Complex & Intelligent Systems (2023) 9:7395–7422
https://doi.org/10.1007/s40747-023-01151-y

ORIGINAL ARTICLE

GIPC-GAN: an end-to-end gradient and intensity joint proportional constraint generative adversarial network for multi-focus image fusion

Junwu Li[1] · **Binhua Li**[1,2] · **Yaoxi Jiang**[1]

Received: 21 June 2022 / Accepted: 11 June 2023 / Published online: 30 June 2023

Abstract
As for the problems of boundary blurring and information loss in the multi-focus image fusion method based on the generative decision maps, this paper proposes a new gradient-intensity joint proportional constraint generative adversarial network for multi-focus image fusion, with the name of GIPC-GAN. First, a set of labeled multi-focus image datasets using the deep region competition algorithm on a public dataset is constructed. It can train the network and generate fused images in an end-to-end manner, while avoiding boundary errors caused by artificially constructed decision maps. Second, the most meaningful information in the multi-focus image fusion task is defined as the target intensity and detail gradient, and a jointly constrained loss function based on intensity and gradient proportional maintenance is proposed. Constrained by a specific loss function to force the generated image to retain the information of target intensity, global texture and local texture of the source image as much as possible and maintain the structural consistency between the fused image and the source image. Third, we introduce GAN into the network, and establish an adversarial game between the generator and the discriminator, so that the intensity structure and texture gradient retained by the fused image are kept in a balance, and the detailed information of the fused image is further enhanced. Last but not least, experiments are conducted on two multi-focus public datasets and a multi-source multi-focus image sequence dataset and compared with other 7 state-of-the-art algorithms. The experimental results show that the images fused by the GIPC-GAN model are superior to other comparison algorithms in both subjective performance and objective measurement, and basically meet the requirements of real-time image fusion in terms of running efficiency and mode parameters quantity.

Keywords Multi-focus image fusion · Gradient-intensity joint proportional constraint · Deep region competition algorithm · Target intensity and detail gradient · Generative adversarial network · GIPC-GAN

Introduction

Due to hardware limitations such as optical lenses, images captured under a single shooting setting cannot comprehensively and effectively image scene information. For example, in digital photography, under the condition of limited depth of field, it is difficult to keep all-in-focus information of the objects at different depths of field in the same image [1, 2]. It is therefore what image fusion aims- to extract and reintegrate effective information from multi-source images to generate a single image that is informative and conducive to other subsequent computer vision tasks [3, 4]. As an important image enhancement method in the field of multimodal image fusion, multi-focus image fusion can effectively fuse different focus areas in multi-source images to generate an all-in-focused and all-clear image, which lays solid foundation for subsequent other computer vision tasks, such as medical diagnosis [5, 6], object detection and recognition [7], image denoising [8] and object segmentation [9, 10].

Multi-focus image fusion, as a research hotspot in image fusion, has witnessed rapid development in recent years [11–13]. For its problem, numerous research schemes have

✉ Binhua Li

Junwu Li
lijunwu@stu.kust.edu.cn

Yaoxi Jiang
jiangyaoxi@kust.edu.cn

[1] Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China

[2] Key Laboratory of Applications of Computer Technologies of the Yunnan Province, Kunming University of Science and Technology, Kunming 650500, China

been proposed by scholars, which can be mainly divided into traditional methods and deep-learning-based methods [14].

Traditional fusion methods manually measure activity levels and design fusion rules in spatial or transform domains through corresponding mathematical transformations, which is therefore divided into spatial domain methods, transform domain methods and hybrid domain methods. Specifically, the spatial domain methods directly perform linear operations on the pixels of the image to complete image fusion [15–17]; the transform domain method decomposes the source image into other domains, and integrates and reconstructs the image coefficients in the decomposed domain to achieve image fusion, in which are further divided into multi-scale transform method [18–20], sparse representation method [21, 22], saliency method [23], subspace method [24] and other methods [25]; and hybrid domain methods combine the advantages of the above two methods, thus having an improved fusion performance [26–28]. Despite their good image fusion results, there are still some limitations of the above methods. Firstly, traditional methods do not take into account the feature differences of multi-source images, and using same feature transformation in the feature extraction process, is likely to cause poor feature expression capabilities. Secondly, the fusion strategy designed by the traditional methods is relatively rough, and there is little space for improvement in image fusion performance. Finally, the design strategies are usually complex, which is not conducive to real-time image fusion tasks.

In recent years, deep learning has attracted extensive attention of scholars in the field of CV due to its powerful feature expression. Introducing deep learning to the image fusion task can overcome the above limitations of traditional fusion methods [29–31]. Liu et al. [32] first used convolutional neural network for multi-focus image fusion, which established a direct mapping between source images and focus images by binary classification of in-focus and de-focus regions. It is worth noting that hand-constructed high-quality image patches and Gaussian blurred versions are employed as supervised training datasets to train the classifiers. Ma et al. [33] proposed a two-stage unsupervised deep learning model for multi-focus image fusion. The method exploits spatial frequencies and adopts a gradient-based method to measure sharp changes in the features extracted by the network. The rate of change is used to characterize activity level measures and generate initial decision maps. Consistency checking are then conducted to refine the decision map and generate the final fused image. Zhang et al. [2] proposed a generative adversarial network with adaptive and gradient combined constraints to fuse multi-focus images. In this model, an adaptive decision block is introduced to determine the focus characteristics of source image pixels according to the principle of repeated blurring differences. Guided by the adaptive decision block, a content loss function is specially designed to dynamically guide the network optimization direction. In the adversarial game of GAN networks, the gradient maps of the generated images are forced to approximate the hand-designed joint gradient maps.

On one hand, most of the existing deep learning-based methods employ decision map to achieve multi-focus image fusion. This kind of method based on decision map directly combines the pixel regions of the source image, and can maintain high pixel fidelity. However, it requires higher classification accuracy of decision map. Misclassification can result in blurred edges and loss of information near the image focus and defocus boundaries. On the other hand, most of the deep learning-based methods need post-processing operations, such as guided filtering, consistency checking, and fully convolutional conditional random fields, after generating the initial decision map, so as to further refine the generated decision map, which increases the complexity of the model. Deep-learning-based supervised models mostly needs hand-designed decision maps and Gaussian blurring of clear images to generate multi-focus image pairs for model training [34–36]. It is such artificial subjectivity that limits the fusion performance of the algorithm.

To fix the above problems, this paper proposes a new gradient-intensity joint constrained generative adversarial network for multi-focus image fusion, named GIPC-GAN. First, inspired by the literature [37], we define the most meaningful information in the multi-focus image fusion task as texture gradient and object intensity. Notably, in the fusion process, both types of information from the source image are treated as equally important. The main purpose of the proposed algorithm is to make the fused image retain the maximum intensity and gradient information in the source image, that is, the focus information of the image. Second, the DRC algorithm in [38] is deployed to automatically construct a decision map on the public MFI-WHU dataset [2], and a new multi-focus image fusion dataset is established under the Gaussian blur principle. Third, a well-designed network model for multi-scale and multi-level feature representation and transfer of images, and a new loss function with gradient and intensity joint proportional constraints for the most meaningful information of multi-focus image fusion tasks, are presented in the paper. Last but not least, through the adversarial game between the generator and the discriminator, an adversarial balance is maintained, and the generated image is forced to keep as much information of target intensity and detail texture of the source image as possible.

The main contributions of this paper are as follows.

(1) This paper proposes a new generative adversarial network based on combined constraints of gradient and intensity for multi-focus image fusion. The proposed network is based on overall reconstruction approach,

instead of decision map. The model achieves an end-to-end multi-focus image fusion through deep feature extraction and overall feature reconstruction. It avoids the problem of boundary blurring effect because the generator directly generates the fused image instead of the decision map, that is, there is almost no information loss and blurring in the focus and defocus boundary line regions of the image. Without using decision maps, the network of the proposed GIPC-GAN algorithm does not require any post-processing operations, and it can quickly achieve multi-focus image fusion in an end-to-end manner.

(2) On the MFI-WHU public dataset of 120 high-quality images, this paper adopts the Deep Region Competition (DRC) algorithm to extract the foreground and background objects of the images, and automatically construct a decision map, which avoids the subjective arbitrariness of manually constructing decision maps and thus improves segmentation accuracy. A training dataset with 120 multi-focus image pairs is established through the Gaussian blur principle and accurate segmentation decision maps that can provide a new training benchmark and options for multi-focus image fusion tasks.

(3) The most meaningful information in the multi-focus image fusion task is defined as texture gradient and target intensity. In other words, the goal of our fusion image is to preserve as much information of target intensity and gradient texture in the source image as possible, and that is consistent with the multi-focus image fusion task of generating an all-in-focused and all-clear image.

(4) Based on the important information of gradient and intensity defined in the above multi-focus fusion task, this paper designs a new combined-constraint adversarial loss function that maintain proportional gradient and intensity information. It is known that, for the first time in image fusion methods, both intensity discrimination loss and gradient discrimination loss are taken into account in the discrimination loss function. A specific loss function is used to guide the direction of network training and optimization, and to further enhance the target intensity and detail texture of the fused image while maintaining the balance of the target intensity and detail texture information retained in the fused image.

(5) The GIPC-GAN model based on overall reconstruction forces the generated image to continuously approximate the probability distribution of the source image at the pixel level or feature level. The fused image also retains as much texture and intensity information of the source image as possible, and has higher pixel fidelity.

(6) Extensive experiments and ablation studies on two multi-focus public datasets and a multi-source multi-focus image sequence dataset demonstrate the importance of the designed intensity and gradient joint proportional-constrained loss function for multi-focus image fusion. Experimental results demonstrate that GIPC-GAN outperforms other state-of-the-art methods in qualitative and quantitative comparisons and meets the requirements of real-time image fusion.

The structure of the rest part is as follows. Section "Related work" gives an overview of relevant theory and work. The proposed network model is elaborated in Section "GIPC–GAN algorithm". Section "Experimental results" is the detail about the experimental configuration and used datasets, and subjective and objective comparative experiments and experimental evaluations with 7 state-of-the-art fusion algorithms on two public datasets, are conducted in this part. In addition, comparison of running efficiency and model calculation parameters, ablation experiments, fusion experiments and generalization experiments of multi-source and multi-focus image sequence pairs are conducted in this section. Section "Conclusions" is the summary and outlook.

Related work

In this section, a brief introduction of deep learning methods relevant to our work is made. For example, the original Generative Adversarial Networks (GANs) [39], LSGANs [40] for stabilizing the GAN training process, and FusionGAN [41], which used GANs for image fusion for the first time.

GANs [39]

GANs were first proposed by Goodfellow in 2014, and received huge attention in the field of deep learning immediately. Based on the principle of two-player zero-sum games, it can evaluate the target distribution and generate new samples without any supervision. An adversarial game between the Generator (G) and the Discriminator (D) is established, in which the former generates samples that deceive the discriminator, and the latter aims to discriminate whether the input samples are from the generator or the real samples, until it cannot distinguish the input samples.

Mathematically, the adversarial process is represented by maximizing D and minimizing G, and the objective function is shown in formula (1).

$$\min_G \max_D V_{GAN}(G, D) = E_{x\sim p_{data}(x)}\left[\log D(x)\right] + E_{z\sim p_z(z)}\left[\log 1\left(-D(G(z))\right)\right], \tag{1}$$

where G and D is the generator and the discriminator, respectively, p_{data} and p_z represent real sample distribution and generated sample distribution, and x and z represent the real sample and the generated sample, respectively. D, is to make $V_{GAN}(G, D)$ as large as possible, that is, strong recognition capability. For G, $V_{GAN}(G, D)$ should be minimized as much as possible, that is, the generated data is close to the real data. The whole training of GAN is the process of continuous and iterative training of G and D.

LSGANs [40]

Conventional GANs deploy a sigmoid cross-entropy loss function in the discriminator, which leads to the problem of vanishing gradients during training. Due to instability in training, it is difficult to obtain better models by alternating training. To address the problem, Mao et al. proposed least squares generative adversarial networks (LSGANs) in 2017. As an improved version of GANs, LSGANs adopts the least squares loss function in the discriminator and introduces labels to stabilize the network optimization process. The objective function is shown in formula (2) and formula (3).

$$\min V_{LSGAN}(D) = \frac{1}{2}E_{x\sim p_{data}(x)}\left[(D(x)-a)^2\right] + \frac{1}{2}E_{z\sim p_z(z)}\left[(D(G(z))-b)^2\right] \tag{2}$$

$$\min V_{LSGAN}(G) = \frac{1}{2}E_{z\sim p_z(z)}\left[(D(G(z))-c)^2\right], \tag{3}$$

where G and D represent the generator and the discriminator, respectively. a, b and c respectively indicate that D is expected to discriminate real data as true labels, D to discriminate generated data as false labels, and G expects D to discriminate generated data as true labels. It is obvious that the true labels a and c should be as close to 1 as possible, while the false label b should be as close to 0 as possible.

FusionGAN [41]

In 2019, Ma et al. applied GAN to image fusion for the first time. FusionGAN sets up an adversarial game between a generator and a discriminator to fuse infrared and visible images, whose generator can generate fused images with primary infrared targets and secondary visible light detail textures at the very beginning. At this time, the fused image contains more infrared target information and less texture information of the visible image, thus having unbalanced infrared target and visible light texture information. To address it, FusionGAN feeds the visible light image and the fused image into the discriminator, respectively, forcing the fused image and the visible light image to keep the distribution consistency through adversarial learning. The loss functions of its generator and discriminator are shown in formula (4) and formula (5), respectively.

$$\ell_G = \frac{1}{N}\sum_{n=1}^{N}\left(D_{\theta_D}\left(I_f^n\right)-c\right)^2 + \lambda\left[\frac{1}{HW}\left(\left\|I_f-I_r\right\|_F^2+\xi\left\|\nabla I_f-\nabla I_v\right\|_F^2\right)\right] \tag{4}$$

$$\ell_D = \frac{1}{N}\sum_{n=1}^{N}\left(D_{\theta_D}(I_v)-b\right)^2+\frac{1}{N}\sum_{n=1}^{N}\left(D_{\theta_D}(I_f)-a\right)^2, \tag{5}$$

where I_f, I_r, and I_v represent fused image, infrared source image, and visible light source image, respectively; H and W signify the height and width of the image; N is the number of fused images; $\|\cdot\|_F$ represents the Frobenius norm; ∇ represents the gradient operator, and λ and ξ represent balance parameters. a, b, and c represent the labels that the discriminator is expected to identify the fused image as fake, the discriminator to identify the visible light image as true, and the generator expects the discriminator to identify the fused image as true, respectively. The least squares loss function follows the minimization of Pearson's χ^2 divergence, which can make the network training process more stable and make the loss function achieve fast convergence.

GIPC-GAN algorithm

Introduction the proposed GIPC-GAN is made in this section. First is an overview of the model. Then, the architecture of the GIPC-GAN network is detailed. Last, the loss function of the model is designed.

