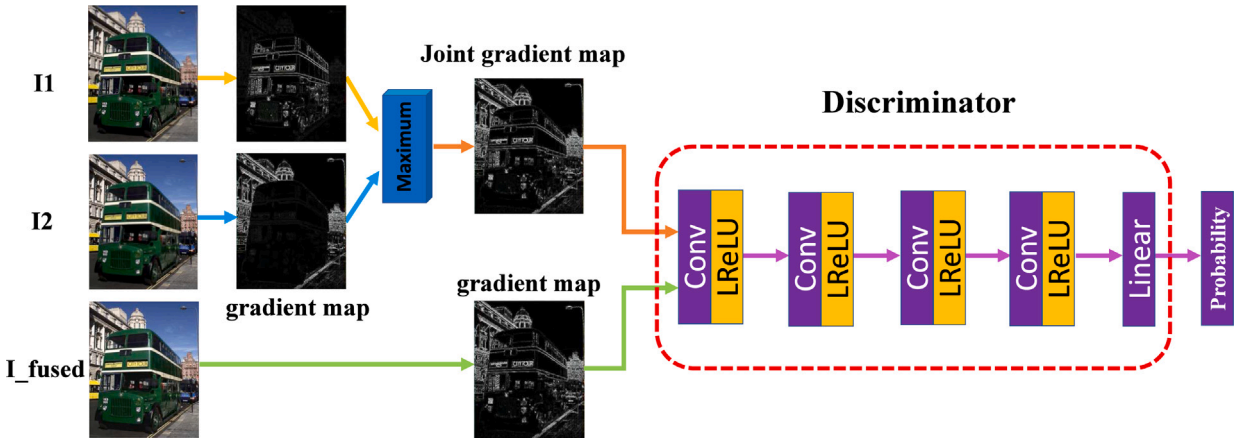

Fig. 4. Network architecture of the discriminator.

with 4 sequences. The MFI-WHU dataset is a multi-focus image fusion dataset established by ourselves, which contains 120 image pairs. These original images are from dataset provided by [35] and public COCO dataset [36]. We use Gaussian blur and a handmade decision map to generate multi-focus image pairs. It is worth noting that these artificially constructed ground truths are not used in our model.

On the Lytro dataset and MFI-WHU dataset, the number of image pairs used for testing is 10 and 30, respectively. For training, in order to obtain more training data, we adopt the expansion strategy of tailoring and decomposition. Specifically, for the Lytro dataset, we cropped the rest of images to 22,090 image patch pairs of size 60×60 for training; for the MFI-WHU dataset, we cropped the rest of images to 202,246 image patch pairs of size 60×60 for training.

4.1.2. Training details

The generator and discriminator are trained iteratively according to an adversarial process. The batch size is set to b, and it takes m steps to train one epoch. The ratio between the discriminator training number and generator training number is p, and a total of M epochs are trained. In our experiment, we set $b = 32$, $p = 2$, $M = 20$, and m is set as the ratio between the whole number of patches and b. The parameters in our MFF-GAN are updated by AdamOptimizer. We summarize the whole training procedure in Algorithm 1. In addition, the parameters of loss terms are set as: $\alpha = 10$, $\beta_1 = 1$ and $\beta_2 = 5$. In this work, we use soft labels for a, b, c, and d to make GAN's training more stable. Specifically, for the labels that should be set to 1 (*i.e.*, a and c), we set them to the random numbers ranging from 0.7 to 1.2. For labels that should be set to 0 (*i.e.*, b), we set them to the random numbers ranging from 0 to 0.3.

Algorithm 1 Training procedure of MFF-GAN

for M epochs **do**
  for m steps **do**
    for p times **do**
      Select b patches of source 1 $\left\{I_1^1, I_1^2 \cdots I_1^b\right\}$;
      Select b patches of source 2 $\left\{I_2^1, I_2^2 \cdots I_2^b\right\}$;
      Select b fused patches $\left\{I_{\text{fused}}^1, I_{\text{fused}}^2 \cdots I_{\text{fused}}^b\right\}$;
      Update the parameters of the discriminator by AdamOptimizer: $\nabla_D(\mathcal{L}_D)$;
    end for
    Select b patches of source 1 $\left\{I_1^1, I_1^2 \cdots I_1^b\right\}$;
    Select b patches of source 2 $\left\{I_2^1, I_2^2 \cdots I_2^b\right\}$;
    Update the parameters of the generator by AdamOptimizer: $\nabla_G(\mathcal{L}_G)$;
  end for
end for

We transform the images from RGB to YCbCr color space. Because the Y channel (luminance channel) can represent structural details and brightness variation, we just devote to fusing the Y channel values. For Cb and Cr channels (chrominance channels), we fuse them in a traditional way. Then, the fused components of these channels are transferred to RGB to obtain the final result. It is worth noting that all

deep learning-based methods run on the same GPU RTX 2080Ti, while other methods run on the same CPU Intel i7-8750H.

4.1.3. Evaluation metrics

In order to access the performance of different methods objectively, we evaluate the fusion results from two aspects, *i.e.*, qualitatively and quantitatively. Qualitative evaluation relies on the subjective visual perception of humans. A good fused result preserves the details while maintaining the sharpness of the source images, especially the details of the junction between the focused and defocused regions. Quantitative evaluation refers to measuring the performance of fused images, using common quality metrics. We select six popular statistics as objective metrics to measure the fusion results, such as standard deviation (SD) [37], entropy (EN) [38], $Q^{AB/F}$, spatial frequency (SF) [39], visual information fidelity (VIF) [40], sum of the correlations of differences (SCD) [41].

