Fusion-UDCGAN: Multifocus Image Fusion via a U-Type Densely Connected Generation Adversarial Network

Yuan Gao, Shiwei Ma, Jingjing Liu, and Xianchao Xiu, *Member, IEEE*

***Abstract*— Multifocus image fusion has attracted considerable attention because it can overcome the physical limitations of optical imaging equipment and fuse multiple images with different depths of the field into one full-clear image. However, most existing deep learning-based fusion methods concentrate on the segmentation of focus–defocus regions, resulting in the loss of the details near the boundaries. To address the issue, this article proposes a novel generation adversarial network with dense connections (Fusion-UDCGAN) to fuse multifocus images. More specifically, the encoder and the decoder are first composed of dense modules with dense long connections to ensure the generated image's quality. The content and clarity loss based on the $L1$ norm and the novel sum-modified-Laplacian (NSML) is further embedded to provide the fused images retaining more texture features. Considering that the previous dataset-making approaches may lose the relation between the overall structure and the information near the boundaries, a new dataset, which is uniformly distributed and can simulate natural focusing boundary conditions, is constructed for model training. Subjective and objective experimental results indicate that the proposed method significantly improves the sharpness, contrast, and detail richness compared to several state-of-the-art methods.**

***Index Terms*— Dense connections, generation adversarial network (GAN), multifocus image fusion, novel sum-modified-Laplacian (NSML).**

I. Introduction

WHEN the foreground and background objects of the imaging scenes are spatially far apart, it is usually impossible to focus simultaneously due to the physical limitations of optical equipment depth of field and light field rendering, resulting in the blurring of the defocus regions [1]. In the cases of security surveillance and military analysis, the blurring can lead to the loss of important information, affecting the subsequent analysis. In this context, multifocus image fusion, as a method of image information enlargement, can fuse multiple images with different focusing regions to obtain a composite image containing all focusing regions, which has been widely used in various fields.

Multifocus image fusion technology has experienced rapid development, and many methods have been proposed and applied. Traditional methods include guided filter [1], fuzzy logic [2], [3], multiscale transformation [4], [5], and sparse representation [6], [7]. Convolutional neural network (CNN)-based methods include generating heat map or segmentation map methods [8], [9] and the end-to-end fusion method [10], [11]. In addition, given the remarkable success of the generation adversarial network (GAN) in image generation tasks, the research on GAN-based image fusion methods has become a hot topic in recent years. Zhang *et al.* [12] proposed a multifocus image fusion GAN based on decision block and continuous blur, which can accurately complete the extraction of clear regions and fusion. Ma *et al.* [13] introduced two discriminators to distinguish the authenticity of the fused image to maintain the maximum similarity between it and the two input images.

Although different fusion methods can obtain meaningful results, some problems still need to be solved. First, traditional and part of CNN-based methods require the manual design of fusion rules and activity level measurement methods. For example, the feature fusion layer in [11] needs to set the fusion weight manually. However, due to the diversity of fusion scenes, manual rules cannot cover all situations, thus affecting the quality of fusion. Second, the indirect deep learning-based multifocus image fusion method regards it as an image segmentation problem, extracts the focusing and defocusing regions, respectively, and then generates decision maps to guide the fusion. However, the focusing and defocusing regions obtained are often distinguished by clear boundaries, significantly different from the blurred boundaries in the actual situation, thus reducing the fusion quality. Third, in the end-to-end fusion method, the task goal of multifocus image fusion is to generate a composite image that is difficult to obtain, so it is essentially a weakly supervised or unsupervised task. Therefore, there is a dilemma: due to the lack of ground truth, it is difficult to carry out supervised learning, and the fusion quality of the unsupervised model is lower than that of the supervised model, which is due to the inadequate optimization objective. Fourth, GAN is difficult to train, and the adversarial loss is often difficult to converge, resulting in low fusion quality. The gradient-based content loss proposed in [14] aims to solve this problem. At the same time, the loss of information

Manuscript received November 4, 2021; revised January 25, 2022; accepted February 25, 2022. Date of publication March 16, 2022; date of current version April 11, 2022. This work was supported in part by the Natural Science Foundation of Shanghai under Grant 19ZR1420800 and in part by the State Key Laboratory of ASIC and System under Grant 2021KF009. The Associate Editor coordinating the review process was Shutao Li. *(Corresponding author: Shiwei Ma.)*

Yuan Gao, Shiwei Ma, and Xianchao Xiu are with the School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China (e-mail: marsness@shu.edu.cn; masw@shu.edu.cn; xcxiu@shu.edu.cn).

Jingjing Liu is with the State Key Laboratory of ASIC and System, School of Microelectronics, Fudan University, Shanghai 201203, China (e-mail: liujingjing@fudan.edu.cn).