Overview of GIPC-GAN method

Image fusion refers to the feature extraction and reintegration of the most meaningful information in source image, to generate a single image with rich information and can be used for subsequent specific fusion tasks. For the multi-focus image, focus pixel is the most meaningful information of the multi-focus image. Inspired by the Literature [37], it is inferred that the intensity of pixels can represent the histogram distribution of the image, the pixel gradient can denote the degree of difference between image pixels and the gradient information represents the detailed texture in the image. Therefore, we characterize the focus information of the source image in the intensity distribution as well as gradient distribution, and further define the most meaningful information in multi-focus

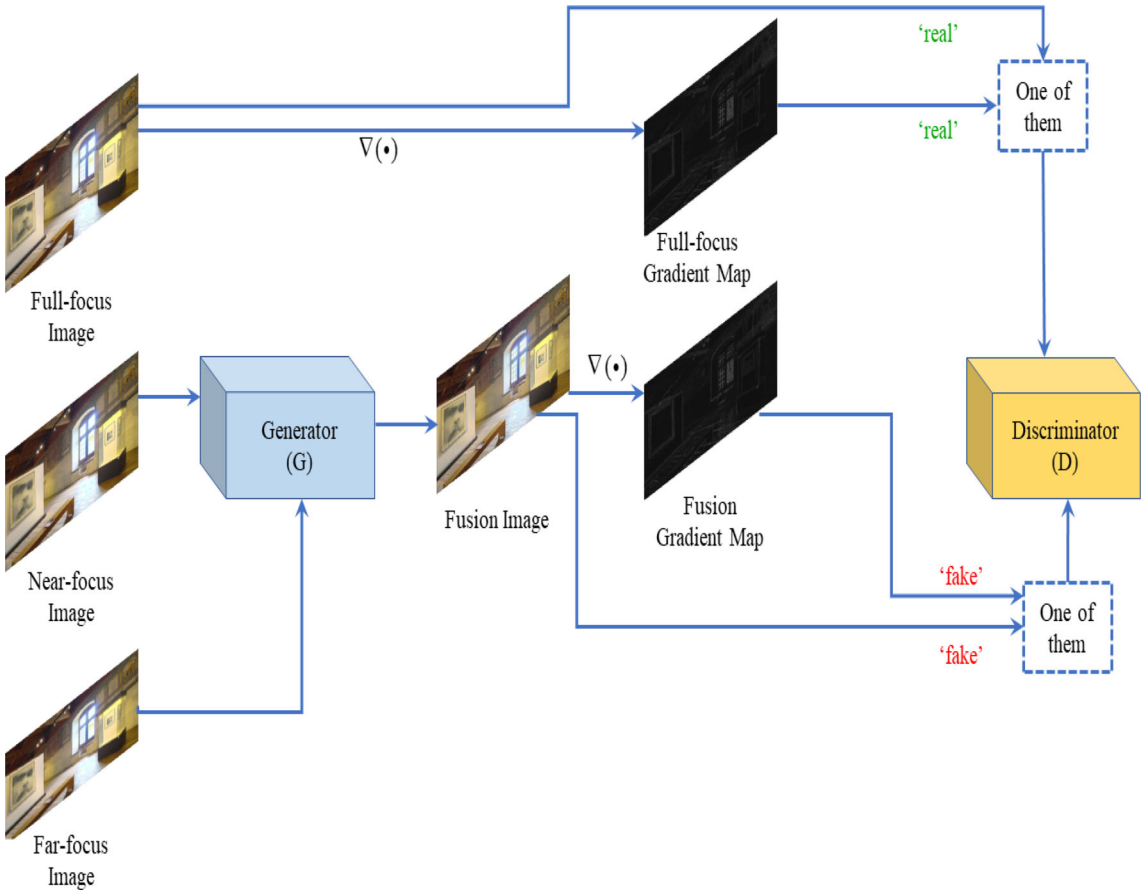

Fig. 1 Overall fusion framework of our GIPC-GAN

image fusion as texture gradient and target intensity. In other words, the purpose of our image fusion is to retain the information of target intensity and gradient texture in the source image as much as possible, so that the fused image can enjoy better visual effect and higher information fidelity. This is consistent with the task of multi-focus image fusion to generate an all-in-focused and all-clear image. Hence, we propose a novel gradient and intensity joint proportional constraint generative adversarial model for multi-focus image fusion. To better retain the intensity and gradient information of the source image, we need to get them in a balanced way. It is mainly reflected in the network architecture and loss function of the subsequent design. The GIPC-GAN overall network architecture is shown in Fig. 1, which is an end-to-end model.

The target intensity and texture details contained in the fused image are further enhanced through an adversarial game between the generator and the discriminator. Specific operations are as follows: (1) For fused image (I_F) and the source all-in-focused image (I_{S_clear}), Laplacian gradient operator is used to obtain the gradient map of the fusion image (GM_F) and the gradient map of source all-in-focused image (GM_{s_clear}). (2) We take I_{S_clear} and GM_{S_clear} as real images and I_F and GM_F as fake images, and then feed them to the discriminator respectively for recognition. (3) I_{S_clear} and GM_{S_clear} are recognized as real data by the discriminator and I_F and GM_F as fake data in terms of pixel intensity and gradient texture. (4) The discriminator continuously guides and optimizes the generator while improving its discriminative ability through continuous learning.

Through continuous adversarial learning and gaming, the generator thus improves its data generation ability to fool the discriminator. Ongoing adversarial learning between the generator and the discriminator can lead the generator to better balance source image texture details while paying attention to source image target intensities. Therefore, the images generated by our model have higher fusion quality.

GIPC-GAN network architecture

As an improved version of GANs, the GIPC-GAN network mainly consists of a generator and a discriminator. Considering that the three source images input by the model are color RGB images, it is of necessity to convert these images from RGB space to YCBCR color space. The Y channel of the near-focus source image and the Y-channel of the far-focus source image are input to the generator, which is designed to

generate an all-in-focused grayscale fusion image. While the input of the discriminator is the Y channel of the all-in-focus source image, the gradient map of the Y-channel of the all-in-focus source image, the grayscale fusion image, and the gradient map of the grayscale fusion image, which aims to distinguish the real source image from the fused image.

Generator architecture

The input of multi-focus image fusion is two different images, and thus have quite different information contained in them. For example, image A shows a clear foreground yet blurred background, while image B has a blurred foreground yet clear background. Referring to [2, 13], our model therefore adopts pseudo-Siamese network architecture that is good at dealing with relatively different inputs. The generator network architecture is shown in Fig. 2.

The generator contains two feature encoding paths, corresponding to near-focus and far-focus source images, respectively, in which the two paths have the same structure but do not share weight parameters. The architecture of the generator is based on the decoder-encoder, which mainly includes three main modules: encoder, feature generator and decoder.

The encoder consists of two image feature extraction paths, and each path consists of four convolutional block layers. Among them, each convolutional block layer consists of a convolutional layer, a BN layer and a LeakRelu layer. The size of the convolution kernel used in the first, second, third and fourth convolutional block layers is all set to 3×3, and the channel dimension of these four convolutional blocks is all set to 16. In the process of convolution, the problem of image feature loss usually occurs with the increase of the number of convolution layers. To address it, we employ dense connections [42] with regularization effect during feature transfer. That is, the output of the previous layer of each convolutional layer is concatenated with the subsequent convolutional layers, so as to compensate for the loss of information during feature transfer. In view of the fact that inputs in these two encoding paths are different, we conduct information exchange on all four convolutional block layers in the encoder to further supplement useful information. The feature generation module has only a convolutional block layer, whose convolution kernel size is set to 1×1 and the channel dimension is 128. It is input with the concatenation of features extracted by the near-focus encoding path and the far-focus encoding path, to further integrate the useful information extracted by these two encoding paths. In order to prevent the sudden change of the extracted features in the channel dimension, we design three convolutional block layers, a convolutional layer and a tanh activation function in the decoder module. Among which the first convolutional block layer contains a convolutional layer and a LeakRelu layer; the second and third contain a convolutional layer, a BN layer and a LeakRelu layer, with the kernel size of these three convolutional block layers of 3×3, and the channel dimensions of 64, 32, and 16, respectively. Then, the fused image is output through a convolutional layer and a tanh activation function; the kernel size of the last convolutional layer is set to 3×3, and the channel dimension is set to 1. Notably, we set Padding to "SAME" for all convolutional layers to prevent information loss during convolution down-sampling. The step length of convolutional kernel sizes 3×3 and 1×1 are set to 1 and 0, respectively, so that the feature dimensions of our input and output images remain unchanged throughout the generator architecture. The specific settings of all convolutional layers in the generator are shown in Table 1.

Discriminator architecture

The architecture of the discriminator is shown in Fig. 3, which mainly consists of four convolutional block layers and one linear layer, where the convolution kernel size of the first, second, third and fourth convolutional block layers is set to 3×3, and the channel dimensions are set to 16, 32, 64 and 128, respectively. In specific, the first convolutional block layer consists of a convolutional layer and a LeakRelu layer; the second, third, and fourth convolutional block layers all consist of a convolutional layer, a BN layer, and a LeakRelu layer, and the channel dimension of the last linear layer is set to 1, aiming to classify the input probabilities. The size length of all convolutional layers in the discriminator is set to 2. In order to better balance the retained intensity and gradient information of fused images, we design the input of the discriminator as all-in-focus image, all-in-focus gradient map, fused image and fused gradient map. The specific settings of all convolutional layers in the discriminator are shown in Table 2.

Loss function

It is known that loss functions play a crucial role in deep learning. The GIPC-GAN network is based on generative adversarial network, which mainly consists of generator loss (ℓ_G) and discriminator loss (ℓ_D).

Generator loss function

The generator loss guides and optimizes the generator, mainly including adversarial loss ($\ell_{G_{adv}}$) and content loss (ℓ_{Cont}). The mathematical definition is shown in formula (6).

$$\ell_G = \ell_{G_{adv}} + \lambda \ell_{Cont}, \tag{6}$$

where λ is a balance coefficient that controls the balance between adversarial loss and content loss.

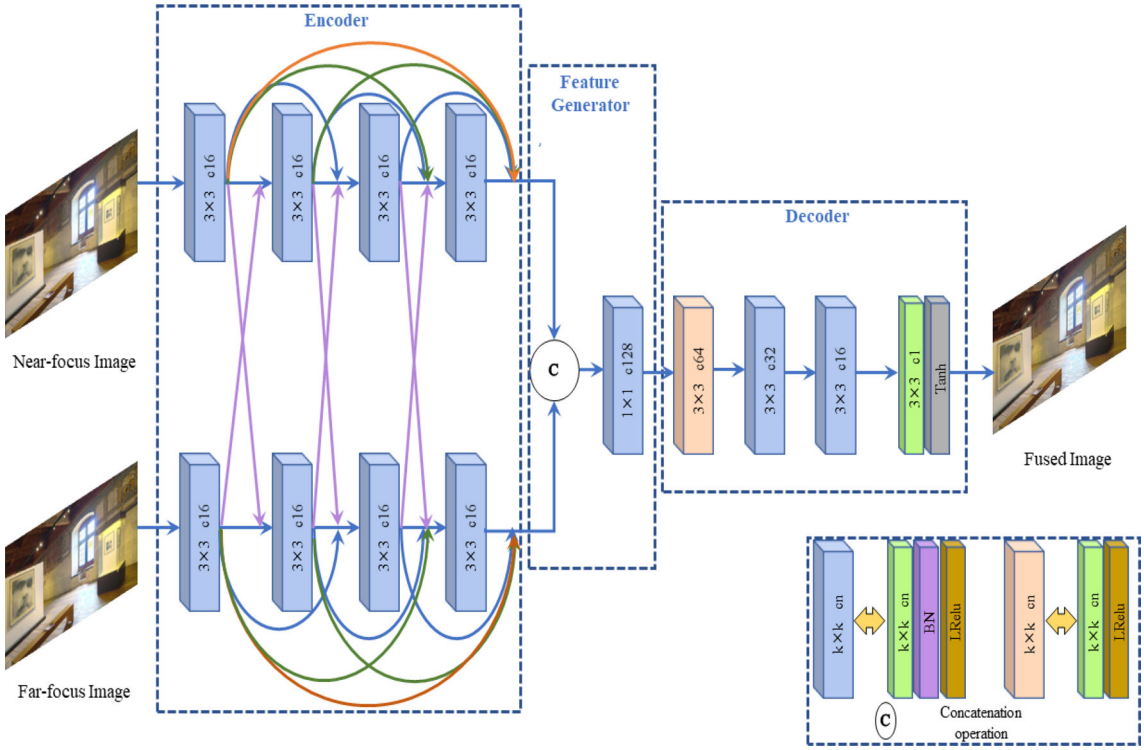

Fig. 2 Network architecture of the Generator

The adversarial loss is used to maintain a balance among the retained information, and can further enhance the target intensity and texture information of the fused image. It consists of a intensity adversarial loss ($\ell_{G_{adv\_i}}$) and a gradient adversarial loss ($\ell_{G_{adv\_g}}$). The mathematical definition is shown in formulas (7)–(9).

$$\ell_{G_{adv}} = \lambda_1 \ell_{G_{adv\_i}} + \lambda_2 \ell_{G_{adv\_g}} \tag{7}$$

$$\ell_{G_{adv\_i}} = \frac{1}{N}\sum_{n=1}^{N}\left[D\left(I^n_{fused}\right) - a_1\right]^2 \tag{8}$$

$$\ell_{G_{adv\_g}} = \frac{1}{N}\sum_{n=1}^{N}\left[D\left(\nabla I^n_{fused}\right) - a_2\right]^2, \tag{9}$$

where N is the number of fused images; $\nabla(\cdot)$ refers to the Laplace gradient operator; λ_1 and λ_2 are balance coefficients of fusion image adversarial loss and fusion image gradient adversarial loss respectively; a_1 and a_2 are the probabilistic labels that the generator expects the discriminator to identify with the fused image and the fused gradient map. These two adversarial losses help maintain the balance of target intensity information and gradient texture information retained in the fused image, and force the generator to further focus on keeping image intensity and texture details. The generator expects a_1 and a_2 to be as large as possible. For the sake of calculation convenience, $a_1 = a_2 = 1$ is set here.

Content loss constrains the extraction and reconstruction of image information. As the most important information of the multi-focus image is defined as the target intensity and texture gradient, the content loss includes two parts: intensity loss (ℓ_{int}) and gradient loss (ℓ_{gad}). The mathematical definition is shown in formula (10)–(12).

$$\ell_{Cont} = \beta_1 \ell_{int} + \beta_2 \ell_{gad} \tag{10}$$

$$\ell_{int} = \left\| I_{fused} - I_{focus} \right\|_F^2 \tag{11}$$

$$\ell_{gad} = \left\| \nabla I_{fused} - \nabla I_{focus} \right\|_F^2, \tag{12}$$

where β_1 and β_2 are the weight coefficients to keep the balance between the loss of the intensity term and the loss of the gradient term, and $\|\cdot\|_F^2$ represents the second norm of Frobenius, ∇I_{fused} and ∇I_{focus} denote the gradient map of the fused image and the gradient map of the all-in-focus image, respectively.