- SD: This metric reflects the distribution of pixel values in the image, that is, the distance between each pixel value and the average value. The SD is defined as:

$$SD = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(F(i,j)-\mu)^2}, \tag{14}$$

  in which M and N represent the size of the image, and μ is the average value of pixels. In general, the larger the SD, the higher the contrast.
- EN: This metric measures the amount of information contained in the fused image. The larger the value of EN, the more information it contains. The definition of EN is as follows:

$$EN = -\sum_{i=0}^{L-1} p_l \cdot \log_2 p_l, \tag{15}$$

  here, L is the number of gray levels of the image, and p_l refers to the normalized probability corresponding to the gradation l.
- $Q^{AB/F}$: This metric measures the amount of edge information that is transferred from source images to the fused image. $Q^{AB/F}$ is defined as:

$$Q^{AB/F} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} Q^{AF}(i,j)w^A(i,j) + Q^{BF}(i,j)w^B(i,j)}{\sum_{i=1}^{M}\sum_{j=1}^{N}(w^A(i,j)+w^B(i,j))}, \tag{16}$$

  where $Q^{XF}(i,j) = Q_g^{XF}(i,j)Q_a^{XF}(i,j)$, $Q_g^{XF}(i,j)$ and $Q_a^{XF}(i,j)$ indicate the edge strength and orientation values at location (i,j), respectively. w^X is the weight that expresses the importance of each source image to the fused image. A large $Q^{AB/F}$ means that considerable edge information is transferred to the fused image.
- SF: This metric measures the structural texture of the fused results. The SF is defined as:

$$SF = \sqrt{RF^2 + CF^2}, \tag{17}$$

  in which $RF = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(F(i,j)-F(i,j-1))^2}$ and $CF = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(F(i,j)-F(i-1,j))^2}$. A fused image with a large SF is sensitive to human perception according to the human visual system and has rich edges and textures.
- VIF: This metric measures the information fidelity of the fused image, which is consistent with the human visual system. VIF aims to compute the distortion between the fused and source images through four steps. First, the source images and fused image are filtered and divided into different blocks. Second, the visual information of each block with and without distortion is evaluated. Third, the VIF for each subband is calculated. Finally, the overall metric based on VIF is calculated.
- SCD: This metric measures the level of correlation between the information transmitted to the fused image and the corresponding source image, which can evaluate the pseudo information contained in the fused image. The larger SCD value indicates better fusion performance and less pseudo information. SCD is defined as:

$$SCD = r_{(D_1,VIS)} + r_{(D_2,IR)}, \tag{18}$$

  where D_1 and D_2 represent the information from visible image and the infrared image, respectively. They are defined as: $D_1 = I_{fused} - I_{ir}$ and $D_2 = I_{fused} - I_{vis}$. The $r(\cdot)$ is the correlation coefficient, which is defined as follows:

$$r_{(X,F)} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}(X(i,j)-\overline{X})(F(i,j)-\mu)}{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(X(i,j)-\overline{X})^2\sum_{i=1}^{M}\sum_{j=1}^{N}(F(i,j)-\mu)^2}}.$$

4.2. Results on the Lytro dataset

4.2.1. Qualitative experiments

We select two representative image pairs to qualitatively demonstrate the superiority of our method. The fusion results are shown in Fig. 5. From the results, we see that our MFF-GAN has clear advantages over other methods. First of all, our method can accurately retain the details of source images, including those near the boundary line of the focused and defocused regions. This is significant as most previous methods only can retain details away from the boundary line but lose those near the boundary line. In addition, our MFF-GAN can maintain regular textures better, such as outline lines.

For the decision map-based methods, such as CNN, SESF and DSIFT, they usually lose details near the junction of focused and defocused regions due to misclassification. For example, in the first set of results in Fig. 5, these methods lost the pipe on the ceiling. For those methods that are not based on the decision map (global), such as dctVar and S-A, they often blur regular edges or contain the blocking artifact. For example, in the second set of results in Fig. 5, these methods do not maintain the edges of the monkey's cheek well enough. In contrast, our method is more like combining the advantages of these two kinds of methods. On the one hand, our method can accurately maintain the details near the boundary line of the focused and defocused regions. On the other hand, the result of our method has good overall visual perception, which maintains regular edges and does not contain the blocking artifact.

4.2.2. Quantitative experiments

We further quantitatively demonstrate the characteristics of our method on the 10 image pairs from the Lytro dataset. The popular statistical results are reported in Fig. 6. Our method achieves the largest average values on five metrics including SD, EN, $Q^{AB/F}$, VIF and SCD. As for SF, our method can achieve the second largest average values. From these results, we can conclude that the result of our MFF-GAN has the best contrast, contains the most information and can maintain the best edge information. In addition, our method has the highest visual fidelity and introduces the least amount of pseudo information. Last but not least, our MFF-GAN can also preserve texture details well, which is only worse than SESF. All these quantitative metrics confirm the excellent visual effect of our method's fused results. Overall, our MFF-GAN performs better than all other comparative methods in objective evaluation.

4.3. Results on the MFI-WHU dataset

4.3.1. Qualitative experiments

We also give two groups of qualitative results from the MFI-WHU dataset, as shown in Fig. 7. It can be clearly seen that our method our can accurately retain details near the boundary line of focused and defocused regions, while CNN, SESF and DSIFT cannot. For example, in Fig. 7, these methods lost the athlete's hand in the first set of results. On the contrary, our method can accurately maintain the details of these