Digital Object Identifier 10.1109/TIM.2022.3159978

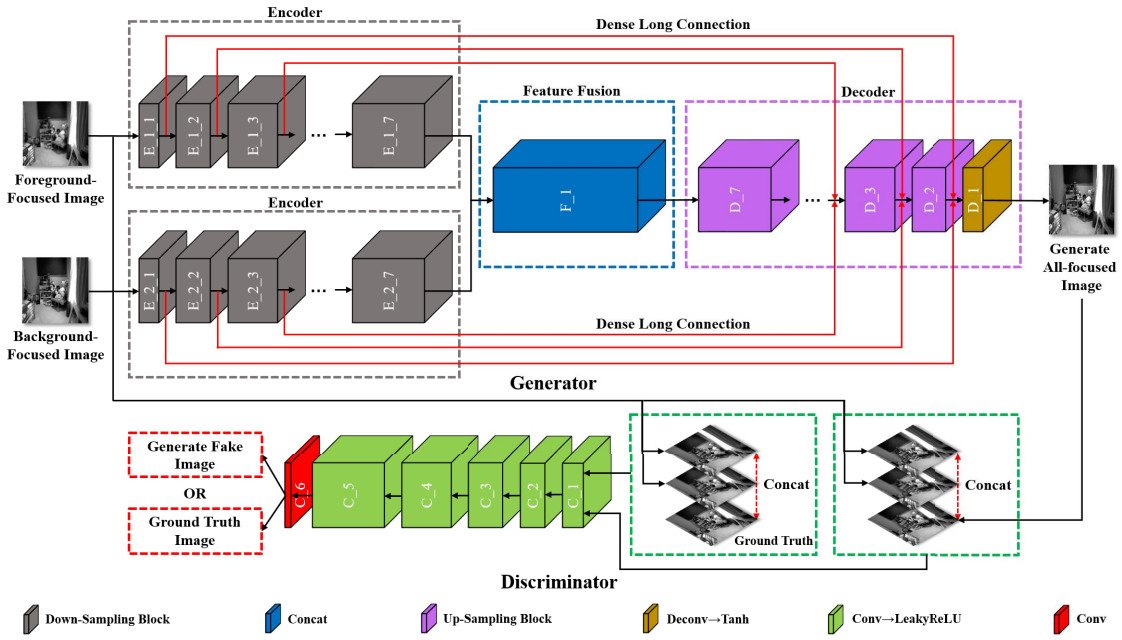

Fig. 1. Overview framework of Fusion-UDCGAN.

in the decoding and encoding process of the generator will also lead to the degradation of image quality.

This article proposes a multifocus image fusion method based on Z-type densely connected GAN to solve the above problems, called Fusion-UDCGAN. The generator of Fusion-UDCGAN encodes the foreground- and background-focused images, respectively, and inputs them into the decoder to obtain the fused images. Fusion-UDCGAN is an end-to-end image fusion method that completely avoids the drawbacks of manual design rules. The discriminator takes the fused image and the real image as input, respectively, and distinguishes the authenticity. To overcome the difficulties of network training of GAN, the method adds dense connections varying with depth during the encoding and decoding operations to increase the number of feature maps and reduce information loss. At the same time, inspired by U-Net [15], the proposed method establishes dense long connections and realizes feature reuse by concatenating low-level features with high-level features to provide more features in the decoding operation, thus further improving the fusion quality. Considering the characteristics of multifocus image fusion and further stabilizing the GAN training, the content loss based on the $L1$ norm and the clarity loss based on novel sum-modified-Laplacian (NSML) is added to make a better training effect. In addition, since the method is supervised, a new multifocus image dataset named CDMFI-SHU[1] based on depth images is made for training. The dataset dynamically segments the scene according to the depth and randomly adds different degrees of blur effect to the foreground and background scenes, respectively, to simulate the focus situation under different states. Guided filtering is also introduced to blur the focus boundaries in the dataset to restore the natural focus situations. Due to the restoration of natural focus situations, the information loss at the focus–defocus boundaries of the fused image is effectively reduced.

Based on solving the problems of the multifocus image fusion mentioned above, the contributions of this article can be summarized as follows.

1) *Novel Fusion-UDCGAN Model:* A new GAN-based supervised multifocus image fusion method with dense Z-type connection and U-type structure is proposed. The method densely encodes foreground- and background-focused images to generate the fused image through decoding in an end-to-end way. The overview of the method is shown in Fig. 1.
2) *Optimization Solver:* The content and clarity loss based on the $L1$ norm and NSML are proposed to force the generator to produce sharp images and minimize the loss of information.
3) *Application:* Since it is difficult to obtain the multifocus image dataset used for supervised training, a new multifocus image dataset CDMFI-SHU based on the depth images is designed.

The rest contents of this article are organized as follows. In Section II, related works, including GAN and its variants, U-Net, and DenseNet, are introduced. The proposed method is described in detail in Section III. Section IV is experiment and evaluation, including the detailed experimental setting and the generating algorithm of the CDMFI-SHU dataset. In addition, the subjective, objective evaluation, and ablation experiments are also elaborated in Section IV. Discussion and conclusions are given in Sections V and VI, respectively.

II. Related Works

A. Generative Adversarial Networks

GAN is first proposed by Goodfellow *et al.* [16] for image generation in 2014. The core concept of GAN is to establish the mapping relationship between the input samples and the

[1]CDMFI-SHU is an open-source dataset available on https://github.com/Marsness/CDMFI-SHU

target samples through the maximum–minimum game so that the network can learn from the former to the generation of the latter without *a priori* assumptions. In recent years, GAN has achieved remarkable results in image fusion [17]–[19].