Discriminator loss function

The discriminator loss guides and optimizes the discriminator. It improves the discriminative ability through continuous training, and can distinguish the fused fake images from the real source images effectively. In our model, the fake image is the fused image and the gradient map of the fused image,

Table 1 Specific settings for all convolutional layers of the generator

Generator	Convolutional layers	Kernel size	Strides	Activation function	Input channals	Output channals
Encoder	Conv-Block layers-1	3 × 3	1	LRelu	1	16
	Conv-Block layers-2	3 × 3	1	LRelu	32	16
	Con-Block layers-3	3 × 3	1	LRelu	48	16
	Conv-Block layers-4	3 × 3	1	LRelu	64	16
Feature Generator	Conv-Block layers-5	1 × 1	0	LRelu	128	128
Decoder	Conv-Block layers-6	3 × 3	1	LRelu	128	64
	Conv-Block layers-7	3 × 3	1	LRelu	64	32
	Conv-Block layers-8	3 × 3	1	LRelu	32	16
	Conv-1	3 × 3	1	Tanh	16	1

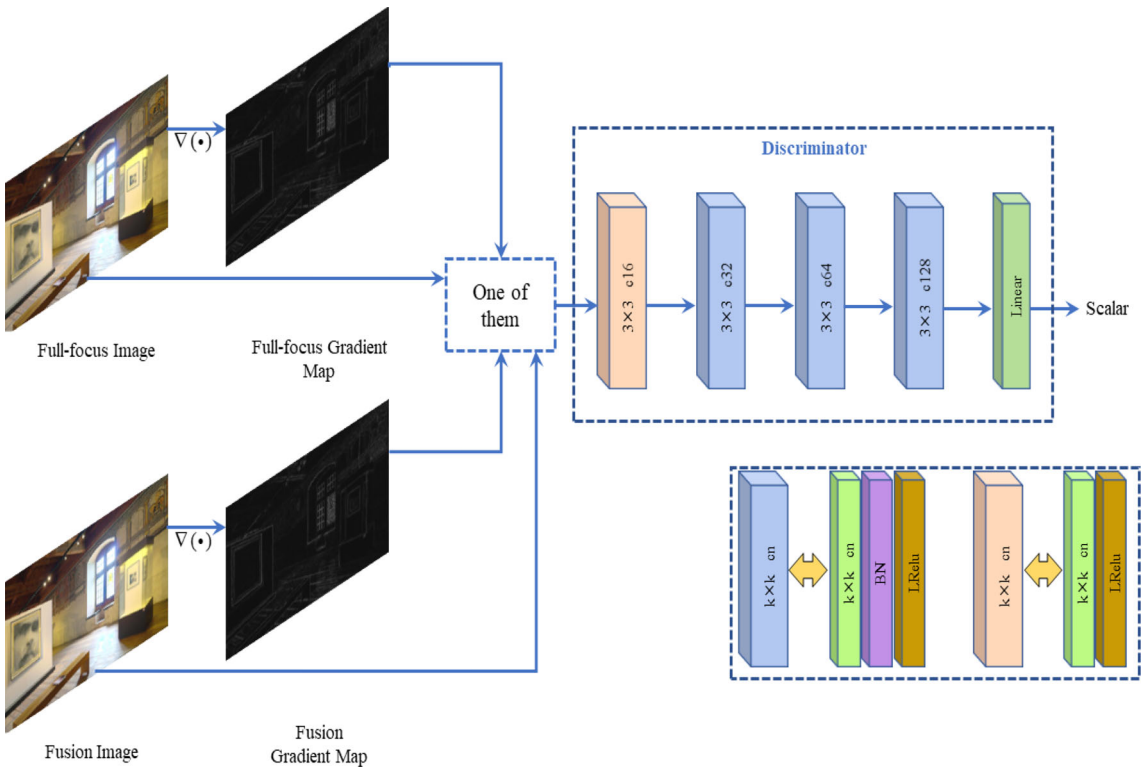

Fig. 3 Network architecture of the discriminator

Table 2 Specific settings for all convolutional layers of the discriminator

Discriminator	Convolutional layers	kernel size	Strides	Activation function	Input channals	Output channals
	Conv-Block layers-1	3 × 3	2	LRelu	1	16
	Conv-Block layers-2	3 × 3	2	LRelu	16	32
	Con-Block layers-3	3 × 3	2	LRelu	32	64
	Conv-Block layers-4	3 × 3	2	LRelu	64	128
	Linear layers-1	1 × 1	0	/	128*H*W	1

and the real image is the all-in-focus source image and the gradient map of the all-in-focus source image. Therefore, our discriminator loss consists of four parts: fused image discrimination loss (ℓ_{F_adv}), fused image gradient discrimination loss (ℓ_{Fg_adv}), all-in-focus source image discrimination loss (ℓ_{FC_adv}), and all-in-focus source image gradient discrimination loss (ℓ_{FCg_adv}). The mathematical definition is shown in formula (13)-formula (17).

$$\ell_D = \lambda_3\ell_{F\_adv}+\lambda_4\ell_{Fg\_adv}+\lambda_5\ell_{FC\_adv}+\lambda_6\ell_{FCg\_adv} \quad (13)$$

$$\ell_{F\_adv} = \frac{1}{N}\sum_{n=1}^{N}\left[D\left(I_{fused}\right)-b_1\right]^2 \quad (14)$$

$$\ell_{Fg\_adv} = \frac{1}{N}\sum_{n=1}^{N}\left[D\left(\nabla I_{fused}\right)-b_2\right]^2 \quad (15)$$

$$\ell_{FC\_adv} = \frac{1}{N}\sum_{n=1}^{N}\left[D\left(I_{foucs}\right)-c_1\right]^2 \quad (16)$$

$$\ell_{FCg\_adv} = \frac{1}{N}\sum_{n=1}^{N}\left[D\left(\nabla I_{foucs}\right)-c_2\right]^2, \quad (17)$$

where b_1, b_2, c_1 and c_2 are the probability labels of the discriminator to identify the fused image, the fused image gradient, the all-in-focus source image, and the all-in-focus source image gradient, respectively. The discriminator is expected to identify fake data and real data accurately, so b_1 and b_2 are expected to be as small as possible, and c_1 and c_2 to be as large as possible. For convenience, here we set $b_1 = b_2 = 0$ and $c_1 = c_2 = 1$. λ_3, λ_4, λ_5 and λ_6 is the balance coefficients of fusion image discrimination loss, fusion image gradient discrimination loss, source all-focus image discrimination loss and source all-focus image gradient discrimination loss respectively.

In order to maintain a balance between the target intensity and texture gradient information retained in the fused image, we set the same ratio of the two types of information in the discriminator and the generator, that is, the gradient and intensity information follow the ratio maintain consistency strategy. Under the constraint of discrimination loss, the discriminator can guide the optimization direction of the generator so as to generate a fused image that contains more and balanced target intensity and texture gradients information.

Experimental results

In this section, we first elaborate the experimental details, including the use of datasets, evaluation metrics, and settings of model parameters. Second, the GIPC-GAN model is compared and evaluated qualitatively and quantitatively with 7 state-of-the-art multi-focus fusion algorithms on two multi-focus public datasets, which are the BF [43] and DSIFT [44] based on spatial domain, the MWGF [45] based on transform domain, and the deep-learning-based CNN [32], SESF [33], ACGAN [54] and MFF-GAN [2]. Third, we conducted model complexity comparison experiments of various algorithms to comprehensively verify the efficiency of our proposed GIC-GAN model from the time complexity and space complexity of the algorithm. Fourth, we also conduct ablation experiments on the model. Final, applying the proposed GIPC-GAN model to multi-focus sequence image pairs to further verify the generalization of the model on multi-source multi-focus image pairs.

Experimental settings

Datasets

The training dataset is constructed on 120 all-focus images of MFI-WHU provided in Literature [2]. It is worth noting that the MFI-WHU dataset utilizes Gaussian blurring and manually-constructed decision maps to generate multi-focus image pairs, and is inevitably subjective. Such artificiality

and subjectivity will unavoidably lead to inaccurate segmentation decision maps and make it difficult to model multi-focus images in real scenes. Literature [38] proposed a Deep Region Competition (DRC) algorithm, which aims to extract foreground objects from images in a completely unsupervised manner. This algorithm treats foreground extraction as a special case of general image segmentation, focuses on identifying and separating objects from background, and is more competitive on complex real-world data and challenging multi-object scenes performance. Taking advantage of the Literature [36], we apply the DRC algorithm on the MFI-WHU public dataset to automatically generate decision maps. In this way, it not only avoids the tedious manual operation of the decision maps, but also improves the accuracy of generating the segmentation decision maps. Then the Gaussian blur algorithm is used to simulate the multi-focus image to generate multi-focus image pairs. The newly generated multi-focus image dataset is named MFI-DRC. The constructed multi-focus image pair can be represented by formula (18) and formula (19).

$$I_A = F * I + (1 - F) * (G(x, y; \sigma) \otimes I) \tag{18}$$

$$I_B = F * (G(x, y; \sigma) \otimes I) + (1 - F) * I, \tag{19}$$

where I_A and I_B represent the source image with a clear foreground and a blurred background and the source image with a blurred foreground and a clear background, respectively; I is the all-clear source image; F is the decision map; $G(x, y; \sigma)$ refers to a Gaussian filter, and $\otimes$ denote the convolution operation.

The MFI-DRC training set contains a total of three sets of images, I_A, I_B, and I, each of which contains 120 images. In view that MFI-DRC is based on the full-focus dataset MFI-WHU, it contains rich scene types such as mountains, houses, buildings, and animals. For model training, a large training dataset is usually required to avoid model overfitting. To get more training data, we employ a strategy of data augmentation to crop and decompose images. Specifically, 90 pairs of source images in the MFI-DRC dataset are selected randomly, and are crop into 184,885 pairs of image patches of resolution 80×80.

For the test dataset, experiments are performed on two datasets, Lytro [46] and MFI-DRC. For Lytro, a commonly used dataset for multi-focus image fusion, it contains 20 image pairs and 4 multi-focus image sequences. Therefore, the 20 image pairs and the rest 30 image pairs of the Lytro and MFI-DRC datasets are selected as our test datasets, respectively. Notably, no data augmentation is required for the test dataset.

Training settings

During the training, the generator and discriminator are optimized alternately. To maintain the stability of training, we set the ratio of discriminator training times to generator training times as t, that is, $t = 2:1$. The total training times epoch is set to n, that is, $n = 10$. The batch size is set to b, that is, $b = 32$. Each epoch requires training samples of batch m, where m denotes the ratio of the total number of training samples n to the number of sample batches b, that is, $m = n/b$. The Adam optimizer is adopted, with two default parameters β_1 and β_2 initialized to 0.5 and 0.999, respectively, to update and optimize the objective function. The initial learning rate of G and D is set to lr, that is, $lr = 0.0001$. In the training process, the learning rate is updated dynamically by using a linear decline strategy. To better understand the process of the algorithm, we summarized the training process of the whole model, as shown in Algorithm 1.

Like most image fusion methods, the hyperparameters of the loss function of GIPC-GAN model are also determined through the empirical values and experimental research of other relevant literature. Referring to literature [2, 54] and mode parameter tuning, we set the weight parameters in the generator loss function as $\lambda = 9.3$, $\beta_1 = 1.3$ and $\beta_2 = 4.5$. To make the model training more stable, inspired from Literature [38], we set a_1, a_2, b_1, b_2, c_1 and c_2 as soft labels. Specifically, for real value labels (a_1, a_2, c_1, and c_2) and fake value labels (b_1 and b_2), we set random numbers between (0.7, 1.2) and (0, 0.3), respectively. As for the fused image adversarial loss and fused image gradient adversarial loss in the generator, we found that when $\lambda_1 = 0.1$ and $\lambda_2 = 1$, the model can achieve better fusion effect through training. Since the information retention ratio of the intensity loss and gradient loss in the generator is 1:10, we implement the strategy of keeping the information ratio consistent for the intensity loss and gradient loss in the discriminator to set the discrimination loss balance parameters in the discriminator, that is, set $\lambda_3 = \lambda_5 = 0.1$ and $\lambda_4 = \lambda_6 = 1$.

Algorithm 1 Training Procedure of GIPC-GAN

Input: source foreground focus image I_{s_ffocus}, source background focus image I_{s_bfocus} and source all-focus image $I_{s\_allfocus}$

Output: fused all-focus image $I_{f\_allfocus}$

for n epochs **do**
  for m steps **do**
    for t times **do**
      Select b source foreground focus patches $\{I^1_{s\_ffocus}, I^2_{s\_ffocus}, ..., I^b_{s\_ffocus}\}$;
      Select b source background focus patches $\{I^1_{s\_bfocus}, I^2_{s\_bfocus}, ..., I^b_{s\_bfocus}\}$;
      Select b source all-focus patches $\{I^1_{s\_allfocus}, I^2_{s\_allfocus}, ..., I^b_{s\_allfocus}\}$;
      Select b fused all-focus patches $\{I^1_{f\_allfocus}, I^2_{f\_allfocus}, ..., I^b_{f\_allfocus}\}$;
      Update the parameters of the discriminator by AdamOptimizer:$\nabla D(\mathcal{L}_D)$ in Eq. (13);
    end for
    Select b source foreground focus patches $\{I^1_{s\_ffocus}, I^2_{s\_ffocus}, ..., I^b_{s\_ffocus}\}$;
    Select b source background focus patches $\{I^1_{s\_bfocus}, I^2_{s\_bfocus}, ..., I^b_{s\_bfocus}\}$;
    Generate fused all-focus patches $\{I^1_{f\_allfocus}, I^2_{f\_allfocus}, ..., I^b_{f\_allfocus}$ by G
    Update the parameters of the generator by AdamOptimizer: $\nabla G(\mathcal{L}_G)$ in Eq. (6);
  end for
end for

Training details and environment configurations

As the source image of multi-focus image fusion is a color RGB image, it needs to convert the input source image from RGB color space to YCBCR color space before training. And our method is to fuse the Y channel of the source image. For the CB and CR color channels of the source image, we use traditional methods to fuse them. And the obtained fused components of YCBCR space are converted to RGB color space to complete the final fusion of color RGB images.

The GPU configuration based on the deep learning method in this paper is: GPU-RTX 3090 24G, and the CPU setting based on the traditional method is: CPU-AMD Ryzen 9 3900 × 12-Core 3.79 GHz memory-32G. The software used in this paper is Tensorflow 2 and Matlab 2022a.

Evaluation metrics

Quality assessment of fused images is an important and complex research issue in image fusion tasks. In order to analyze the fusion performance of different methods in comprehensive way, it is of necessity to integrate both qualitative and quantitative analysis aspects to evaluate the fusion results. Among which, the qualitative evaluation is based on the human subjective visual system, and judges the performance of the fused images with human subjective awareness. For multi-focus image fusion tasks, the goal of fusion is to obtain all-in-focused and all-clear images within a limited depth of field range; while quantitative evaluation is based on mathematical statistical indicators to analyze and evaluate the quality of fused images from different statistical perspectives. In this paper, we select 6 objective quantitative indicators to comprehensively analyze the fusion performance of those algorithms, and they are: Entropy (EN) [47], Spatial Frequency (SF) [48], Standard Deviation (SD) [49], correlation coefficient (CC) [50], and Multiscale Structural Similarity (MS-SSIM) [51].