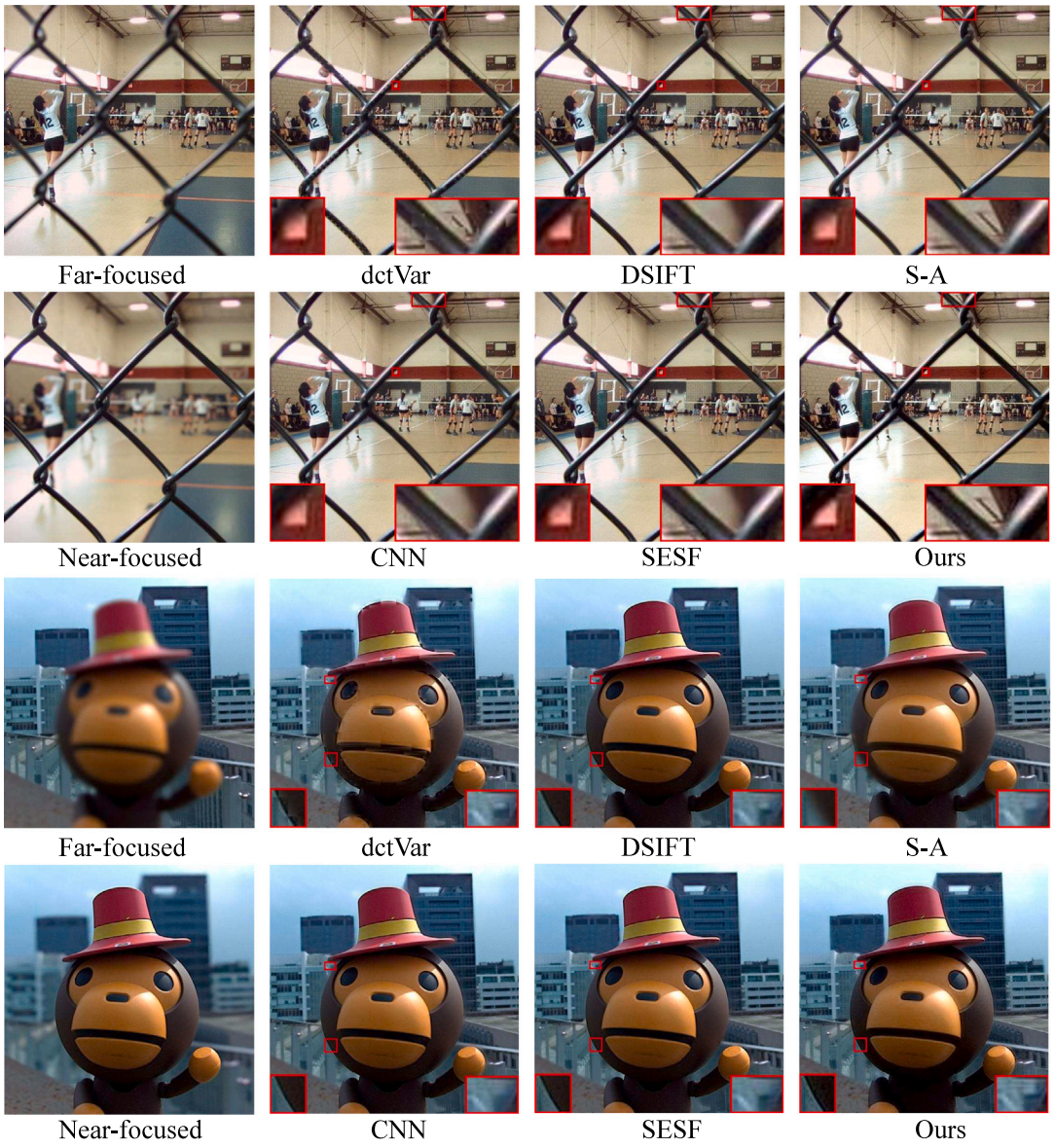

Fig. 5. Qualitative results on the Lytro dataset. In each group: source images, the results of dctVar [32], DSIFT [9], S-A [33], CNN [22], SESF [20] and our MFF-GAN.

parts. In addition, our method can maintain regular textures better, while dctVar and S-A cannot. For example, in the second set of results in Fig. 7, they blur the outline of the giraffe, which are sharpened in the result of our MFF-GAN. In general, our method not only has good overall clarity, but also can maintain local details, especially at the junction of the focused and defocused regions.

4.3.2. Quantitative experiments

We next quantitatively demonstrate the characteristics of our method on 30 pairs of images from the MFI-WHU dataset, and the results are reported in Fig. 8. Our method achieves the largest average value on all six metrics including SD, EN, $Q^{AB/F}$, SF, VIF and SCD. Different from the results from the Lytro dataset, our MFF-GAN also achieves the largest average value on SF. The reason is that the MFI-WHU dataset contains more image pairs, which enables the network to be fully trained. These results indicate that our method also has the best quantitative performance on the MFI-WHU dataset. Therefore, our MFF-GAN has superiority over all other methods in objective evaluation.

4.4. Comparison of efficiency

Our model consists of two parts: generator and discriminator. In the training phase, both the generator and the discriminator need to be optimized to improve performance. At this time, the total number of parameters is 0.431 M. In contrast, during the testing phase, only the generator is retained to produce the desired full-clear image, and the number of parameters used for testing is only 0.04 M, which is very lightweight. In order to evaluate our model more comprehensively, we also provide the comparative experiments on the running time in testing phase, as shown in Table 1. It can be seen that our method achieves the highest running efficiency in both two datasets, almost an order of magnitude faster than the comparative methods. Therefore, we can conclude that our MFF-GAN has the significant superiority in running time.

4.5. Ablation experiments

4.5.1. Decision block analysis

We use the decision block to control the loss function, so as to dynamically adjust the optimization target of the network. In other