The original GAN uses a set of random noise z as input, but the effect is unstable and only applicable to generating random results. Later, some improved models, such as deep convolutional GANs (DCGANs) [20] and least-squares GAN (LSGAN) [21], try to introduce the funnel-type networks to take the image as the input and then establish the image-to-image mapping relationship by subsampling and upsampling operation. The introduction of funnel-type networks does the translation between images possible, which lays a foundation for the introduction of GAN in image fusion.

B. U-Net

U-Net is an image generate method proposed by Ronneberger *et al.* [15], which is proposed to generate semantic segmentation maps in medical image segmentation and achieved impressive results. Given the excellent performance of U-Net in the field of image segmentation, more studies try to modify it to other image generation tasks and achieve good results [22]–[24].

Inspired by the full convolutional neural network (FCN) [25], U-Net adds skip connections. The skip connections concatenate the downsampling and upsampling operations of the same layer, instead of directly monitoring and loss backpropagation on the high-level semantic features used in the funnel-type network of GAN. This approach ensures that the generated images contain more low-level features, which reduces the loss of features in the deep networks.

C. DenseNet

Inspired by ResNet [26] and inception networks [27], Huang *et al.* [28] proposed DenseNet in 2017 and achieved better results than ResNet. Subsequently, more studies introduced DenseNet as backbones for various image tasks, further verifying its effectiveness [29]–[31].

The core idea of DenseNet is similar to that of ResNet, but it establishes dense connections between all the layers in front and the layers behind, and through it to achieve more efficient feature reuse, so that it achieves better performance than the latter with fewer parameter and computing costs. DenseNet consists of three main components: dense block, bottleneck, and transition layer. The dense block is the core of DenseNet, which provides the networks with dense connections.

III. Proposed Method

A. Overview of the Proposed Method

The significance of multifocus image fusion is to extract the meaningful information of foreground- and background-focused images generated by the difference in focus depth and fuse them into a composite image with more information. Specifically, the significant information of multifocus images refers to the sharp regions in the images, and the most critical work in the fusion process is to accurately extract them from multifocus images and discard the fuzzy regions. Therefore, it is necessary to introduce an appropriate mechanism to encode the foreground- and background-focused images, respectively, and extract and eliminate the codes of the focus and defocus regions according to certain constraints, providing a necessary premise for accurate fusion. In addition, due to the complexity of the edge of the focus and defocus regions, multifocus image fusion should be different from the image generation task, which generates smooth and determined edges, such as image segmentation. Considering the above situation, the overall framework of the proposed method is shown in Fig. 1.

First, a Siamese neural network without weight sharing is introduced as an encoder to extract and encode features of foreground- and background-focused images, respectively. Then, the codes of the two branches are input into the feature fusion layer, which concatenates them with the 1×1 convolution kernels. After that, the fusion codes are passed through the decoder to generate the all-clear image. In addition, the proposed method establishes dense long connections between the downsampling blocks and the upsampling blocks to combine the low- and high-level features, and improves the efficiency of feature reuse to decrease the information loss in the downsampling process and increase the information amount of the features in the upsampling process. Unlike the long connections in U-net, the dense long connections transfer more low-level features through the dense blocks.

Fusion-UDCGAN uses a discriminator with six convolutions layers, and an adversarial game is established between it and the generator. In particular, the proposed method first concatenates the input multifocus images with the generated images and the ground truth, respectively, to produce two sets of high-dimensional images. Inspired by cGAN [32], this approach can eliminate the interference of input observations to the discriminator, thus improving the performance of it, which has been verified in [33]. Then, the discriminator is forced to distinguish the two during training according to the adversarial loss. Finally, in the continuous adversarial game, the generator generates images that can fool the discriminator, which can no longer distinguish it from the ground truth.

B. Network Architecture

1) Generator Architecture: The detailed architecture of the generator in Fusion-UDCGAN is shown in Fig. 2(a). The downsampling block is an essential module of Fusion-UDCGAN, which consists of a 3×3 convolution layer and a dense block, and uses the ReLU as the activation function. Compared with the pooling layer in U-Net, the downsampling block can generate more feature maps, reduce the information loss in the downsampling process to make the network deeper, and will not be affected by the size of input images. In the trunk network, the encoded features are decoded by seven upsampling blocks, which have the opposite structure to the downsampling blocks. In addition, the final fused image is obtained using the Tanh activation function at the last layer of the trunk network. It is worth noting that the dense blocks in this article include the dense layer and the translation layer