(1) EN measures the richness of information contained in images. The higher the value, the more information the fused image contains and the higher the fusion quality. The definition of EN is shown in formula (20).

$$EN = -\sum_{i=0}^{L-1} P_i \log_2(P_i), \tag{20}$$

where L is the total gray level of the image, and P_i is the normalized histogram corresponding to the gray level i.

(2) SF reflects the gray level change rate of the image. The higher the value, the clearer the fused image, the richer the edge and texture details, and the better the fused image quality. The definition of SF is shown in formula (21).

$$SF = \sqrt{RF^2 + CF^2} \tag{21}$$

where $RF = \sqrt{\frac{1}{M\times N}\sum_{i=1}^{M}\sum_{j=2}^{N}[F(i,\ j) - F(i,\ j-1)]^2}$ is the spatial row frequency, $CF = \sqrt{\frac{1}{M\times N}\sum_{i=2}^{M}\sum_{j=1}^{N}[F(i,\ j) - F(i-1,\ j)]^2}$ is the spatial column frequency, and M and N are the image sizes.

(3) SD measures the information richness of the image. The larger the value, the more dispersed the gray level distribution of the image, the higher the contrast, and the better the subjective visual quality of the image. The definition of SD is shown in formula (22).

$$SD = \sqrt{\frac{1}{M\times N}\sum_{i=1}^{M}\sum_{j=1}^{N}(F(i,\ j) - \mu)^2}, \tag{22}$$

where $F(i,\ j)$ and μ are the gray value and mean value of the image at $(i,\ j)$.

(4) CC measures the linear correlation between the source image and the fused image. The higher the value, the more similar the fusion image is to the source image. The definition of CC is shown in formula (23).

$$CC = \omega_A r_{AF} + \omega_B r_{BF}, \tag{23}$$

where $r_{XF} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}(X_{i,j}-\mu_X)(F_{i,j}-\mu_F)}{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(X_{i,j}-\mu_X)^2\sum_{i=1}^{M}\sum_{j=1}^{N}(F_{i,j}-\mu_F)^2}}$, μ_X and μ_F are the mean values of source image X and fused image F respectively, and ω_A and ω_B are the weight coefficients of r_{AF} and r_{BF} respectively.

(5) $Q^{AB/F}$ uses local metrics to evaluate the amount of edge information transferred from the source image to the fused image. The higher the value, the more prominent the edge of the fused image, and the better the fusion quality. The definition of $Q^{AB/F}$ is shown in formula (24).

$$Q^{AB/F} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}(Q_{AF}(i,\ j)\times\omega_A(i,\ j)+Q_{BF}(i,\ j)\times\omega_B(i,\ j))}{\sum_{i=1}^{M}\sum_{j=1}^{N}(\omega_A(i,\ j)+\omega_B(i,\ j))}, \tag{24}$$

where $(i,\ j)$ is the pixel position, Q_{AF} and Q_{BF} represent the edge intensity between source image A, B and fused image F respectively, and ω_A and ω_B represent the quantization weight of Q_{AF} and Q_{BF} respectively.

(6) MS-SSIM evaluates the structural similarity between the fused image and the source image from a multi-scale perspective. MS-SSIM can be better consistent with the human visual perception system, and its evaluation effect is usually better than SSIM. The larger the value, the more similar the structure between the fused image and the source image. The definition of MS-SSIM is shown in formula (25).

$$MS-SSIM = \left[l_S(A,\ B)\right]^{\alpha S}\prod_{i=1}^{S}[c_i(A,\ B)]^{\beta i}[s_i(A,\ B)]^{\gamma_i}, \tag{25}$$

where l is the comparative brightness between image A and B, c is the image contrast, s is the image structure, α, β and γ are the relative importance of adjusting the image brightness, contrast and structure, respectively, and S is the image scale.

Notably, for the above six indicators, the larger the value, the higher the quality of image fusion.

Experimental comparisons

Experimental results on the MFI-DRC dataset

Qualitative experiments: To verify the advantages of our GIPC-GAN model over other state-of-the-art algorithms, we selected three representative image pairs from the MFI-DRC dataset and performed qualitative analysis on them. The fusion results are shown in Figs. 4, 5 and 6. For the sake of observation, we selected two details in these three sets of image pairs for analysis, marked with red and green rectangles respectively, and zoomed in on these two details.

As can be seen from Figs. 4, 5 and 6, the GIPC-GAN method and the other 7 state-of-the-art algorithms can achieve better fusion results. But among all, the GIPC-GAN model has obvious advantages. First, the proposed method preserves the target intensity and texture details of the source image accurately, with clear textures at the boundaries of in-focus and de-focus regions. Second, the GIPC-GAN can better retain the edge contour of the source image, and the texture of the fused image is more prominent overall. While the other 7 algorithms have problems of blurring and texture loss at the boundaries of in-focus and de-focus regions. Three decision map-based methods, such as CNN, SESF and DSIFT, lose details at the boundaries of in-focus and de-focus regions due to decision map classification errors, while the other two traditional methods, BF and MWGF, have the problem of blurred details due to the limitations of extracting and fusing features. ACGAN and MFF-GAN, which are based on deep learning, will cause the imbalance of the two kinds of information retained in the fused image because their discriminator loss function only contains a single intensity loss term or gradient loss term. Therefore, the images fused by these two models will lose some contour and detail information to some extent.

These problems can be seen in the red and green rectangles marked in Figs. 4, 5 and 6. As can be seen from the enlarged part in Fig. 4, our model has the most prominent contours and clear texture details at the grass, while the contour texture is relatively blurry in the grass in other six methods. It can also be seen from the enlarged image in Fig. 5, our model has the most prominent contours at the soil walls marked with red boxes and the clearest textures at the trees marked with green boxes. And it can be found that in Fig. 6, our model has the clearest texture details at the grass and the most prominent edge contour at the wall slate, while the other 7 methods lose details to some extent, especially in CNN, SESF, ACGAN, MFF-GAN and DSIFT models, outlines at the wall slate are blurred. Overall, the GIPC-GAN model has the best fusion performance compared to other state-of-the-art algorithms.

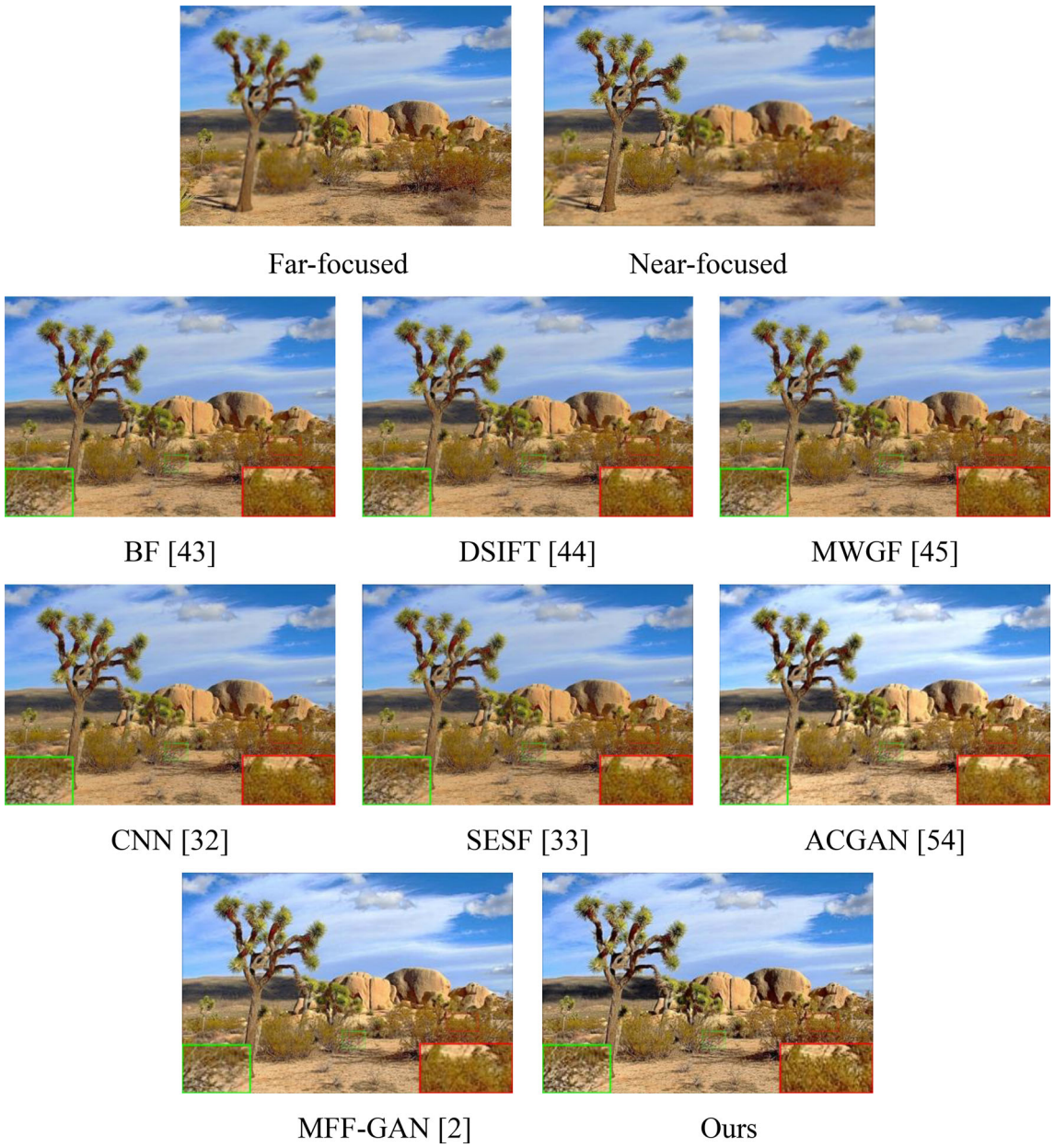

Fig. 4 Qualitative comparison of GIPC-GAN with 7 state-of-the-art methods on the "Desert plants" image pair from the MFI-DRC dataset

Quantitative experiments: To further verify the fusion advantage of the GIPC-GAN model, we quantitatively compare it with other 7 algorithms on the rest 30 image pairs in the MFI-DRC dataset. The results of various algorithms are shown in Fig. 7. For the convenience of observation, we highlight the mean values of the top three indicators in red, green and blue fonts respectively in the statistical table. As can be seen from Fig. 7, the GIPC-GAN model ranks first in the three statistical mean indicators of SF, SD and CC, second in the EN mean indicator, and third in the $Q^{AB/F}$ and MS-SSIM mean indicators, and it is only 0.0188, 0.003, and 0.0022 less than the top-ranked metrics on EN, $Q^{AB/F}$, and MS-SSIM metrics, respectively, which is very small.

Through careful analysis, it can be concluded that the images fused by the GIPC-GAN method have high contrast, clear images, with prominent contours and clear texture details, and have the most similar linear correlation with the source images. In addition, the images fused by the proposed algorithm have rich information, clear edge contours and high structural similarity with the source images, which are indistinguishable from the fusion performance of MFF-GAN, DSIFT and SESF models in these three aspects respectively.

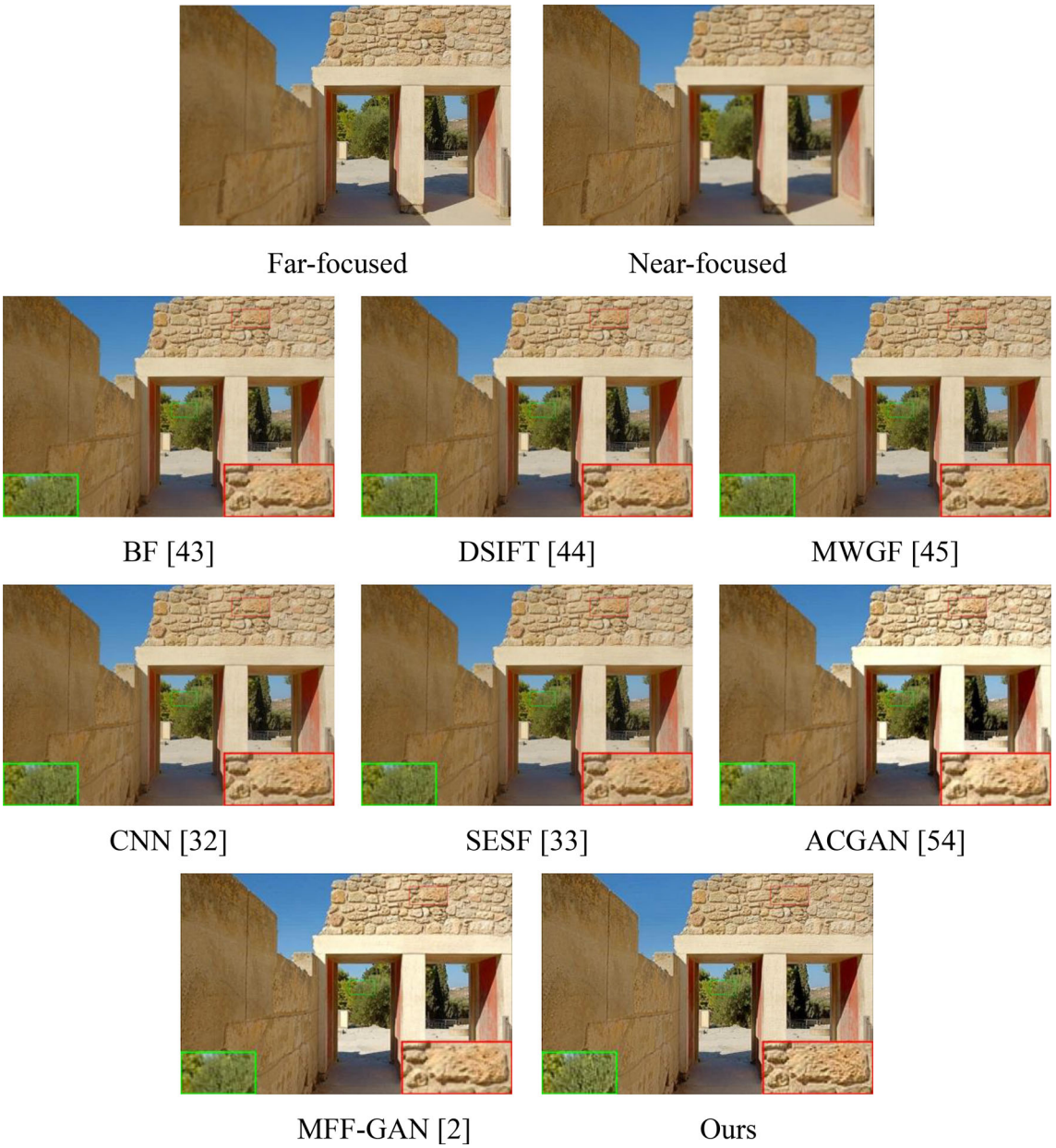

Fig. 5 Qualitative comparison of GIPC-GAN with 7 state-of-the-art methods on the "Dirt wall" image pair from the MFI-DRC dataset

Overall, the GIPC-GAN model performs the best in objective statistical evaluation.