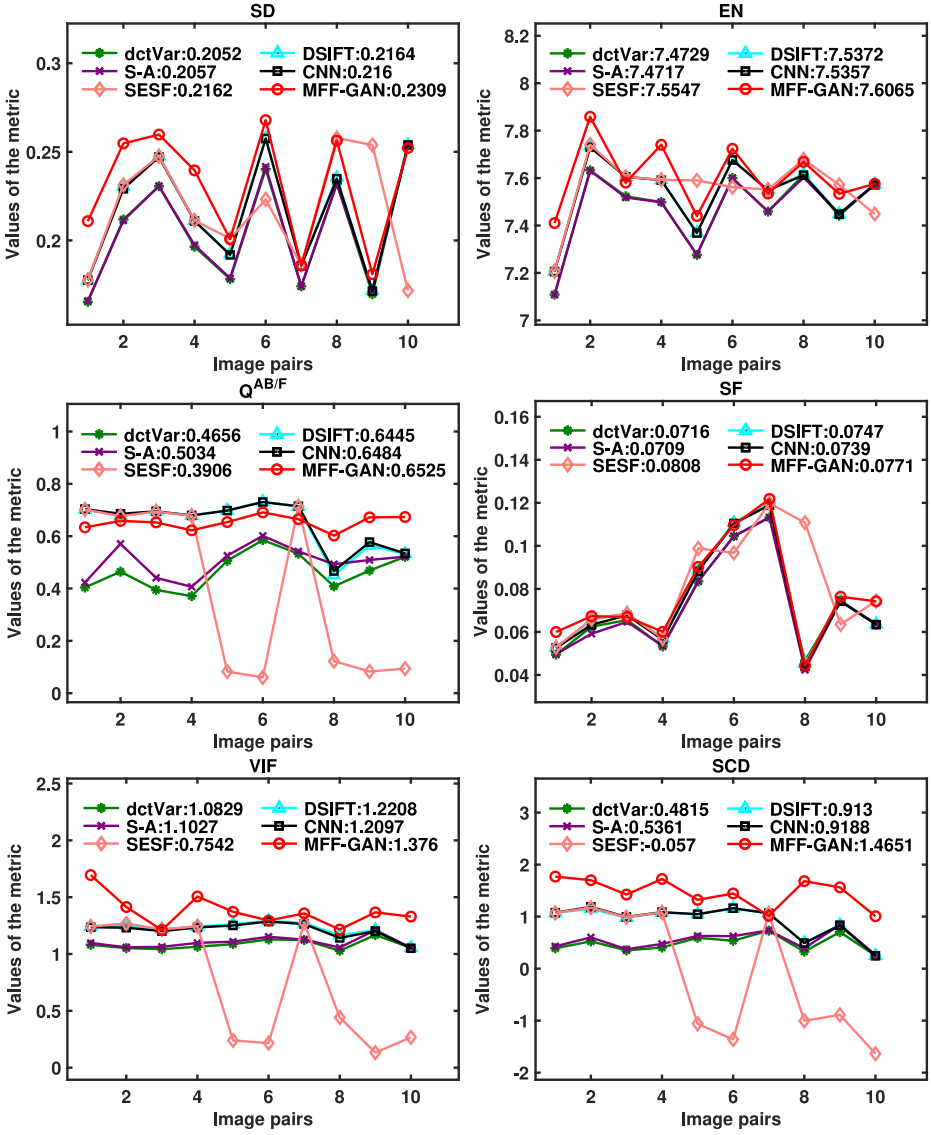

Fig. 6. Quantitative comparisons of the six metrics, *i.e.*, SD, EN, $Q^{AB/F}$, SF, VIF and SCD, on ten image pairs from the Lytro dataset.

Table 1
The mean and standard deviation of running time of different methods on two datasets (unit: second).

Dataset	dctVar [32]	DSIFT [9]	S-A [33]	CNN [22]	SESF [20]	Ours
Lytro	0.514 ± 0.030	5.086 ± 1.346	0.189 ± 0.013	106.807 ± 2.089	0.272 ± 0.015	**0.025 ± 0.001**
MFI-WHU	0.636 ± 0.139	6.963 ± 1.088	0.263 ± 0.060	137.538 ± 27.899	0.355 ± 0.098	**0.030 ± 0.006**

words, the decision block can make the distribution of the generated fused image approximate to the clear source image at the pixel scale. In order to fully verify the role of the decision block, we train our MFF-GAN without it. Specifically, we adopt the equally proportional setting strategy like PMGI [42] to realize multi-focus image fusion from the global perspective of patch. We randomly select two results shown in Fig. 9. When there is no decision block, the clarity of the fused result is not enough, which is more like the image between clarity and blur, such as the traffic board and text on the bus in highlighted regions. On the contrary, fused results with the decision block are full-clear. This demonstrates that the decision block plays an important role in the fusion process. Obviously, this pixel-scale control strategy is more appropriate than the global one in multi-focus image fusion.

4.5.2. Discriminator analysis

In our method, we employ the discriminator to further enhance the texture details of the fused image through an adversarial process with the generator. To validate its effectiveness, we train our MFF-GAN without the discriminator. A typical example is shown in Fig. 10. It can be clearly seen that the fused result with discriminator has more obvious and complete window details, while the result without discriminator suffers from some details weakening. As a result, this proves that the discriminator can further enhance the texture.

4.5.3. Parameter analysis

The optimization of the generator relies on two types of loss functions, namely content loss $\mathcal{L}_{G_{\text{con}}}$ and adversarial loss $\mathcal{L}_{G_{\text{adv}}}$. The role of $\mathcal{L}_{G_{\text{adv}}}$ is to further enhance the texture details in addition to the content loss, as demonstrated in the experiment above. Once the corresponding coefficient α enables the generator to establish an effective

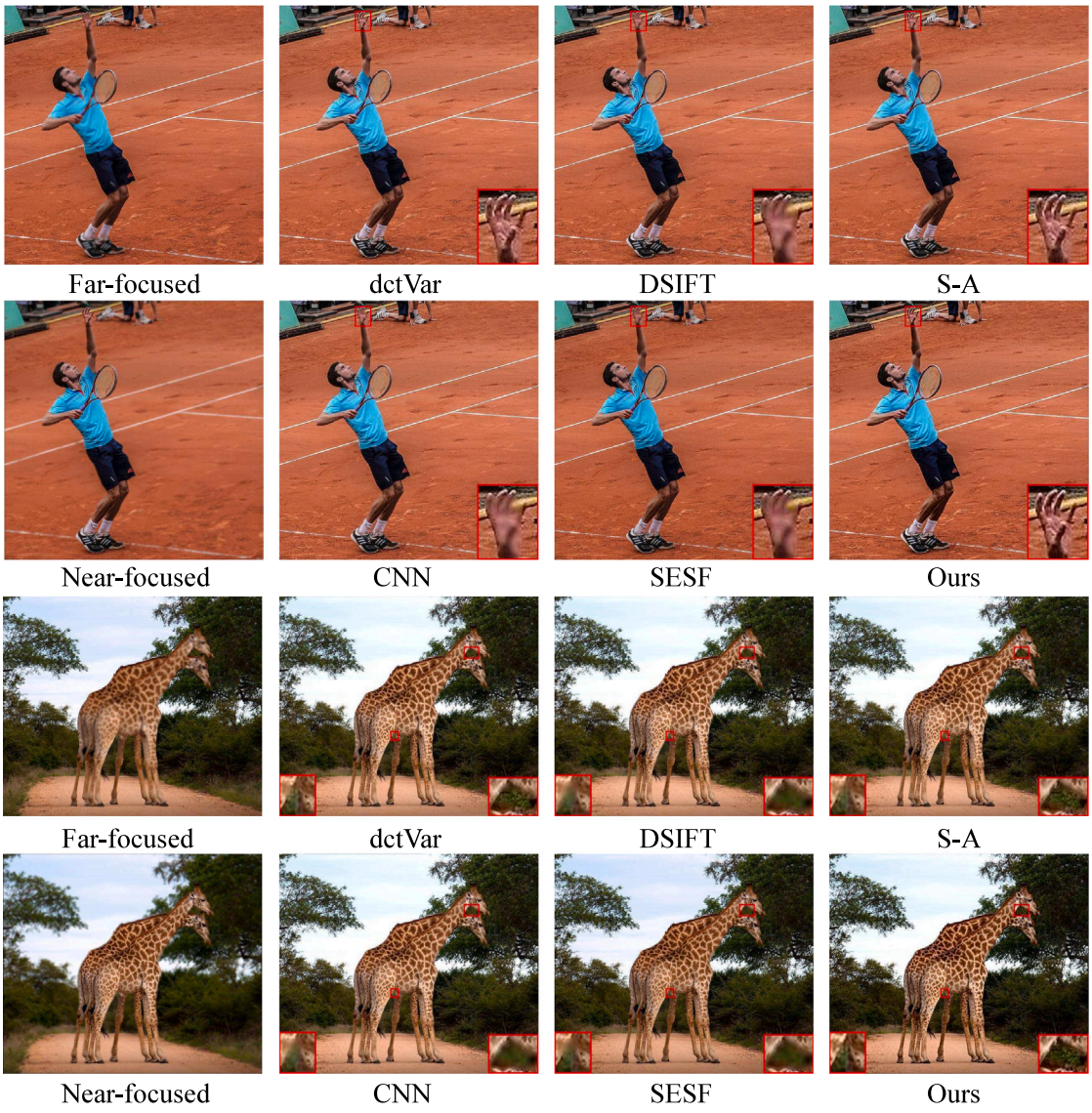

Fig. 7. Qualitative results on the MFI-WHU dataset. In each group: source images; the results of dctVar [32], DSIFT [9], S-A [33], CNN [22], SESF [20] and our MFF-GAN.