TABLE I
DETAILED PARAMETER SETTING OF G AND D

		Module	Down-sampling block											
			Conv					Dense Block						Activation
			I_c	O_c	K_s	S_s	P_s	I_c	O_c	T_c	G_r	B_s	D_n	
G	Encoder	E_1_1, E_2_1	1	64	7	2	1	64	160	80	32	4	3	ReLU
		E_1_2, E_2_2	80	128	3	2	1	128	224	112	32	4	3	ReLU
		E_1_3, E_2_3	112	256	3	2	1	256	448	224	32	4	6	ReLU
		E_1_4, E_2_4	224	512	3	2	1	512	704	352	32	4	6	ReLU
		E_1_5, E_2_5	352	512	3	2	1	512	896	448	32	4	12	ReLU
		E_1_6, E_2_6	448	512	3	2	1	512	896	448	32	4	12	ReLU
		E_1_7, E_2_7	448	512	3	2	1	-	-	-	-	-	-	ReLU
			Up-sampling block											
			ConvTranspose					Dense Block						
	Decoder	D_1	512*2	512	3	2	1	-	-	-	-	-	-	ReLU
		D_2	448*3	512	3	2	1	512	896	448	32	4	12	ReLU
		D_3	448*3	512	3	2	1	512	896	448	32	4	12	ReLU
		D_4	352*3	256	3	2	1	512	704	352	32	4	6	ReLU
		D_5	224*3	128	3	2	1	256	448	224	32	4	6	ReLU
		D_6	112*3	64	3	2	1	128	224	112	32	4	3	ReLU
		D_7	80*3	1	7	2	1	64	160	80	32	4	3	Tanh
D			Conv					-						
		C_1	1	64	7	2	1	-						LeakyReLU
		C_2	64	128	3	2	1							LeakyReLU
		C_3	128	256	3	2	1							LeakyReLU
		C_4	256	512	3	2	1							LeakyReLU
		C_5	512	512	3	1	1							LeakyReLU
		C_6	512	1	3	1	1							-

* Conv, ConvTranspose indicates the convolution and transposed convolution layer, respectively. C_i, C_o, K_s, S_s, and P_s represent input channels, output channels, kernel size, padding size of Conv or ConvTranspose, respectively. D_i, D_o, G_r, B_s, and D_n represent input channels, output channels, growth rate, bn size, and the number of dense layers of dense blocks, respectively. T_o represents the output channels of the transition layer, which is half of the output channels of dense layers.

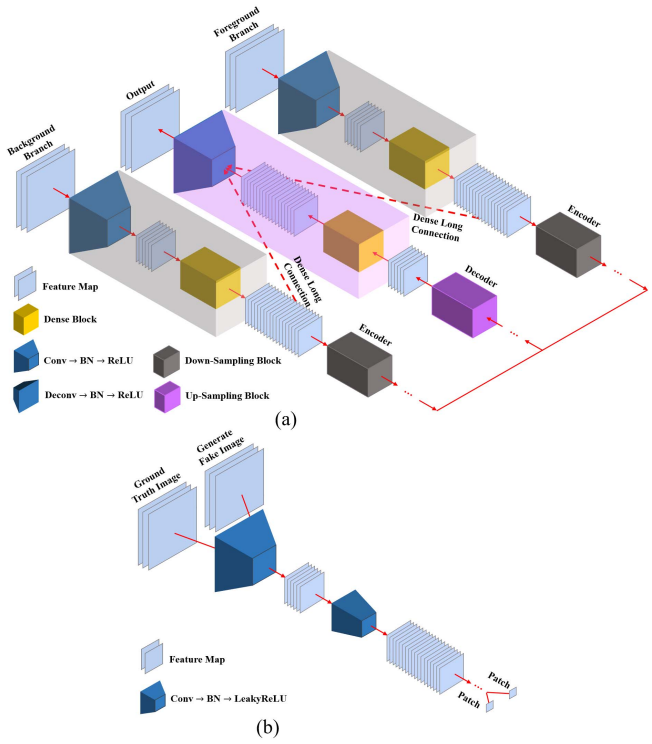

Fig. 2. Architecture of (a) generator and (b) discriminator.

in [28]. The transition layer in the dense block reduces the number of input feature maps by half to reduce the number of network parameters. The detailed parameter setting for each layer of the generator can be found in Table I.

Unlike U-net, Fusion-UDCGAN builds a denser and more efficient long connection between low- and high-level features. Specifically, the dense long connections concatenate the dense feature maps generated by the downsampling block at the lower level with the dense feature maps generated by the dense block in the upsampling block at the higher level. Then, input them into the transposed convolution layer (because the downsampling block and the upsampling block have opposite construction, the dense long connection appears Z-type) to provide dense low-level information for transposed convolution to generate images that contain both low- and high-level information, improve the quality, and accelerate the convergence speed.

2) Discriminator Architecture: The design of the discriminator is based on patchGAN in [33], and the detailed architecture is shown in Fig. 2(b). The discriminator mainly contains six 3×3 convolution layers and uses LeakyReLU as the activation function. The generated fake image and the ground-truth image are input into the discriminator in turn. After the convolution operation, two corresponding patches are generated, and the loss between them is calculated. Unlike traditional methods, patchGAN does not use the linear layer as the end of the network but uses the output patches to form a Markov random field to calculate the loss, assuming independence between pixels separated by more than a patch diameter. As a kind of texture and detail loss, patchGAN is more suitable for judging the focus regions of multifocus images (through the degree of clarity and the amount of texture information) than traditional methods. The detailed parameter setting for each layer of the discriminator can be found in Table I.