Experimental results on the Lytro Dataset

Qualitative experiment: Another indicator to evaluate the quality of the model is the generalization of the model. In order to verify that the GIPC-GAN model has good generalization performance compared to other algorithms, we selected three representative image pairs in the Lytro dataset and conduct qualitative analysis on them. The fusion results are shown in Figs. 8, 9 and 10. For the convenience of observation, we selected two details in these three sets of image pairs for analysis, marked with red and green rectangles respectively, and zoomed in on these two details.

It can be seen from Figs. 8, 9 and 10 that the image fused by the GIPC-GAN model can retain as much target intensity and texture gradient information as possible, which is reflected on the multi-focus image with prominent target contour and clear texture details. Since our model adopts an end-to-end approach, without using decision maps, the fused images can better maintain regular textures near the boundaries of in-focus and de-focus regions. To be specific, the pipe on the

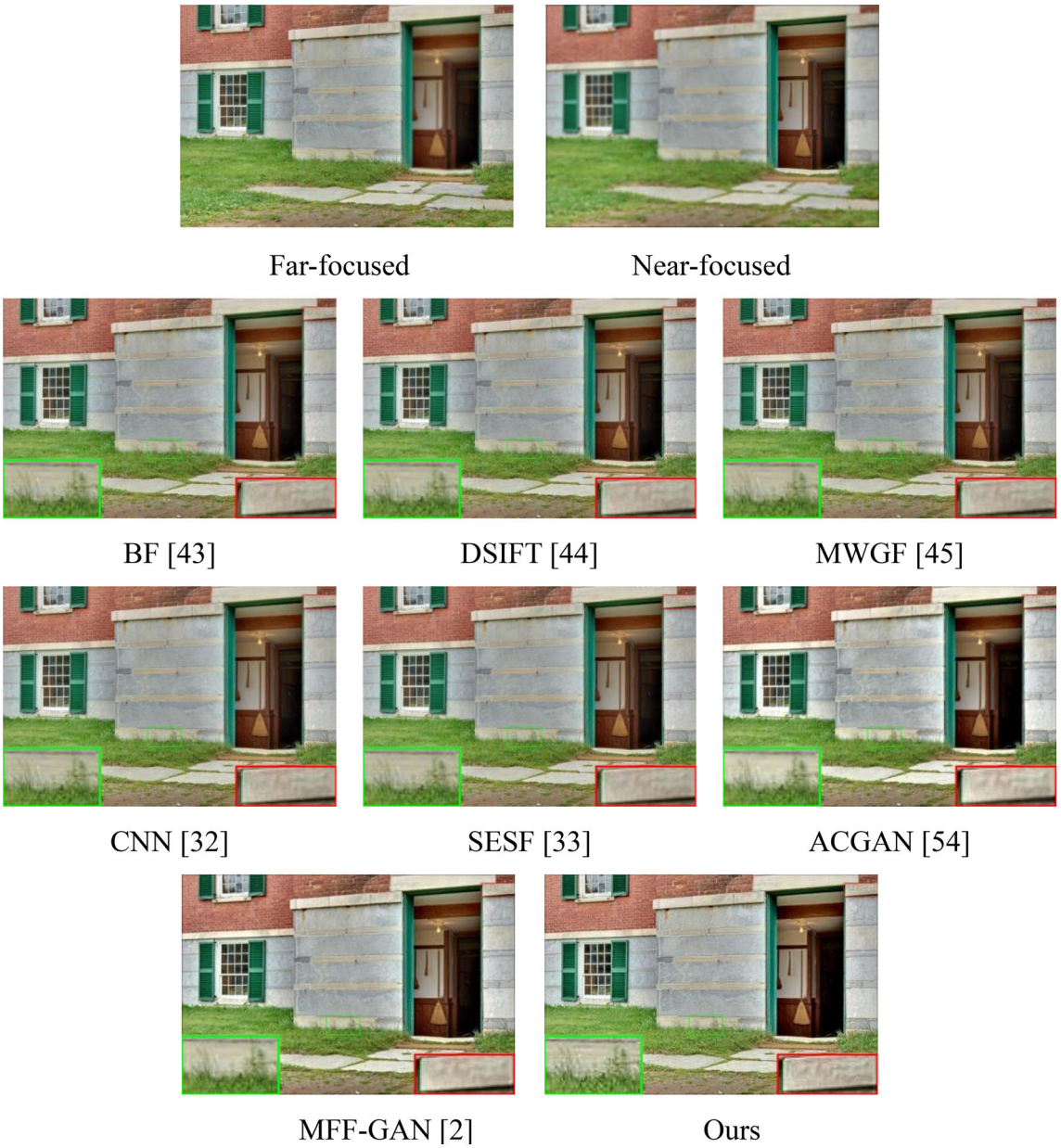

Fig. 6 Qualitative comparison of GIPC-GAN with 7 state-of-the-art methods on the "Apartment" image pair from the MFI-DRC dataset

ceiling in Fig. 8 and the golf ball in Fig. 9. For DSIFT, CNN, and SESF decision-map-based methods, they usually lose details near the boundary of in-focus and de-focus regions due to misclassification. For example, there lack of a rail of the pipe on the ceiling in Fig. 8, and a ball in golf ball area in Fig. 9. The traditional BF and MWGF methods also suffer from blurring or loss of details. For example, the golf ball in Fig. 9 appears blurry and the hat folds in Fig. 10 are lost in detail. In contrast, the ACGAN, MFF-GAN and GIPC-GAN algorithms achieve better fusion results. However, ACGAN and MFF-GAN also has the problem of loss of details, such as the indistinct texture and low contrast at the folds of the hat in Fig. 10. Overall, the images fused by the GIPC-GAN have the best subjective visual effect, high contrast and clear texture details. Compared with the other 7 algorithms, the GIPC-GAN model achieves the best generalization performance.

Quantitative experiments: To further verify the generalization performance of the GIPC-GAN model, quantitative comparison of 20 image pairs on the Lytro dataset with other 7 algorithms are conducted. The results are shown in Fig. 11. For convenience, we highlight the top three index means in the chart with red, green and blue fonts respectively.

As can be seen from Fig. 11, the GIPC-GAN model ranks first in the three statistical mean indicators of EN, SF, SD,

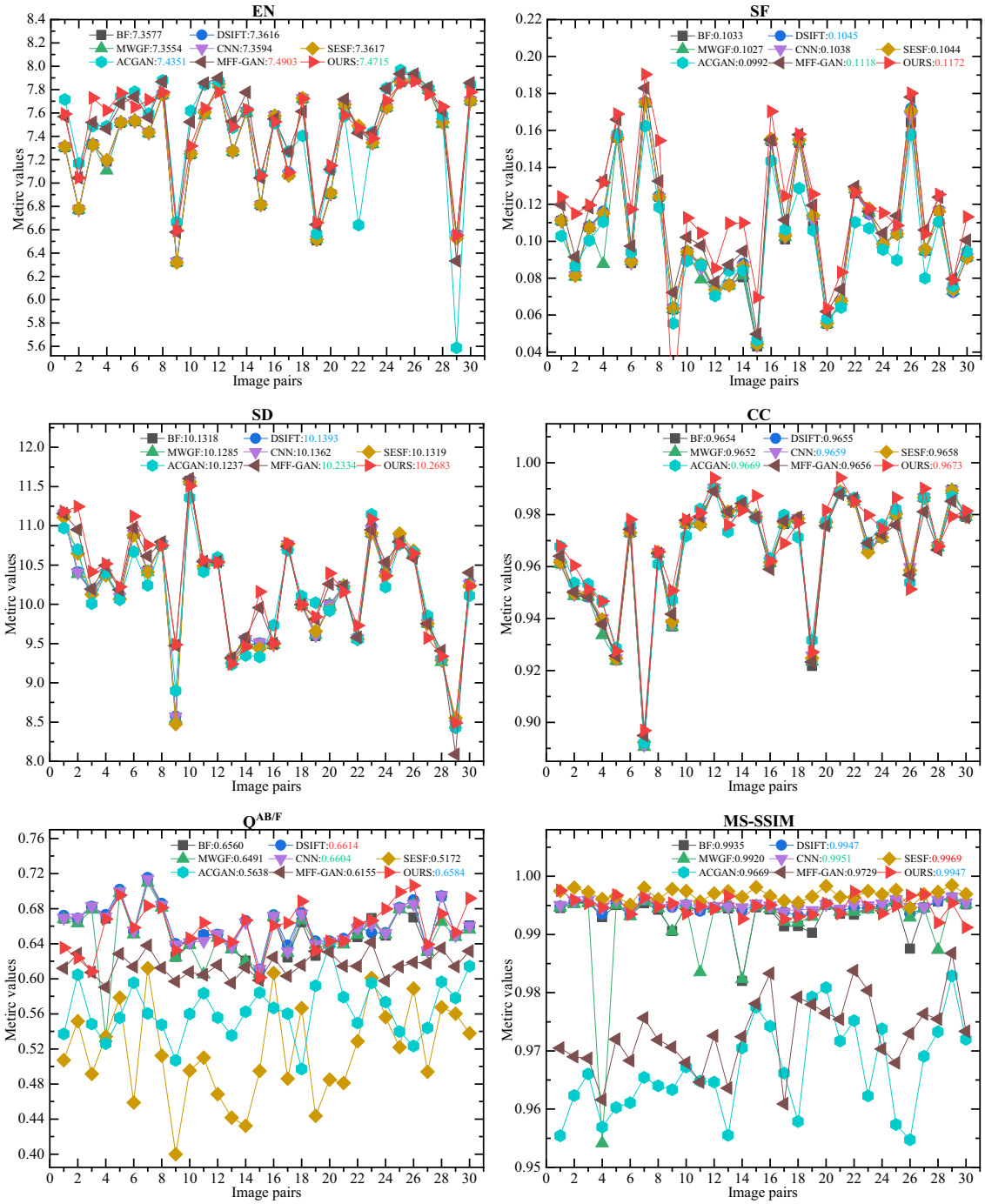

Fig. 7 Quantitative comparisons of the six metrics, on 30 image pairs from the MFI-DRC dataset

second in the MS-SSIM mean indicator, third in the CC mean indicator and Ranked fourth on the $Q^{AB/F}$ Means metric, and it is only 0.0039, 0.0076, and 0.0012 less than the top-ranked metrics on CC, $Q^{AB/F}$, and MS-SSIM metrics, respectively, which is very small. The results show that the GIPC-GAN algorithm also has the best fusion performance on the Lytro dataset. To sum up, our proposed GIPC-GAN model has better generalization performance than other 7 state-of-the-art algorithms.

Summary of experimental results

It can be seen from Section "Experimental results on the MFI-DRC dataset" and Section "Experimental results on

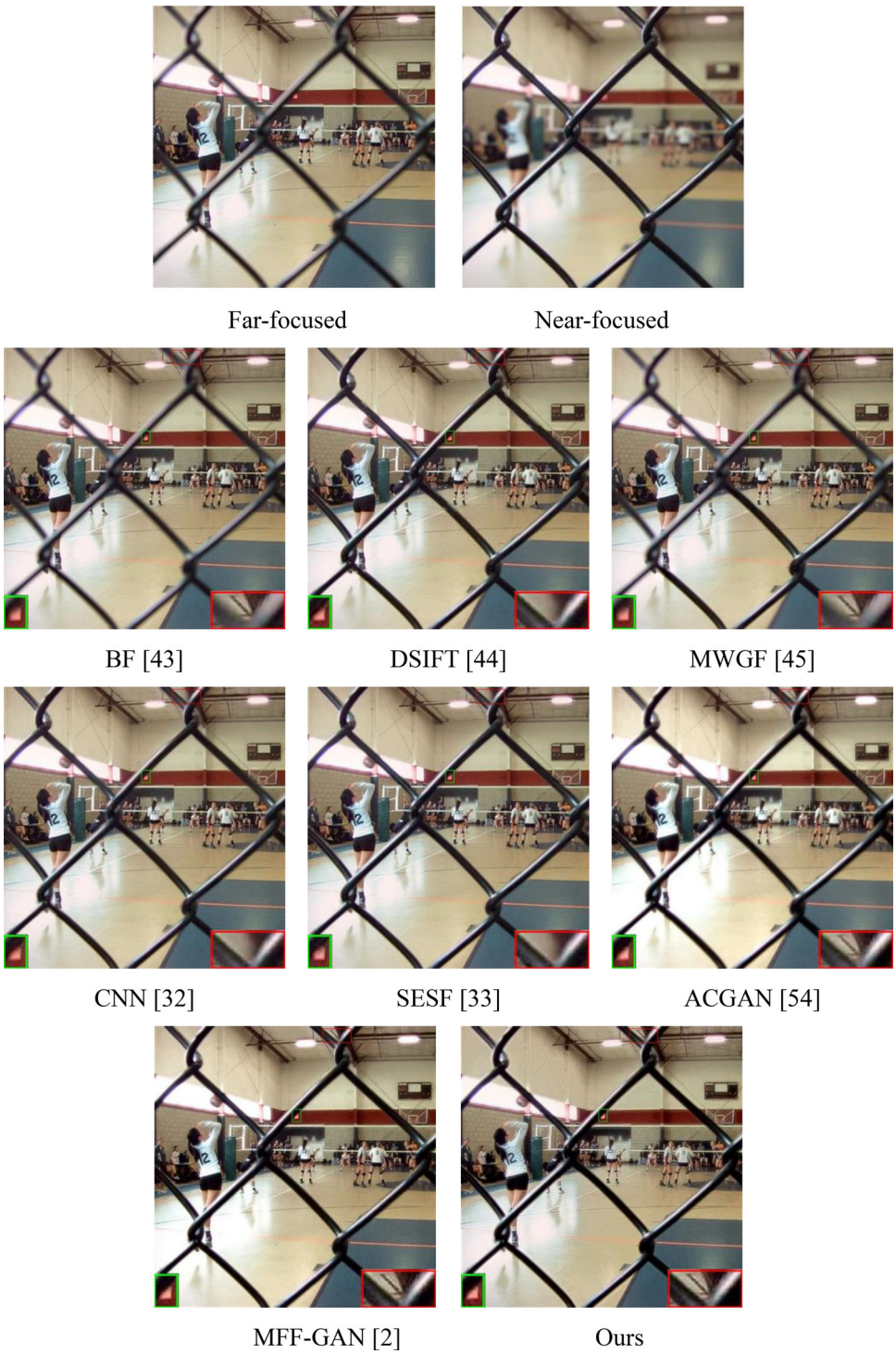

Fig. 8 Qualitative comparison of GIPC-GAN with 7 state-of-the-art methods on the "Volleyball court" image pair from the Lytro dataset

the Lytro Dataset" that our GIPC-GAN has achieved better fusion effect and generalization performance than other state-of-the-art models in MFI-DRC dataset and Lytro dataset, owing to the good network architecture design and a new joint proportional maintain constraint to adversarial loss function with gradient and intensity. Specifically, it can be summarized into the following 5 points:

(1) Our GIPC-GAN model is a fusion method based on global reconstruction, which helps eliminate the boundary blurring effect because it directly generates fusion images instead of decision maps. Therefore, there is almost no information loss and blurring in the focusing and defocusing boundary areas. However, the fusion methods of DSIFT, CNN and SESF based on decision

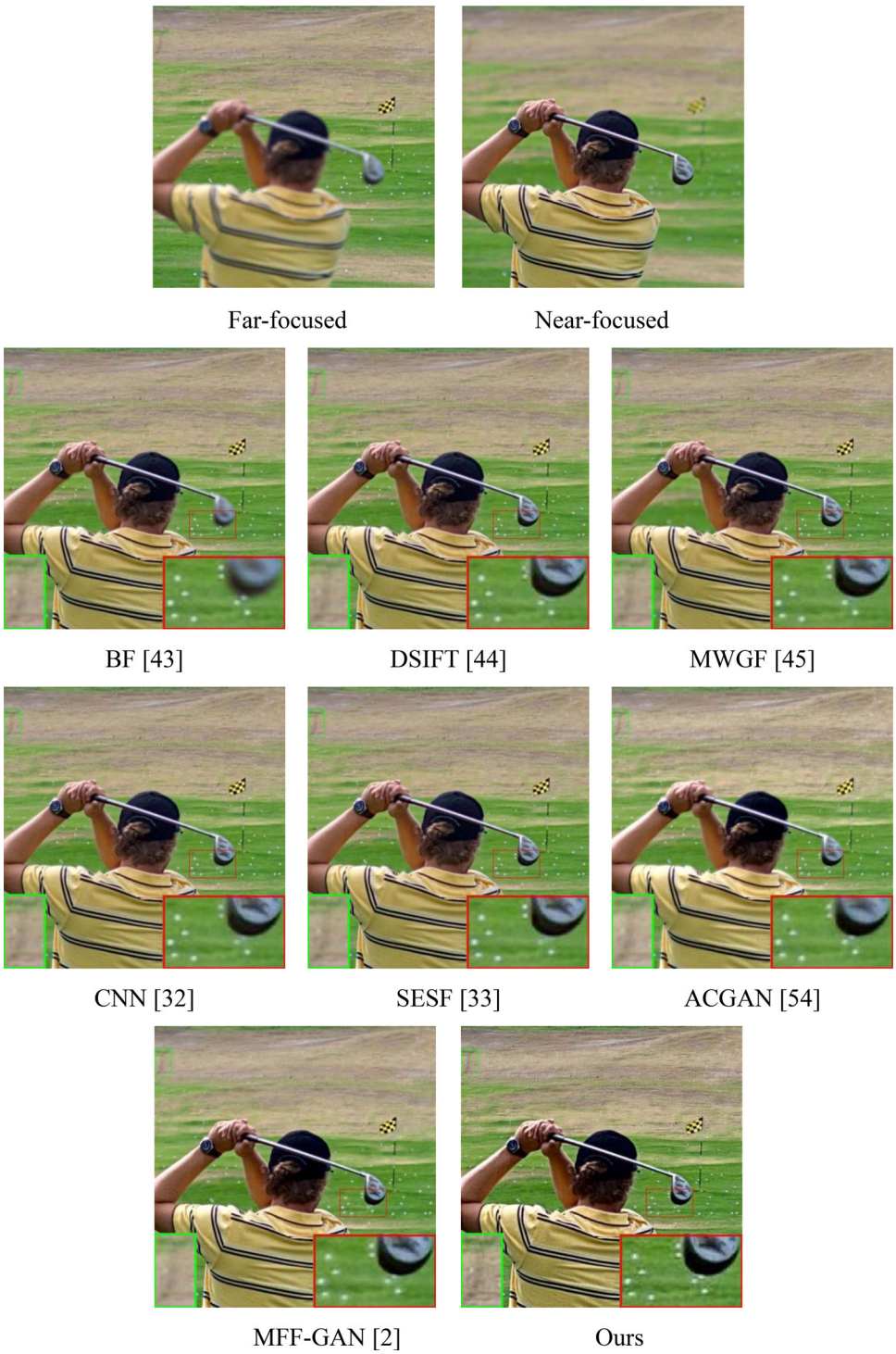

Fig. 9 Qualitative comparison of GIPC-GAN with 7 state-of-the-art methods on the "Golf course" image pair from the Lytro dataset

maps usually lose details near the boundary of focusing and defocusing regions due to classification errors, while the three traditional methods—BF, DSIFT and MWGF, have the problem of fuzzy details due to the limitations of activity level measurement and fusion features.

(2) The generator designed in our model is based on the architecture of encoder, feature generator and decoder. Compared with deep-learning-based SESF, ACGAN and MFF-GAN, its training model is more stable than that of the above methods. Therefore, the features extracted by GIPC-GAN model are more balanced.

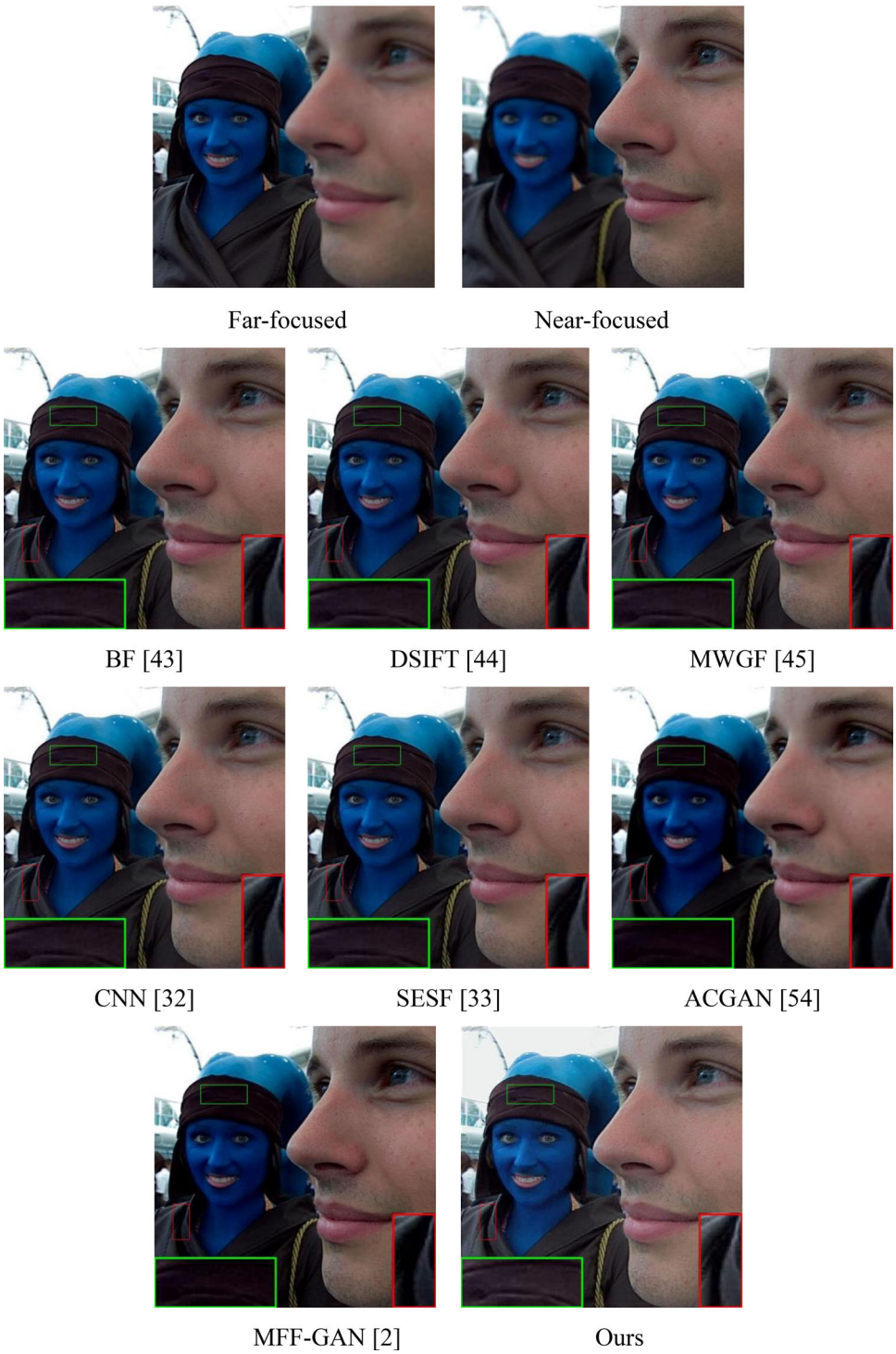

Fig. 10 Qualitative comparison of GIPC-GAN with 7 state-of-the-art methods on the "Amusement park" image pair from the Lytro dataset

(3) The most meaningful information in the multi-focus image fusion task is defined as texture gradient information and target intensity information, which sheds new light on the optimization of GIPC-GAN model.
(4) A new joint proportional maintain constraint adversarial loss function with gradient and intensity, is designed. The specific loss function is used to guide the training and optimization direction of the network and further enhance the target intensity and detail texture of the fused image while maintaining the balance of the target intensity and detail texture information retained in the fused image. However, ACGAN and MFF-GAN are based on GAN, because their discriminator loss function only contains a single intensity loss term or gradient

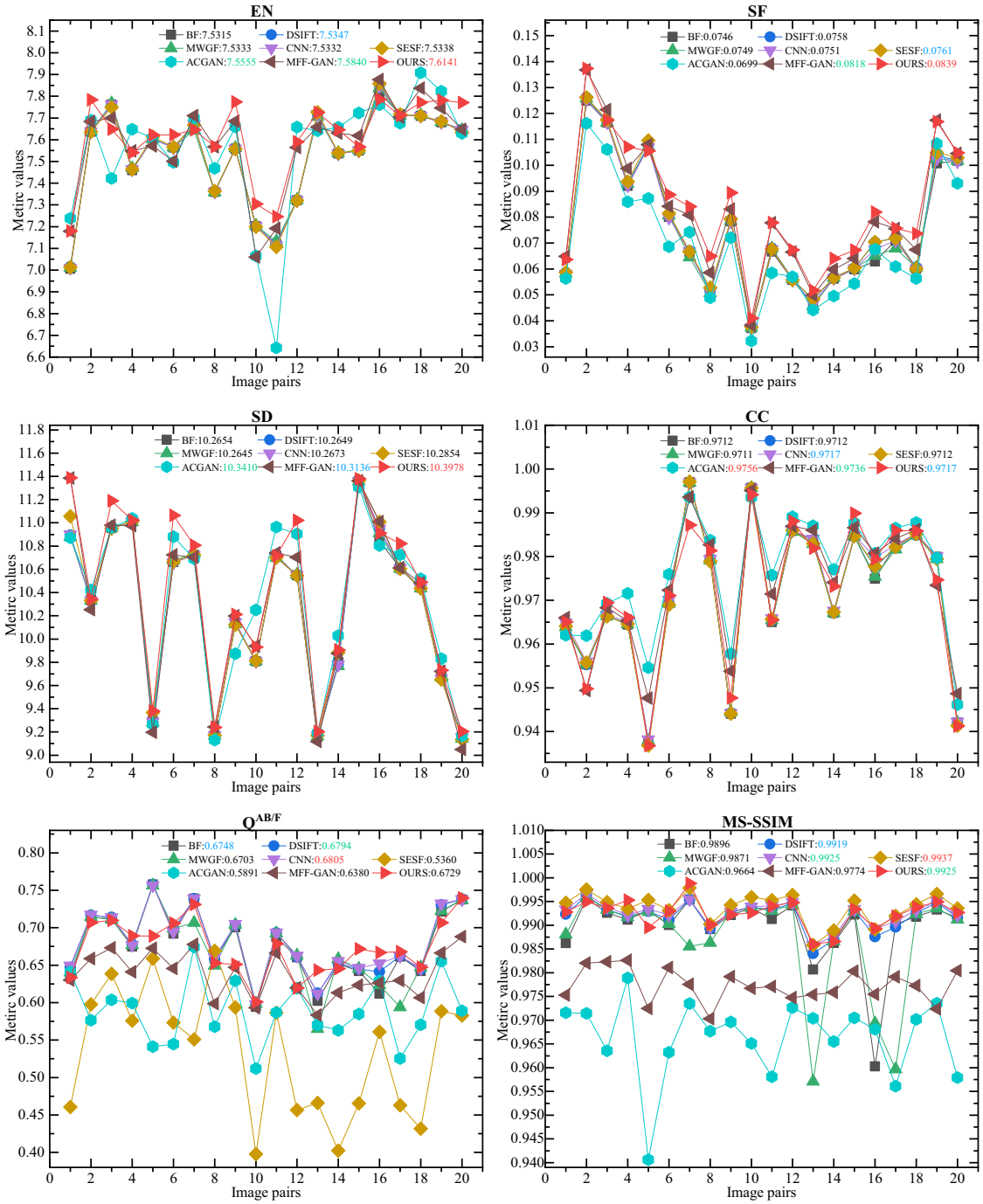

Fig. 11 Quantitative comparisons of the six metrics, on 20 image pairs from the Lytro dataset

loss term, it will cause the imbalance of the two kinds of information retained in the fused image.

(5) The network does not need any post-processing operations since there are no decision maps. GIPC-GAN can quickly achieve multi-focus image fusion in an end-to-end manner.

Model complexity analysis

As we all know, computational model complexity is viewed as an important measure to evaluate model performance in deep-learning-based methods. Specifically, the computational complexity of the model mainly includes the time complexity and the space complexity of the model. A quantitative comparative evaluation of the proposed GIC-GAN

and other state-of-the-art models is conducted from these two aspects.

Mode running efficiency comparison

In the image fusion methods based on deep learning [2, 13, 33], the time complexity based on the average running time of the algorithm is an important indicator to evaluate the quality of the model. In order to evaluate the GIPC-GAN model in a more comprehensive and objective way, we conduct comparative experiments on the running time of various advanced algorithms on the MFI-DRC dataset and Lytro dataset, as shown in Table 3. For the convenience, we highlight the top three statistics for average runtime in red, green, and blue fonts, respectively. Note that the traditional algorithms of BF, DSIFT and MWGF and deep-learning algorithm of CNN run on CPU, while SESF, ACGAN, MFF-GAN and the GIPC-GAN model operate on GPU.

It can be seen from Table 3 that our GIC-GAN algorithm achieves the second fastest running efficiency on MFI-DRC and Lytro datasets compared with the other 7 state-of-the-art comparison methods. On MFI-DRC dataset and Lytro dataset, GIC-GAN model is 0.1882s and 0.0985 s faster than ACGAN model which ranks first, respectively and the difference is very small. ACGAN, MFF-GAN and GIC-GAN are all based on end-to-end methods without post-processing, so the running time of these three algorithms is relatively short. SESF is a based decision maps method that requires post-processing operations, which will undoubtedly increase the running time of the algorithm. In the process of model training, MFF-GAN needs to calculate the weight screening maps of the source image, which will inevitably increase the running time of the algorithm. Due to the powerful graphics matrix acceleration capability of the GPU, it has advantages over the three traditional algorithms that run on the CPU: BF, DSIFT and MWGF. GIC-GAN model runs efficiently among the 8 fusion methods and basically meets the requirements of real-time image fusion tasks, when not taking the differences between CPU and GPU hardware environments into consideration.