confrontation with the discriminator, the dependence of the model on α is not significant. In other words, the fused result is more sensitive to the $\mathcal{L}_{G_{\text{con}}}$. $\mathcal{L}_{G_{\text{con}}}$ consists of the intensity and gradient terms, and the corresponding weight parameters are β_1 and β_2. The intensity loss term constrains the fused result to maintain a similar intensity distribution with source images, so as to avoid contrast distortion. In contrast, the gradient loss term is dedicated to making the fused image contain rich texture details. Therefore, these two loss terms transform the image fusion into a multi-task optimization problem [43], and it is necessary to analyze the influence of β_1 and β_2 on the fusion. Specifically, we keep $\beta_1 = 1$ unchanged and assign different values to β_2. Because the factors affecting fusion process are actually the relative values of β_1 and β_2, this setting strategy is equivalent to studying these two parameters at the same time [42]. The results are shown in Fig. 11.

It can be clearly seen that as the coefficient β_2 of the gradient loss term gradually increases, the texture in the fused image becomes sharper. But at the same time, contrast distortion has become more obvious. This phenomenon is intuitively consistent with the effects of the intensity and gradient loss terms. When $\beta_2 = 5$, a good balance is realized in the preservation of contrast and texture structure. In other words, the fused result at this time not only contains clear texture details, but also has a contrast distribution very similar to source images.

4.6. Sequence multi-focus image fusion

When the number of multi-focus source images exceeds two, our method is also applicable. To confirm this point, we implement our method on a sequence with three multi-focus source images. Specifically, we first fuse two of the source images as before, and then fuse this intermediate result with the last source image to obtain the final fused image. The results are shown in Fig. 12. It can be seen that the fused result of our method contains all the clear regions in the three source images, which is a full-clear image with good visual effects.

4.7. Generalization experiment

The generalization ability of deep learning-based methods is an important basis for measuring the performance of a method. Due to differences in imaging equipment, the performance of the transferred model is often limited. However, our method is based on the extraction and reconstruction of information in sharpened areas, which is less affected by data differences. To verify this, we conduct the generalization experiment of our model. Concretely, we train it on the MFI-WHU dataset and then test it on the Lytro dataset. The qualitative and quantitative results are shown in Fig. 13 and Table 2. It can be seen

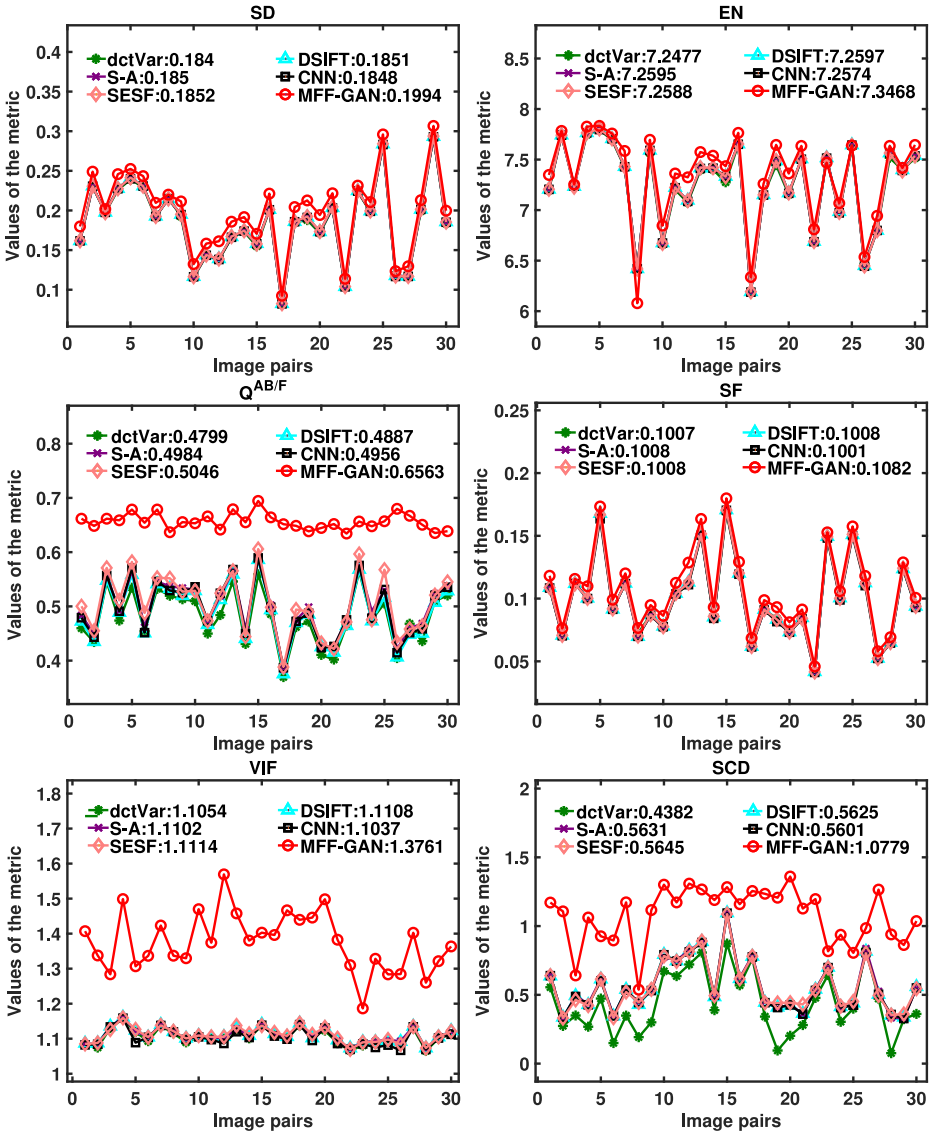

Fig. 8. Quantitative comparisons of the six metrics, *i.e.*, SD, CC, $Q^{AB/F}$, SF, VIF and SCD, on ten image pairs from the MFI-WHU dataset.