C. Loss Functions

1) Generator Loss: The generator loss $\boldsymbol{L}_G$ has three parts: the adversarial loss $\boldsymbol{L}_{\text{adv}}$, the content loss $\boldsymbol{L}_{\text{cont}}$, and the clarity loss $\boldsymbol{L}_{\text{clar}}$, which can be expressed as

$$\boldsymbol{L}_G = \boldsymbol{L}_{\text{adv}}(G) + \alpha_1 \boldsymbol{L}_{\text{cont}}(G) + \alpha_2 \boldsymbol{L}_{\text{clar}}(G) \tag{1}$$

where α_1 and α_2 are weights of content loss and clarity loss, respectively, to adjust the importance.

Adversarial loss is the classic loss in GAN, which is set to extract texture features. The proposed method introduces the generator adversarial loss with label proposed in LSGAN [20], which is expressed as follows:

$$\boldsymbol{L}_{\text{adv}}(G) = \frac{1}{N}\sum_{n=1}^{N} \log(D(x_1, x_2, G(x_1, x_2)) - a) \tag{2}$$

where x_1 and x_2 are the input multifocus images, N represents a batch of input images, and a is the possibility that the generated image will be judged as real by the discriminator. It is important to note that the input images and the generated image are concatenated and input into the discriminator to exclude the error caused by the former to the loss. The adversarial loss constraint generator produces images more similar to the ground truth in terms of texture and style, but it cannot make it close to the ground truth in terms of content. Therefore, this article proposes to use the $L1$ loss as the content loss of the generator

$$\boldsymbol{L}_{\text{cont}}(G) = \frac{1}{N}\sum_{n=1}^{N} \|y - G(x_1, x_2)\|_1 \tag{3}$$

where $\|\cdot\|_1$ is to calculate the $L1$ norm and y is the ground truth. The content loss forces the generated image to be as close as possible to the ground truth in the $L1$ sense, which is necessary for multifocus images fusion because it encourages less blur.

The key of multifocus image fusion is to fuse the sharp regions of the images, that is, to combine the high-frequency subbands. The content loss can promote the reduction of fuzzy regions to a certain extent, but it cannot meet the needs of high-frequency subbands' information restoration of multifocus images fusion. Therefore, based on NSML, this article proposed the clarity loss

$$\boldsymbol{L}_{\text{clar}}(G) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{H \times W} \times \sum_{h=1}^{H}\sum_{w=1}^{W} \begin{pmatrix} \text{NSML}(y(h,w)) \\ -\text{NSML}(G_{(x_1,x_2)}(h,w)) \end{pmatrix}^2 \tag{4}$$

where NSML$(\cdot)$ represents computing the NSML value of an image whose size is $H \times W$ on a certain pixel (h, w)

$$\text{NSML}(h,w) = \sum_{p=-P}^{P}\sum_{q=-Q}^{Q} \omega(p,q)[ML(h+p, w+q)]^2 \tag{5}$$

$$\begin{aligned} ML(h,w) = {} & |I(h,w) - I(h+\text{step}, w)| \\ & + |I(h,w) - I(h-\text{step}, w)| \\ & + |I(h,w) - I(h, w+\text{step})| \\ & + |I(h,w) - I(h, w-\text{step})| \end{aligned} \tag{6}$$

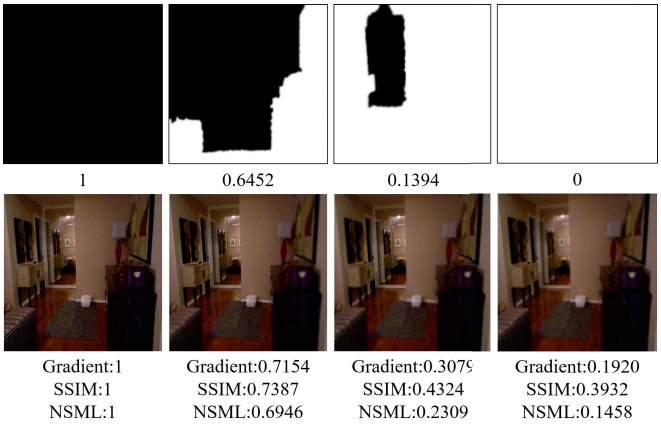

Fig. 3. Comparison of gradient, SSIM, and NSML of multifocus images.

where (h, w) is a pixel of the input image, $\omega(p, q)$ is a weighted window, which represents the influence of the neighbor pixels on the center pixel, and step is the variable spacing between the pixel while computing the derivative, which is typically equal to 1. The size of the sliding window is $P \times Q$. Moreover, to better measure the effect of corner pixel on the central pixel, based on the classic NSML algorithm in [34], the proposed method adds the influence of corner pixel

$$\text{ML}(h,w) = \text{ML}(h,w) + S + T \tag{7}$$

$$\begin{aligned} S = 0.707 \cdot \big[& |I(h,w) - I(h+\text{step}, w+\text{step})| \\ & + |I(h,w) - I(h-\text{step}, w-\text{step})| \big] \end{aligned} \tag{8}$$

$$\begin{aligned} T = 0.707 \cdot \big[& |I(h,w) - I(h+\text{step}, w-\text{step})| \\ & + |I(h,w) - I(h-\text{step}, w+\text{step})| \big] \end{aligned} \tag{9}$$

where S and T, respectively, represent the influence of main diagonal and secondary diagonal pixels on the central pixel.