Model parameters quantity comparison

As a classic and commonly used space complexity evaluation metric in deep learning-based methods [2, 55], the model parameter plays a vital role in evaluating the model performance of image fusion. As the image fusion method based on deep learning usually runs in GPU environment, we only compare and analyze the model parameters of SESF, ACGAN, MFF-GAN and GIC-GAN, which are four models based on deep learning. Generally, methods based on deep learning can be divided into two stages: training and testing. In order to analyze the space complexity of the models in a comprehensive way, we compared the parameters of the training model and the test model for these four models. The experimental results are shown in Table 4. For the convenience of observation, the top two statistical values of the number of model parameters are highlighted with red and green fonts respectively.

It can be seen from Table 4 that in the model training stage, our GIPC-GAN model parameters ranked third, 0.3403 M more than the SESF model parameters ranks first, and 0.0052 M more than the ACGAN and MFF-GAN model parameters ranks second. In the model testing phase, our GIPC-GAN model parameters ranked second, only 0.0022 M more than the ACGAN and MFF-GAN models, which ranks first, and the difference is very small. The parameters of SESF model reach 0.0748 M. Because the task of multi-focus image fusion is to integrate multiple partially focused source images into an all-focus fusion image, and this image fusion process is exactly completed in the test phase. Through comprehensive analysis, it can be concluded that the GIPC-GAN fusion model proposed by us has achieved a low complexity comparable to other state-of-the-art comparison models both in the model training stage and in the model testing stage. In terms of model space complexity, GIPC-GAN also basically achieves the performance of real-time image fusion.

Model ablation study

For the two important information of target intensity and texture gradient defined above, this paper designs a combined-constraint loss function based on the proportional information of intensity and gradient in the generator and the discriminator loss function. As far as we know, the existing fusion methods in multi-focus image or infrared and visible light image [2, 37], only take into consideration of the intensity information and gradient information of the retained source image in the content loss function of the model, but not in the discriminator loss function. Previous methods are of a certain degree of human subjectivity when setting the weight of the intensity and gradient loss in the content loss, and such a design will inevitably cause the imbalance of the intensity and gradient information retained in the fused image and the loss of detail information. To solve the above problem, we not only consider the intensity loss and gradient loss in the generator, but also put these two kinds of information in the discriminator. The ratio of intensity loss and gradient loss weights in the generator to the image pixel discrimination loss, and image gradient discrimination loss weights in the discriminator are required to follow a proportional maintenance strategy, i.e., the ratio of weights remains the same order of magnitude (10^1 order of magnitude). With such design, not only the generator is able to maintain a balance between the source image target intensity and gradient details during the process of fusing images, but also in

Table 3 The mean and standard deviation of running time of different methods on MFI-DRC and Lytro datasets (unit: second)

	MFI-DRC	Lytro
BF [41]	0.7221±0.3442	0.5190±0.0527
DSIFT [42]	5.0708±2.0453	3.6310±0.6978
MWGF [43]	1.9136±0.7291	1.5505±0.2849
CNN [30]	123.8767±41.0987	99.5241±1.9322
SESF [31]	0.4133±0.5306	0.3858±0.5970
ACGAN [54]	0.0344±0.0125	0.0282±0.0069
MFF-GAN [2]	0.2622±0.4059	0.1338±0.4950
Ours	0.2226±0.4039	0.1267±0.5056

Table 4 Analysis of model parameters based on four deep learning fusion methods (unit: M)

	Train	Test
SESF [31]	0.0748	0.0748
ACGAN [54]	0.4099	0.0382
MFF-GAN [2]	0.4099	0.0382
Ours	0.4151	0.0404

the process of discrimination, the discriminator could continuously guide and optimize the generator with balanced information of intensity and gradient so as to deceive the discriminator.

To verify the effectiveness of the combined-constrained loss function with proportional information of intensity and gradient, ablation experiments is conducted on the model. To be specific, in the ablation model, the intensity discrimination loss in fused image and the all-in-focus source image are removed in the discrimination loss function, i.e., set $\lambda_1 = \lambda_3 = 0$, with the name of Without_int model. Also, the gradient discrimination loss in the fused image gradient and the all-in-focus source image are removed in the discrimination loss function, that is, set $\lambda_2 = \lambda_4 = 0$, with the name of Without_gra model. To verify the efficiency of the designed combined-constraint loss function, subjective ablation studies on two multi-focus datasets—MFI-DRC and Lytro, are conducted. The subjective ablation results of the model are shown in Figs. 12 and 13.

Figures 12 and 13 show the subjective ablation comparison results of the GIPC-GAN model on a typical image pair of the MFI-DRC and Lytro datasets, respectively. For the convenience of analysis, we select two prominent details for analysis in Figs. 12 and 13, respectively, and mark them with red and green rectangles. It can be seen from Figs. 12 and 13 that, both Without_int and Without_gra and the GIPC-GAN model can achieve to a relatively ideal fused result. However, most of the results of the fusion of Without_int and Without_gra models have problems of texture blurring and loss of details. For example, at the window marked by the red box and the bench marked by the green in Fig. 12, the Without_int model loses the contour of the window railing and the stripe information in the middle of the bench because of its high brightness, while the Without_int model blurs the contour of the window railing and the stripe information in the middle of the bench because it retains too much intensity information of the source image. In Fig. 13, the hair texture marked by the red box and the watch outline marked by the green box also have the same problem as Fig. 12. It is because the Without_int model and the Without_gra model only focus on retaining either gradient or intensity information in the source image, resulting in an imbalance between the texture gradient and target intensity retained in the fused image. Therefore, these two models inevitably lose important information in the source image.

In contrast, our GIPC-GAN model can give attention to both the intensity and gradient information of the source image during the adversarial gaming, and make the fused image better balance and retain the target intensity and texture detail information in the source image and thus have better subjective visuals, which can be concluded from the red and green rectangles marked in Figs. 12 and 13. The desirable results of the GIPC-GAN model benefit from our designed combined-constraint loss function based on the proportional information of intensity and gradient. It indicates

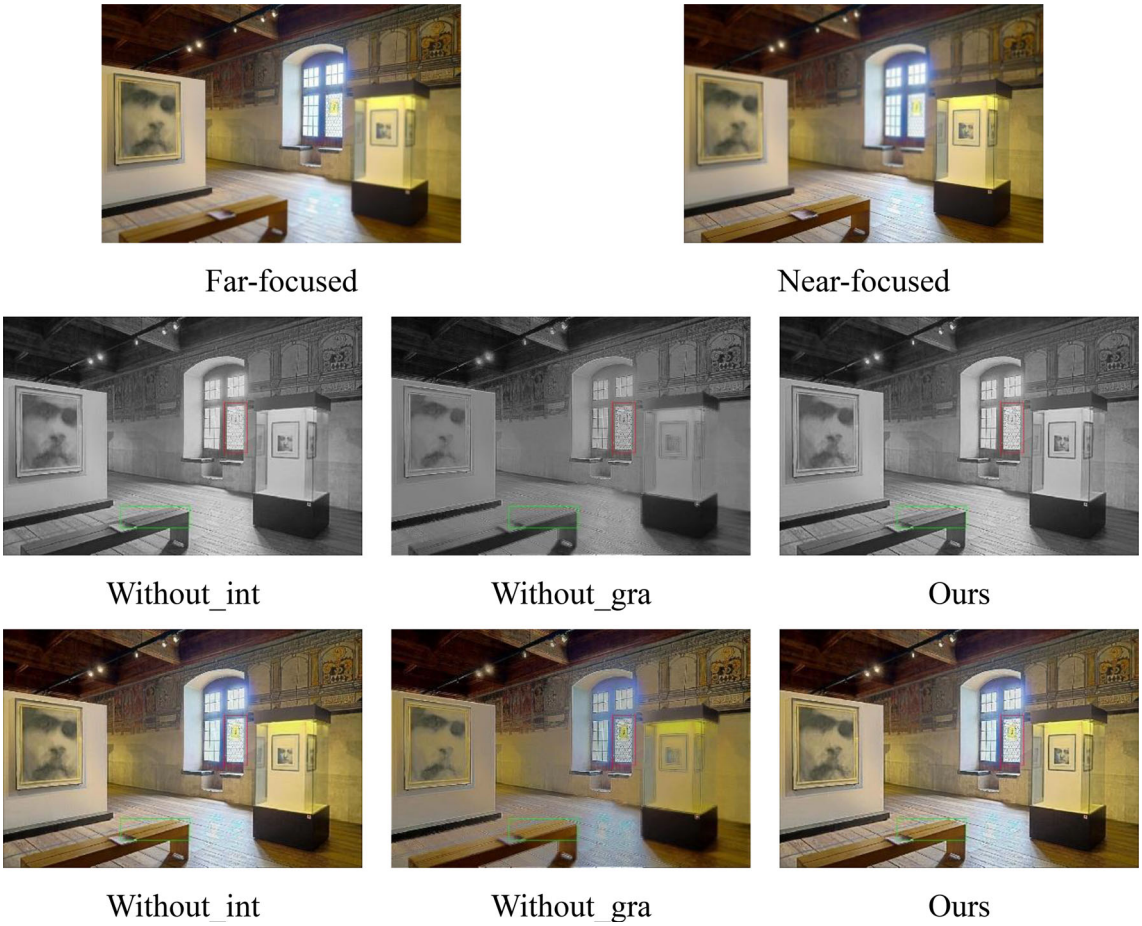

Fig. 12 Subjective experiments of ablation on the "Art gallery" image pair of the MFI-DRC dataset (The first row to the third row is the source image, the fused grayscale image and the fused RGB image respectively)

this specialized combined-constraint loss function plays a critical role in image fusion.

Fusion of multi-focus sequential image pairs

In order to verify that our GIPC-GAN model has good fusion performance and generalization on multi-focus sequence image pairs, we conducted fusion comparison experiments with other 7 algorithms on the multi-source multi-focus image sequence provided by Lytro. Specifically, fusion operations on the multi-source and multi-focus images are perform in turn, to obtain the final fused image. The experimental results are shown in Figs. 14 and 15. For convenience, two prominent details are selected for analysis in Figs. 14 and 15, respectively, and marked with red and green rectangles.

As can be seen from Figs. 14 and 15, both the GIPC-GAN model and the other 7 algorithms can achieve good fusion results on multi-source multi-focus image pairs. Yet, the images fused by the GIPC-GAN model have the best subjective visual effect and retain more target intensity and texture detail information of the source image, while the other 7 algorithms loss details loss or blurring. For example, at the red box marked in Fig. 14, the images fused by the BF, DSIFT, MWGF, CNN, ACGAN and SESF models have blurry logos and texts on the oxygen tank. At the green box marked in Fig. 14, the contours of the images fused by the CNN, SESF, ACGAN and MFF-GAN models are blurred at the island. Similar situation happens in Fig. 15. Through comprehensive analysis, conclusion can be drawn that the fusion result of the GIPC-GAN model has the clearest image and the best all-focus on the whole and has the best generalization on the multi-source multi-focus image sequence of the Lytro dataset.

Conclusions

In this paper, a novel gradient and intensity joint proportional constraint generative adversarial network (GIPC-GAN) is proposed for multi-focus image fusion. First, Deep Region Competition (DRC) algorithm is used to automatically generate decision maps and a set of labeled multi-focus image

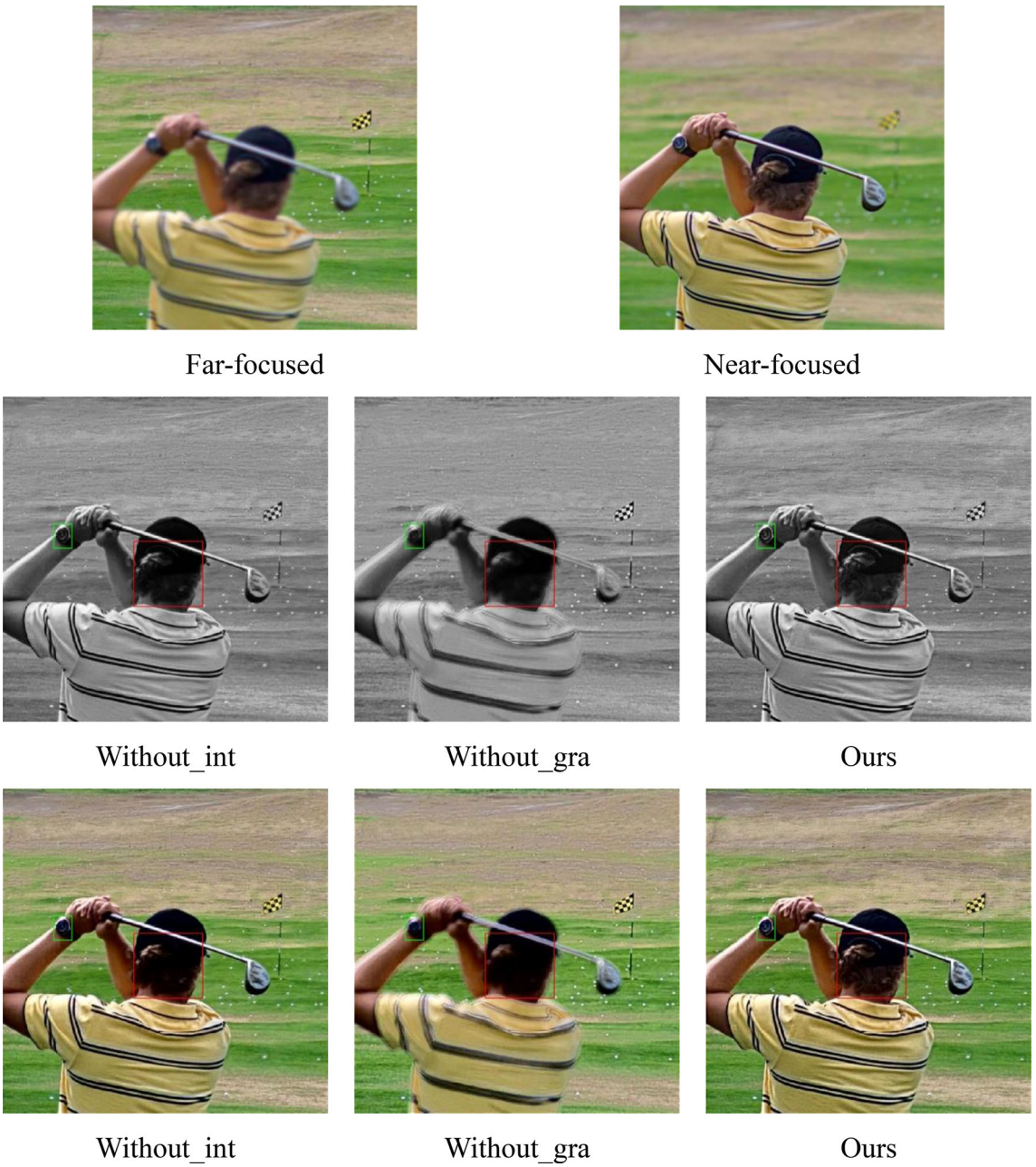

Fig. 13 Subjective experiments of ablation on the "Golf course" image pair of the Lytro dataset (The first row to the third row is the source image, the fused grayscale image and the fused RGB image respectively)

datasets on a public dataset is constructed, which avoids the boundary errors that occur in artificially constructing decision maps. Second, the most meaningful information for the multi-focus image fusion task is defined as the target intensity and detail gradient. With this regard, we propose a combined constraint loss function of proportional intensity and gradient. Third, we take the source image, the gradient map of the source image, the fusion image and the gradient of the fusion image as the input of the discriminator in the GAN network to further preserve the target intensity and detail information of the fusion image in a more balanced way. Final, through experimental verification on two multi-focus public datasets and a multi-source multi-focus image sequence dataset, GIPC-GAN model stands out among other state-of-the-art algorithms in terms of efficiency. It is worth noting that our GIC-GAN model has fast operation efficiency and low overall parameters, which basically meets the requirements of real-time image fusion.