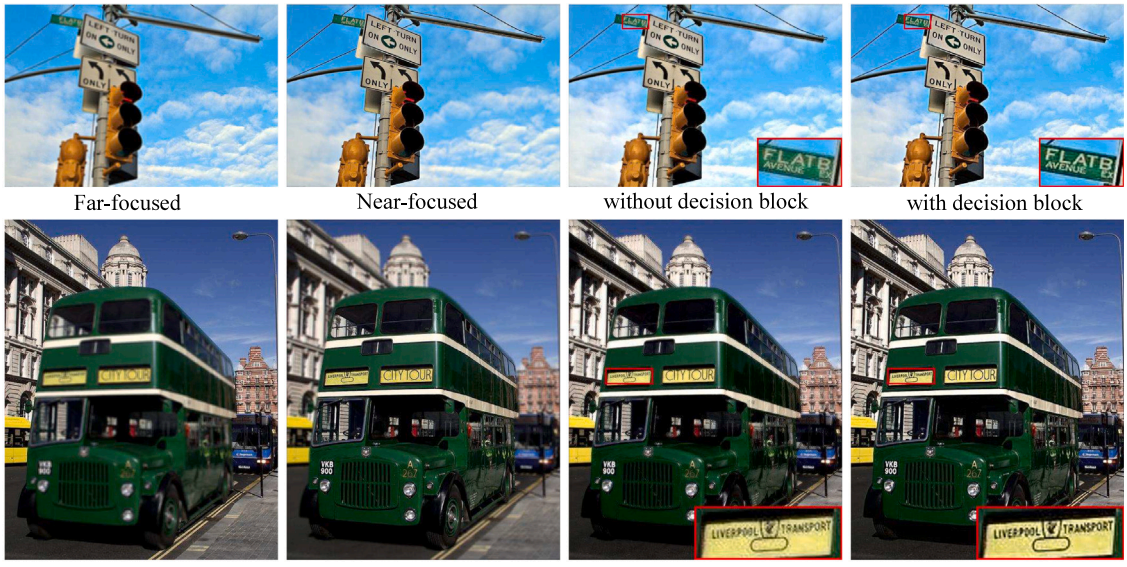

Fig. 9. Ablation experiment of the decision block. From left to right: far-focused images, near-focused images, results without discriminator and results with discriminator.

Fig. 10. Ablation experiment of the discriminator. From left to right: far-focused image, near-focused image, the result without discriminator and the result with discriminator.

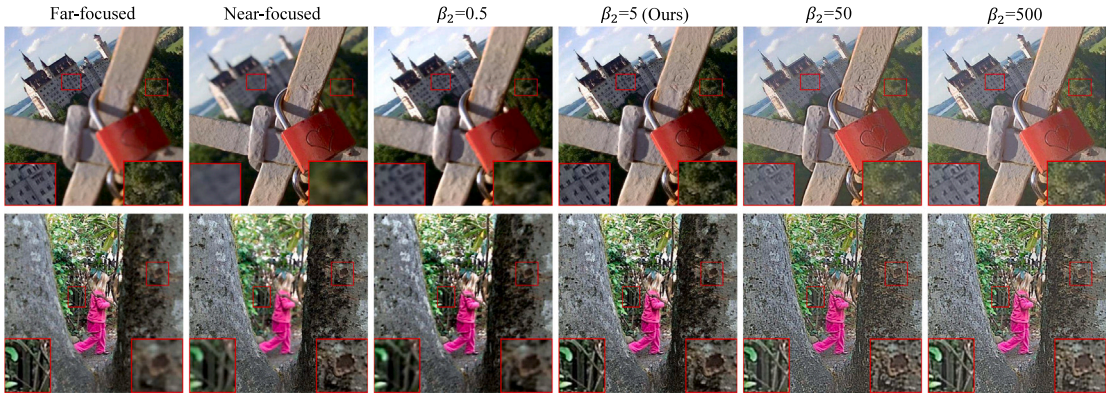

Fig. 11. Results of parameter analysis. We fix β_1 to 1 and assign different values to β_2. From left to right, β_2 gradually increases.

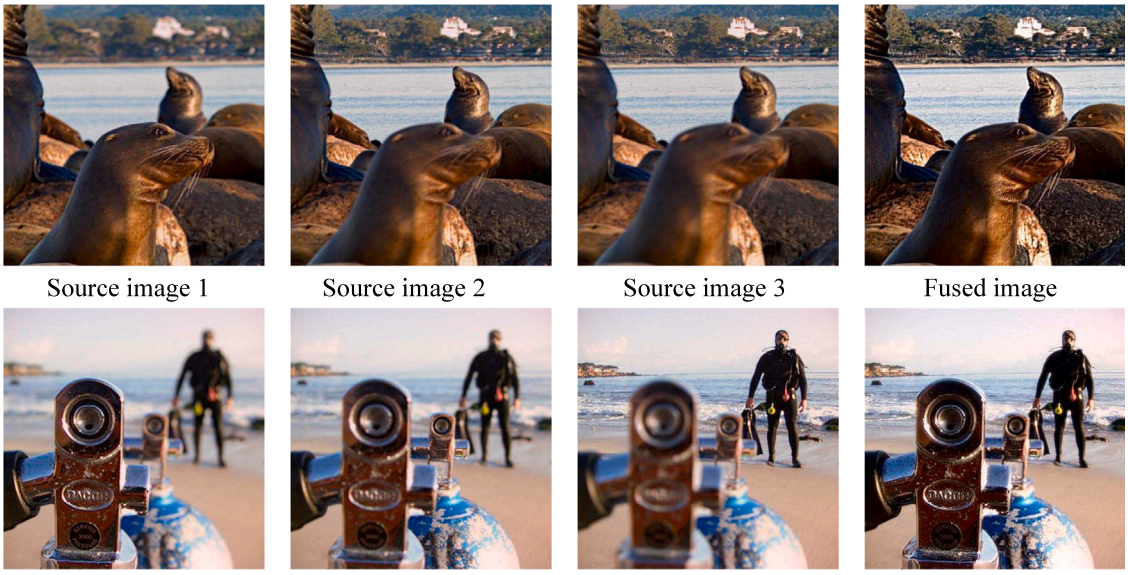

Fig. 12. Sequence multi-focus image fusion. From left to right: source image 1, source image 2, source image 3 and the result of our MFF-GAN.