Fig. 3 compares the gradient and the SSIM employed as loss function in [12] and [35] with the NSML introduced as clarity loss by the proposed method. The data in the figure are all from the self-built multifocus dataset CDMFI-SHU, where the first row is the weight maps and the second row is the multifocus images. The numbers below the first row represent the proportion of clear regions, descending from left to right, while the fourth column is the full blur image. The numbers below the second row show the proportion of clear regions calculated by the three algorithms mentioned above (the values are normalized to $[0, 1]$). SSIM in [35] not only contains the measurement of sharpness but also includes the structure information of the image, which will lead to its insensitivity to sharpness in the multifocus image fusion of the same scene. It can be seen from the numerical value that the value of NSML is more consistent with the actual focusing situation by comparing the gradient and the SSIM. Therefore, NSML is more suitable for multifocus image fusion than gradient to measure the loss of sharpness.

2) Discriminator Loss: The discriminator loss makes it possible to distinguish the ground truth from the generated

fake one, so it is mainly accomplished by adversarial loss

$$\begin{aligned} \boldsymbol{L}(D) &= \boldsymbol{L}_{\text{adv}}(D) \\ &= \frac{1}{N}\sum_{n=1}^{N} log(D(x_1, x_2, G(x_1, x_2)) - b) \\ &\quad + \log(D(x_1, x_2, y) - c) \end{aligned} \tag{10}$$

where b and c are the expected labels of the discriminator for the fake images and the ground truth, respectively. The introduction of labels b and c makes it possible for the discriminator to judge the authenticity of the input images.

The discriminator loss identifies the authenticity of the input images by classifying the generated fake images and the ground truth through labels b and c, leading the generator to reserve more texture features and forcing it to generate images with higher quality in the continuous game.

D. CDMFI-SHU Dataset

GAN is a data-driven model. Large and diverse data are necessary to train a result with better generalization. As far as we know, currently available public multifocus image datasets are few, which is insufficient to support the training of GAN. Now, there are three kinds of construction methods.

1) Segmenting natural images into small patches and Gaussian filtering is added to simulate the condition of multifocus [8]. This method can obtain a large amount of training data, but the size of these patches is often minimal to ignore the overall texture relationship of multifocus images.
2) Using the image segmentation or semantic segmentation algorithm to divide the natural images into the foreground and background scenes and then generate the training set by simulating defocus situation [9], [35]. However, the segmentation of this method is based on the object boundaries, which does not comply with the physical reality of optical imaging.
3) Utilizing the depth data of the RGBD images to divide foreground and background scenes and then generate the multifocus image by local blur [36]. This method can produce multifocus images conforming to the optical characteristics, but the image tends to be consistent due to the improper selection of depth threshold. At the same time, the fuzzy boundary phenomenon in natural multifocus is ignored.

Considering the shortcomings of the three methods mentioned above, this article proposes a method to generate the CDMFI-SHU multifocus image dataset: 1) instead of dividing the image into patches, the whole scene is used for training to learn the texture details of the whole, which is helpful to the generalization of the model; 2) using the NYU-D2 indoor RGBD dataset [37], the foreground and background scenes are segmented based on the dynamic threshold calculated by the gray histogram of the depth map to simulate the actual state of optical focus and make the dataset more diversified; and 3) considering the excellent performance of guided filtering in edge-preserving and smoothing [38], it is introduced to simulate the fuzzy phenomenon of the natural multifocus image at the boundaries of focus and defocus regions. The detailed generation process of CDMFI-SHU is given as follows.

1) Calculate the Dynamic Segmentation Threshold: Calculate the gray histogram of the depth map I_d, and take the maximum and minimum values $\max(I_d)$ and $\min(I_d)$ in the histogram as the total interval to calculate the dynamic segmentation intervals

$$v_i = \text{Hist}_{i-1} + \text{Hist}\left(\frac{W \cdot H}{N_s}\right) \tag{11}$$

where W and H are the width and the height of the depth map, $N_s = 4$ is the number of segmentation of the depth map, and $v = \{\min(I_d), v_1, v_2, v_3, \max(I_d)\}$ are the dynamic interval nodes. Hist represents the gray value corresponding to the number of pixels in the gray histogram. Then, the dynamic segmentation thresholds are generated randomly in each dynamic interval, respectively,

$$t_i = \text{randn}(v_{i-1}, v_i) \tag{12}$$

where randn generate random numbers within intervals.

2) Generate a Segmentation Mask: First, a preliminary segmentation mask is generated based on I_d according to the dynamic segmentation threshold t

$$M_p^t(x, y) = \begin{cases} 0, & M_p(x, y) < t_i \\ 1, & M_p(x, y) \geq t_i \end{cases} \tag{13}$$

where $M_p(x, y)$ is a pixel on the preliminary segmentation mask. There may be some singularities in the depth image due to the error of the shooting equipment, so the morphological-based open and close operation is introduced to eliminate these errors

$$M_m^t = \text{dilate}\big(\text{erode}\big(M_p\big)\big) \tag{14}$$

where dilate$(\cdot)$ and erode$(\cdot)$ represent dilation and erosion operations, respectively. After that, the guided filtering is imported to smooth and blur the boundaries between the focus and defocus regions

$$M_g^t = \text{guided}(M_m, I_{\text{focus}}, \text{randint}(2, 8)) \tag{15}$$

where guided$(\cdot)$ stands for guided filtering, I_{focus} is any real multifocus image, and randint$(2, 8)$ generates a random integer between 2 and 8.