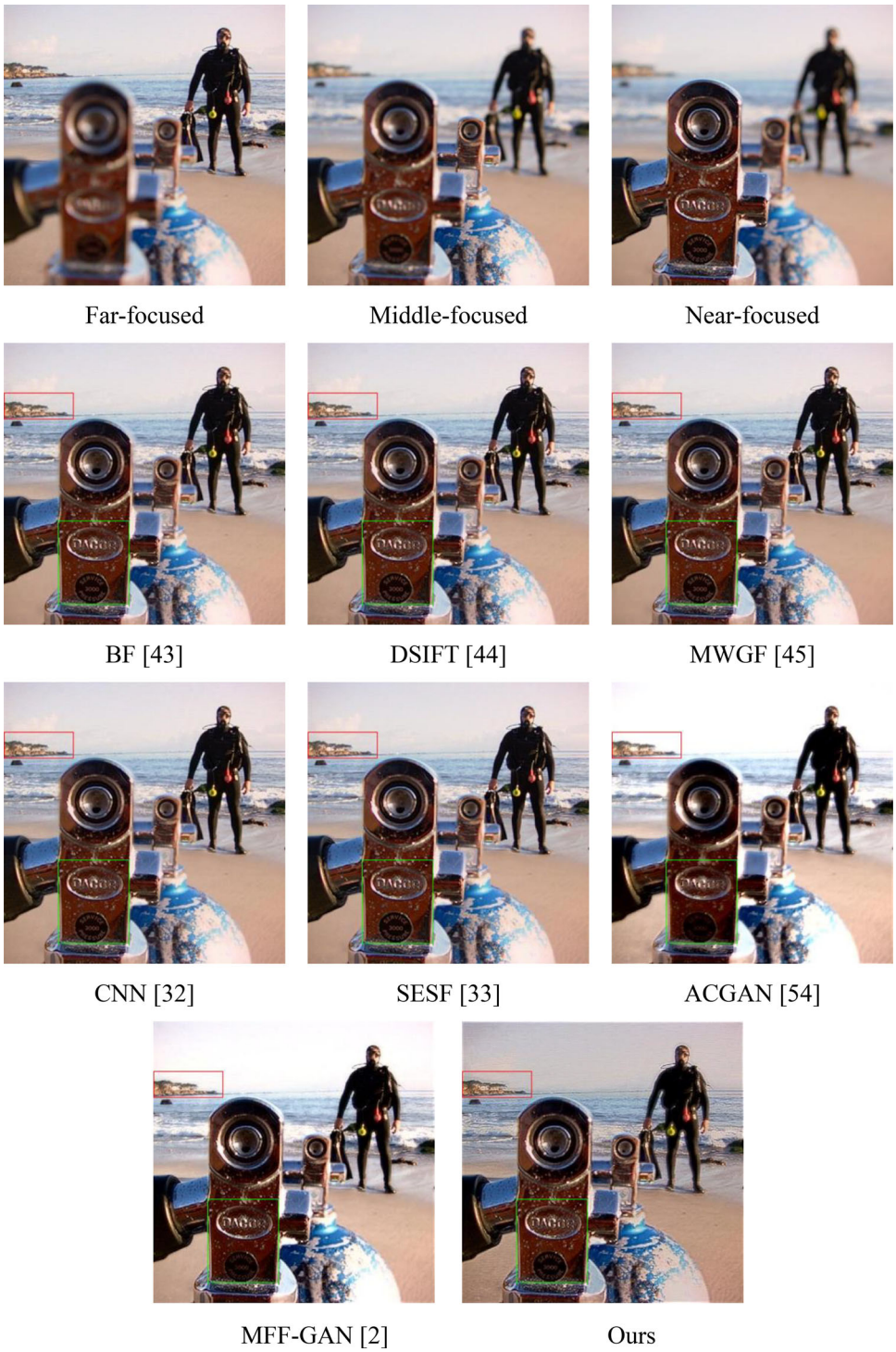

Fig. 14 Qualitative comparison of GIPC-GAN with 7 state-of-the-art methods on the first multi-focus image sequence pair on the Lytro dataset

The GIPC-GAN model proposed in this paper is consistent with most mainstream multi-focus image fusion methods and is also trained on noise-free images. At present, there are relatively few publicly available public multi-focus image datasets, and all datasets are noise free. So far, we have not conducted training or testing on multi-focus image pairs with noise and this makes our GIPC-GAN model have limitations in processing multi-focus source images with noise to some extent. In the future, we plan to conduct research on image denoising and image fusion as a whole, which it to design a unified model for image denoising and image fusion, as well as a specialized loss function for image denoising and image fusion to achieve the mutual promotion of image denoising and image fusion in multi-focus image fusion tasks.

Fig. 15 Qualitative comparison of GIPC-GAN with 7 state-of-the-art methods on the second multi-focus image sequence pair on the Lytro dataset

How to avoid the defocus diffusion effect at the boundary of the in-focus and de-focus regions, is another popular research issue in multi-focus image fusion, yet with little attention in most of the existing models. Among the existing methods, general image fusion methods [52, 53] have achieved good results not only in multi-focus images, but also in other image fusion tasks such as infrared and visible light images, medical images, and remote sensing images. Therefore, to design a general fusion network framework and take into account defocus diffusion problem in multi-focus image fusion, is put on future research plan, which can complete various image fusion tasks by using one general fusion framework.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grant 11673009. The authors would like to thanks Prof. K. Ji from Yunnan Observatory, Chinese Academy of Sciences, for their valuable comments and suggestions for this study.

Data availability The data used to support the findings of this study are available from the corresponding author upon request.

Declarations

Conflict of 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.

References

1. Li S, Kang X, Fang L et al (2017) Pixel-level image fusion: a survey of the state of the art. Inform Fusion 33:100–112
2. Zhang H, Le Z, Shao Z et al (2021) MFF-GAN: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion. Inform Fusion 66:40–53
3. Zhang H, Xu H, Tian X et al (2021) Image fusion meets deep learning: a survey and perspective. Inform Fusion 76:323–336
4. Ma J, Ma Y, Li C (2019) Infrared and visible image fusion methods and applications: a survey. Information Fusion 45:153–178
5. Dai Y, Song Y, Liu W et al (2021) Multi-focus image fusion based on convolution neural network for Parkinson's disease image classification. Diagnostics 11(12):2379
6. Basak H, Kundu R, Sarkar R (2022) MFSNet: a multi focus segmentation network for skin lesion segmentation. Pattern Recogn 128:108673
7. Liu D, Teng W (2022) Deep learning-based image target detection and recognition of fractal feature fusion for BIOmetric authentication and monitoring. Netw Model Anal Health Inform Bioinform 11(1):1–14
8. Ilesanmi AE, Ilesanmi TO (2021) Methods for image denoising using convolutional neural network: a review. Complex Intell Syst 7(5):2179–2198
9. Saleem S, Amin J, Sharif M, et al (2022) A deep network designed for segmentation and classification of leukemia using fusion of the transfer learning models, Complex Intell Syst 8:3105–3120
10. Li D, Peng Y, Guo Y, et al (2022) TAUNet: a triple-attention-based multi-modality MRI fusion U-Net for cardiac pathology segmentation. Complex Intell Syst 8:2489–2505
11. Wang J, Qu H, Wei Y et al (2022) Multi-focus image fusion based on quad-tree decomposition and edge-weighted focus measure. Signal Process 198:108590
12. Ma L, Hu Y, Zhang B et al (2023) A new multi-focus image fusion method based on multi-classification focus learning and multi-scale decomposition. Appl Intell 53:1452–1468
13. Wang Y, Xu S, Liu J et al (2021) MFIF-GAN: A new generative adversarial network for multi-focus image fusion. Signal Process Image Commun 96:116295
14. Liu Y, Wang L, Cheng J et al (2020) Multi-focus image fusion: a survey of the state of the art. Information Fusion 64:71–91
15. Zhang Y, Wei W, Yuan Y (2019) Multi-focus image fusion with alternating guided filtering. SIViP 13(4):727–735
16. Qiu X, Li M, Zhang L et al (2019) Guided filter-based multi-focus image fusion through focus region detection. Signal Process Image Commun 72:35–46
17. Bouzos O, Andreadis I, Mitianoudis N (2019) Conditional random field model for robust multi-focus image fusion. IEEE Trans Image Process 28(11):5636–5648
18. Zhang Z, Xi X, Luo X et al (2021) Multimodal image fusion based on global-regional-local rule in NSST domain. Multimed Tools Appl 80(2):2847–2873
19. Li X, Zhou F, Tan H et al (2021) Multi-focus image fusion based on nonsubsampled contourlet transform and residual removal. Signal Process 184:108062
20. Junwu L, Li B, Jiang Y (2020) An infrared and visible image fusion algorithm based on LSWT-NSST. IEEE Access 8:179857–179880
21. Yu L, Zeng Z, Wang H et al (2022) Fractional-order differentiation based sparse representation for multi-focus image fusion. Multimed Tools Appl 81(3):4387–4411
22. Tan J, Zhang T, Zhao L et al (2021) Multi-focus image fusion with geometrical sparse representation. Signal Process Image Commun 92:116130
23. Babahenini S, Charif F, Cherif F et al (2021) Using saliency detection to improve multi-focus image fusion. Int J Signal Imaging Syst Eng 12(3):81–92
24. Zhang B, Lu X, Pei H et al (2016) Multi-focus image fusion algorithm based on focused region extraction. Neurocomputing 174:733–748
25. Amin-Naji M, Aghagolzadeh A, Ezoji M (2019) Ensemble of CNN for multi-focus image fusion. Inform fusion 51:201–214
26. Li L, Si Y, Wang L et al (2020) A novel approach for multi-focus image fusion based on SF-PAPCNN and ISML in NSST domain. Multimed Tools Appl 79(33):24303–24328
27. Kong W, Miao Q, Lei Y et al (2022) Guided filter random walk and improved spiking cortical model based image fusion method in NSST domain. Neurocomputing 488:509–527
28. Ma X, Wang Z, Hu S (2021) Multi-focus image fusion based on multi-scale sparse representation. J Vis Commun Image Represent 81:103328
29. Li J, Li B, Jiang Y, et al (2022) MSAt-GAN: a generative adversarial network based on multi-scale and deep attention mechanism for infrared and visible light image fusion. Complex Intell Syst 8:4753–4781
30. Ma B, Yin X, Wu D et al (2022) End-to-end learning for simultaneously generating decision map and multi-focus image fusion result. Neurocomputing 470:204–216
31. Ma J, Le Z, Tian X et al (2021) SMFuse: multi-focus image fusion via self-supervised mask-optimization. IEEE Trans Comput Imaging 7:309–320
32. Liu Y, Chen X, Peng H et al (2017) Multi-focus image fusion with a deep convolutional neural network. Inform Fusion 36:191–207
33. Ma B, Zhu Y, Yin X et al (2021) Sesf-fuse: An unsupervised deep model for multi-focus image fusion. Neural Comput Appl 33(11):5793–5804
34. Li J, Guo X, Lu G et al (2020) DRPL: deep regression pair learning for multi-focus image fusion. IEEE Trans Image Process 29:4816–4831

35. Xiao B, Xu B, Bi X et al (2020) Global-feature encoding U-Net (GEU-Net) for multi-focus image fusion. IEEE Trans Image Process 30:163–175
36. Tang H, Xiao B, Li W et al (2018) Pixel convolutional neural network for multi-focus image fusion. Inf Sci 433:125–141
37. Zhang H, Xu H, Xiao Y, et al (2020) 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 (AAAI), 34(07), pp 12797–12804
38. Yu P, Xie S, Ma X et al (2021) Unsupervised foreground extraction via deep region competition. Adv Neural Inf Process Syst 34:14264–14279
39. Goodfellow I, Pouget-Abadie J, Mirza M, et al (2014) Generative adversarial nets. Adv Neural Inf Process Syst 27:2672–2680
40. Mao X, Li Q, Xie H, et al (2017) Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on computer vision (ICCV), 2017, pp 2794–2802
41. Ma J, Yu W, Liang P et al (2019) FusionGAN: A generative adversarial network for infrared and visible image fusion. Inform Fusion 48:11–26
42. Huang G, Liu Z, Van Der Maaten L, et al (2017) Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp 4700–4708
43. Zhang Y, Bai X, Wang T (2017) Boundary finding based multi-focus image fusion through multi-scale morphological focus-measure. Inform fusion 35:81–101
44. Liu Y, Liu S, Wang Z (2015) Multi-focus image fusion with dense SIFT. Inform Fusion 23:139–155
45. Zhou Z, Li S, Wang B (2014) Multi-scale weighted gradient-based fusion for multi-focus images. Inform Fusion 20:60–72
46. Nejati M, Samavi S, Shirani S (2015) Multi-focus image fusion using dictionary-based sparse representation. Inform Fusion 25:72–84
47. Roberts JW, Van Aardt JA, Ahmed FB (2008) Assessment of image fusion procedures using entropy, image quality, and multispectral classification. J Appl Remote Sens 2(1):023522
48. Eskicioglu AM, Fisher PS (1995) Image quality measures and their performance. IEEE Trans Commun 43(12):2959–2965
49. Rao YJ (1997) In-fibre Bragg grating sensors. Meas Sci Technol 8(4):355
50. Deshmukh M, Bhosale U (2010) Image fusion and image quality assessment of fused images. Int J Image Process (IJIP) 4(5):484
51. Wang Z, Simoncelli EP, Bovik AC (2003) Multiscale structural similarity for image quality assessment In: The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2 (2003), pp 1398–1402
52. Zhang Y, Liu Y, Sun P et al (2020) IFCNN: A general image fusion framework based on convolutional neural network. Inform Fusion 54:99–118
53. Zhang H, Ma J (2021) SDNet: A versatile squeeze-and-decomposition network for real-time image fusion. Int J Comput Vision 129(10):2761–2785
54. Huang J, Le Z, Ma Y et al (2020) A generative adversarial network with adaptive constraints for multi-focus image fusion. Neural Comput Appl 32(18):15119–15129
55. Liu Z, Lin Y, Cao Y, et al (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp 10012–10022