Table 2
Quantitative comparison of generalization on 10 test images from Lytro dataset.

Method	SD	EN	$Q^{AB/F}$	SF	VIF	SCD
Ours Normal	0.231 ± 0.033	7.606 ± 0.140	0.652 ± 0.026	0.077 ± 0.024	1.376 ± 0.142	1.465 ± 0.266
Ours Transfer	0.222 ± 0.031	7.561 ± 0.147	0.655 ± 0.030	0.077 ± 0.024	1.259 ± 0.136	1.135 ± 0.391
Degradation	0.009	0.045	−0.003	0	0.117	0.330

from the qualitative results that the transferred model can also generate quite good fused results, in which texture details are maintained very well. Quantitative results further prove this. The transferred model has very little degradation in objective metrics compared with normal results, especially on $Q^{AB/F}$ and SF. Even compared to the results in Fig. 6, the transferred model also performs better than most of the comparative methods on these objective metrics. We further provide an extreme fused result as the example. Specifically, we randomly select a full-clear image (not from the previous two datasets), and then use the Gaussian kernel to degenerate it into a full-blur image. Then, the trained model is tested on the full-clear image and full-blur image, which is shown in Fig. 14. Obviously, our MFF-GAN can still accurately reconstruct the full-clear image, which is not affected by the full-blur source image. All of these show that our fully trained model can be easily transferred to other data and performs well.

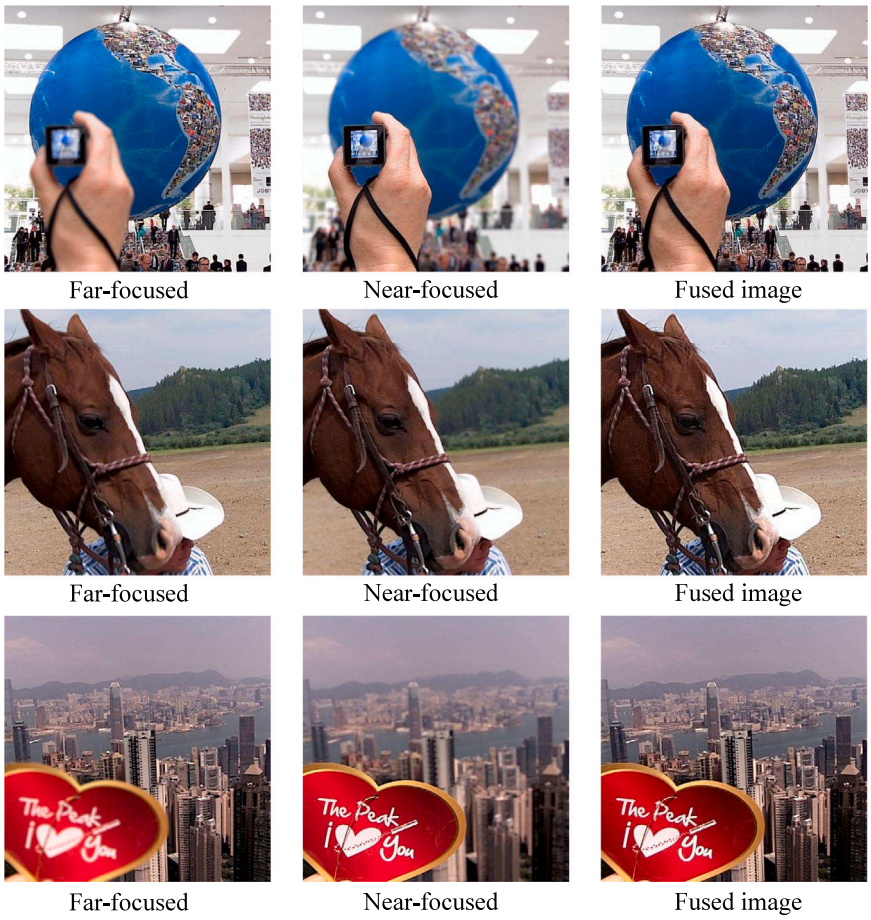

Fig. 13. Visualization of generalization experiment.

Fig. 14. Extreme fused results.

5. Discussion and conclusion

In this paper, a new unsupervised generative adversarial network with adaptive and gradient joint constraints called MFF-GAN is proposed to fuse multi-focus images. We employ the adaptive decision block based on the repeated blur principle to perform focus detection in pixel units, which can dynamically guide the network to produce a fused image of the same distribution as the focused source images, so as to avoid generating the result between clarity and blur. In addition, a specific adversarial loss function based on the joint gradient constraint is introduced to further enhance the texture details of the fused result. Moreover, we create a new multi-focus image fusion dataset for benchmark evaluation. Qualitative experiments show that our MFF-GAN not only has good overall clarity, but also can maintain local details, especially those near the junction of the focused and defocused regions. Quantitative experiments demonstrate that our model performs better than the existing state-of-the-art methods on six widely used metrics. In addition, our method is about one order of magnitude faster than the comparative methods.

Multi-focus images are captured under different shooting settings, which is a type of multi-modal images. In the future, we will focus on broader multi-modal deep learning [44,45], so as to develop a unified model to realize multiple multi-modal image fusion tasks, including infrared and visible image fusion, medical image fusion, multi-exposure image fusion, multi-focus image fusion, and so on. In addition, we will further explore the decomposition process from the fused result to the multi-modal source image, which can realize the interaction between fusion and decomposition processes.

CRediT authorship contribution statement

Hao Zhang: Conceived and designed the research, Performed the experiments, Analyzed the results and wrote the manuscript. **Zhuliang Le:** Conceived and designed the research, Performed the experiments, Provided insightful advices to this work and revised the manuscript. **Zhenfeng Shao:** Conceived and designed the research, Provided insightful advices to this work and revised the manuscript. **Han Xu:**

Analyzed the results and wrote the manuscript. **Jiayi Ma:** Conceived and designed the research, Provided insightful advices to this work and revised the manuscript.

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.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 61773295 and 41890820, and in part by the Natural Science Foundation of Hubei Province, China under Grant No. 2019CFA037.