3) Obtain the Multifocus Image Pairs: First, a full blur image is generated by Gaussian filtering

$$I_g = K_g * \mathrm{I_r} \tag{16}$$

where $*$ is the convolution operation, I_r is the RGB images, K_g is the Gaussian kernel random selected in four intervals: $\{(1.0, 1.5), (2.5, 3.0), (4.0, 4.5), (5.5, 6.0)\}$, and I_g is the images with different degrees of blur. Then, the foreground- and background-focused images can be generated, respectively,

$$I_{\text{fore}} = \begin{cases} \mathrm{I_r} \odot M_g^t + I_g \odot \big(1 - M_g^t\big), & t = t_1, t_2 \\ \mathrm{I_r} \odot (1 - M_g^t) + I_g \odot M_g^t, & t = t_3, t_4 \end{cases} \tag{17}$$

$$I_{\text{back}} = \begin{cases} \mathrm{I_r} \odot \big(1 - M_g^t\big) + I_g \odot M_g^t, & t = t_1, t_2 \\ \mathrm{I_r} \odot M_g^t + I_g \odot \big(1 - M_g^t\big), & t = t_3, t_4 \end{cases} \tag{18}$$

where $\odot$ denotes the elementwise product.

(a) (b) (c)

Fig. 4. Two sets of generated multifocus images and ground truth in CDMFI-SHU. (a) Foreground-focused. (b) Background-focused. (c) Ground truth.

4) Generate CDMFI-SHU Dataset: The generated multifocus image pairs are combined with I_r after random geometric transformation to generate a set of data of CDMFI-SHU. The multifocus image pairs are input into the generator during training, while I_r is input into the discriminator as ground truth. Fig. 4 shows two sets of multifocus images in CDMFI-SHU, where Fig. 4(a) is the foreground-focused images, Fig. 4(b) is the background-focused images, and Fig. 4(c) is the ground truth.

IV. Experiments and Results

A. Experimental Settings

1) Parameter Settings: During the training phase, the generator G and the discriminator Dof Fusion-UDCGAN have trained alternately using the Adam optimizer with $\beta = 0.5$. The batch size $B = 32$, total epochs $E = 100$, and in the first 50 epochs, the learning rate lr $= 0.2$, while, in the last 50 epochs, lr decreases linearly to 0.05. In the loss function, labels a and c range from 0.8 to 1.2, and label b ranges from 0 to 0.2. The weighted window ω to calculate NSML value in clarity loss $\boldsymbol{L}_{\text{clar}}(G)$ is

$$\omega(p,q) = \frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}. \tag{19}$$

Since the adversarial loss $\boldsymbol{L}_{\text{adv}}(G)$ shows a trend of oscillation with slight fluctuation, the value principle of α_1 and α_2 is to control the content loss $\boldsymbol{L}_{\text{cont}}(G)$ and clarity loss $\boldsymbol{L}_{\text{clar}}(G)$ in the same order of magnitude with it through weight control after a certain number of epochs so that each loss can be optimized at the same time, and the model can learn the optimal parameters. After a lot of experimentation, α_1 and α_2 of content loss $\boldsymbol{L}_{\text{cont}}(G)$ and clarity loss $\boldsymbol{L}_{\text{clar}}(G)$ are 80 and 100, respectively. Curves of $\boldsymbol{L}_{\text{adv}}(G)$ under different values of α_1 and α_2 are shown in Fig. 5. During the testing phase, the generator loads the trained parameters to generate the fused image, and the discriminator is deactivated. Finally, the implementation of this method is mainly based on the Pytorch framework of Ubuntu OS, and all experiments run on a server with Intel XEON CPU and NVIDIA TITAN Xp GPU.

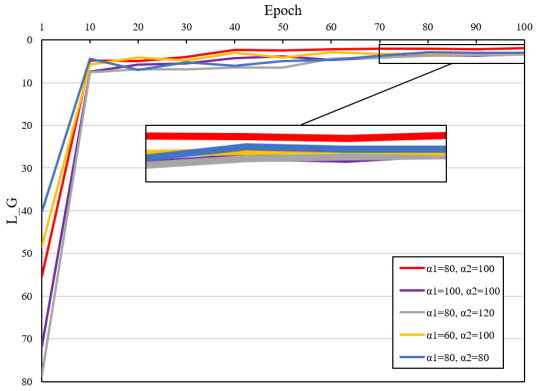

Fig. 5. Generator loss $\boldsymbol{L}_G$ curves of different values of α_1 and α_2.