References

[1] S. Li, X. Kang, L. Fang, J. Hu, H. Yin, Pixel-level image fusion: A survey of the state of the art, Inf. Fusion 33 (2017) 100–112.
[2] M. Chandana, S. Amutha, N. Kumar, A hybrid multi-focus medical image fusion based on wavelet transform, Int. J. Res. Rev. Comput. Sci. 2 (4) (2011) 948.
[3] S. Li, X. Kang, J. Hu, B. Yang, Image matting for fusion of multi-focus images in dynamic scenes, Inf. Fusion 14 (2) (2013) 147–162.
[4] S. Li, J.T. Kwok, Y. Wang, Combination of images with diverse focuses using the spatial frequency, Inf. Fusion 2 (3) (2001) 169–176.
[5] S. Li, B. Yang, Multifocus image fusion using region segmentation and spatial frequency, Image Vis. Comput. 26 (7) (2008) 971–979.
[6] J. Ma, Y. Ma, C. Li, Infrared and visible image fusion methods and applications: A survey, Inf. Fusion 45 (2019) 153–178.
[7] W. Huang, Z. Jing, Multi-focus image fusion using pulse coupled neural network, Pattern Recognit. Lett. 28 (9) (2007) 1123–1132.
[8] S. Li, X. Kang, J. Hu, Image fusion with guided filtering, IEEE Trans. Image Process. 22 (7) (2013) 2864–2875.
[9] Y. Liu, S. Liu, Z. Wang, Multi-focus image fusion with dense sift, Inf. Fusion 23 (2015) 139–155.
[10] B.S. Kumar, Multifocus and multispectral image fusion based on pixel significance using discrete cosine harmonic wavelet transform, Signal Image Video Process. 7 (6) (2013) 1125–1143.
[11] Z. Liu, K. Tsukada, K. Hanasaki, Y.-K. Ho, Y. Dai, Image fusion by using steerable pyramid, Pattern Recognit. Lett. 22 (9) (2001) 929–939.
[12] B. Yang, S. Li, Multifocus image fusion and restoration with sparse representation, IEEE Trans. Instrum. Meas. 59 (4) (2009) 884–892.
[13] Q. Zhang, Y. Liu, R.S. Blum, J. Han, D. Tao, Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review, Inf. Fusion 40 (2018) 57–75.
[14] N. Paramanandham, K. Rajendiran, Multi sensor image fusion for surveillance applications using hybrid image fusion algorithm, Multimedia Tools Appl. 77 (10) (2018) 12405–12436.
[15] L. Yang, B. Guo, W. Ni, Multifocus image fusion algorithm based on contourlet decomposition and region statistics, in: Proceedings of the International Conference on Image and Graphics, 2007, pp. 707–712.
[16] M. Amin-Naji, A. Aghagolzadeh, M. Ezoji, Ensemble of cnn for multi-focus image fusion, Inf. Fusion 51 (2019) 201–214.
[17] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, L. Zhang, Ifcnn: A general image fusion framework based on convolutional neural network, Inf. Fusion 54 (2020) 99–118.
[18] H. Xu, J. Ma, J. Jiang, X. Guo, H. Ling, U2fusion: A unified unsupervised image fusion network, IEEE Trans. Pattern Anal. Mach. Intell. (2020) http://dx.doi.org/10.1109/TPAMI.2020.3012548.
[19] X. Qiu, M. Li, L. Zhang, X. Yuan, Guided filter-based multi-focus image fusion through focus region detection, Signal Process., Image Commun. 72 (2019) 35–46.
[20] B. Ma, X. Ban, H. Huang, Y. Zhu, Sesf-fuse: An unsupervised deep model for multi-focus image fusion, 2019, arXiv preprint arXiv:1908.01703.
[21] 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.
[22] Y. Liu, X. Chen, H. Peng, Z. Wang, Multi-focus image fusion with a deep convolutional neural network, Inf. Fusion 36 (2017) 191–207.
[23] C. Du, S. Gao, Image segmentation-based multi-focus image fusion through multi-scale convolutional neural network, IEEE Access 5 (2017) 15750–15761.
[24] X. Guo, R. Nie, J. Cao, D. Zhou, L. Mei, K. He, Fusegan: Learning to fuse multi-focus image via conditional generative adversarial network, IEEE Trans. Multimed. 21 (8) (2019) 1982–1996.
[25] 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.
[26] 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.
[27] 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.
[28] 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.
[29] X. Mao, Q. Li, H. Xie, R.Y. Lau, Z. Wang, S. Paul Smolley, Least squares generative adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2794–2802.
[30] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, 2015, arXiv preprint arXiv:1511.06434.
[31] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[32] M.B.A. Haghighat, A. Aghagolzadeh, H. Seyedarabi, Multi-focus image fusion for visual sensor networks in dct domain, Comput. Electr. Eng. 37 (5) (2011) 789–797.
[33] W. Li, Y. Xie, H. Zhou, Y. Han, K. Zhan, Structure-aware image fusion, Optik 172 (2018) 1–11.
[34] M. Nejati, S. Samavi, S. Shirani, Multi-focus image fusion using dictionary-based sparse representation, Inf. Fusion 25 (2015) 72–84.
[35] J. Cai, S. Gu, L. Zhang, Learning a deep single image contrast enhancer from multi-exposure images, IEEE Trans. Image Process. 27 (4) (2018) 2049–2062.
[36] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 740–755.
[37] Y.-J. Rao, In-fibre bragg grating sensors, Meas. Sci. Technol. 8 (4) (1997) 355.
[38] J.W. Roberts, J.A. Van Aardt, F.B. Ahmed, Assessment of image fusion procedures using entropy, image quality, and multispectral classification, J. Appl. Remote Sens. 2 (1) (2008) 023522.
[39] A.M. Eskicioglu, P.S. Fisher, Image quality measures and their performance, IEEE Trans. Commun. 43 (12) (1995) 2959–2965.
[40] H.R. Sheikh, A.C. Bovik, Image information and visual quality, IEEE Trans. Image Process. 15 (2) (2006) 430–444.
[41] V. Aslantas, E. Bendes, A new image quality metric for image fusion: the sum of the correlations of differences, Aeu-Int. J. Electron. Commun. 69 (12) (2015) 1890–1896.
[42] H. Zhang, H. Xu, Y. Xiao, X. Guo, J. Ma, Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 12797–12804.
[43] G. Aceto, D. Ciuonzo, A. Montieri, A. Pescapé, Toward effective mobile encrypted traffic classification through deep learning, Neurocomputing 409 (2020) 306–315.
[44] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal deep learning, in: Proceedings of the International Conference on International Conference on Machine Learning, 2011, pp. 689–696.
[45] G. Aceto, D. Ciuonzo, A. Montieri, A. Pescapè, Mimetic: Mobile encrypted traffic classification using multimodal deep learning, Comput. Netw. 165 (2019) 106944.