Algorithm 1 Color Space Recovery Algorithm

INPUT:

background and foreground-focused RGB images: I_{back}^{RGB}, I_{fore}^{RGB}

Trained fusion networks: $UDCGAN$

OUTPUT:

Transfer RBG images to YCbCr color space: I_{back}^{YCbCr}, I_{fore}^{YCbCr}

Input the Y-component into the fusion network: $I_{fuse}^{Y} = UDCGAN(I_{fore}^{Y}, I_{back}^{Y})$

Use weighted average to fuse Cb components:

$$I_{fuse}^{Cb} = \frac{I_{fore}^{Cb}\cdot\left|I_{fore}^{Cb}-128\right| + I_{back}^{Cb}\cdot\left|I_{back}^{Cb}-128\right|}{\left|I_{fore}^{Cb}-128\right| + \left|I_{back}^{Cb}-128\right|}$$

Use weighted average to fuse Cr components:

$$I_{fuse}^{Cr} = \frac{I_{fore}^{Cr}\cdot\left|I_{fore}^{Cr}-128\right| + I_{back}^{Cr}\cdot\left|I_{back}^{Cr}-128\right|}{\left|I_{fore}^{Cr}-128\right| + \left|I_{back}^{Cr}-128\right|}$$

Convert the YCbCr image to RGB color space: I_{fuse}^{RGB}

Return: I_{fuse}^{RGB}

2) Datasets: The proposed method employs the self-built CDMFI-SHU multifocus image dataset for training. The original CDMFI-SHU contains 23 184 pairs of multifocus images and ground truth, which is increased to 30 000 by random geometric transformation during training. It is worth noting that, in order to enable the model to focus on the extraction and fusion of clear areas, the input images during training are transformed into single-channel gray-scale images. Meanwhile, to verify the generalization ability of the proposed method, the trained model is tested on the Grayscale [39] and Lytro [40] datasets. Besides, all images are preprocessed, made gray, and scaled to 512×512 before being input into the network during the testing phase. Since the output fusion image is single-channel, Algorithm 1 is set for color space restoration of experiment results on the Lytro dataset.

3) Evaluation Indexes: This article selects the six most commonly used indexes for evaluation.

SF: It reflects the change rate of the gray scale by calculating the difference between the central pixel and the neighbor

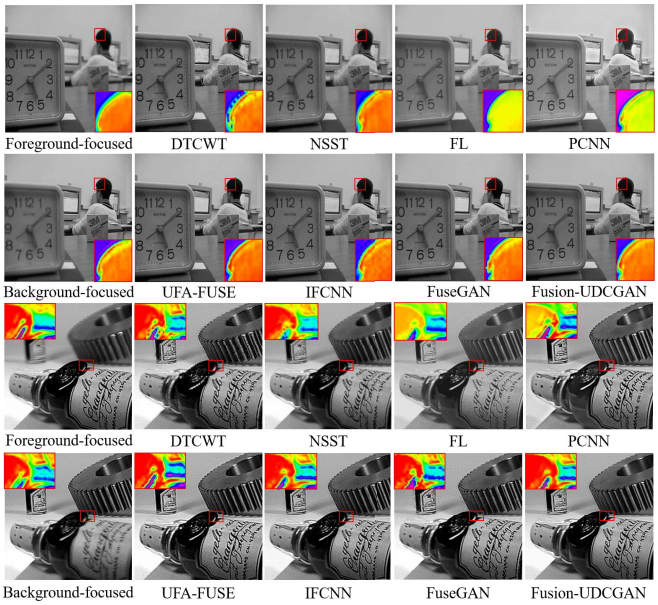

Fig. 6. Experimental results of the Grayscale dataset and pseudocolor graphs of the enlarged regions of DTCWT [42], NSST [43], FL [44], PCNN [45], UFA-FUSE [35], IFCNN [36], FuseGAN [9], and Fusion-UDCGAN.

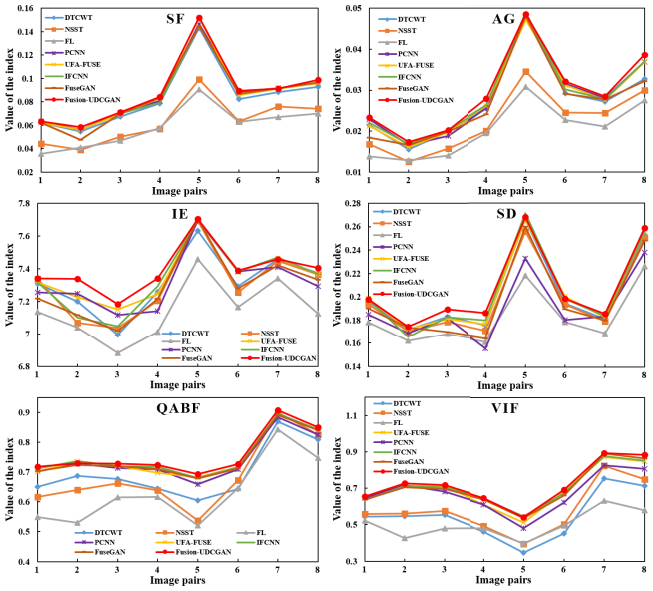

Fig. 7. Experimental results of six objective indexes of ten sets of image pairs on the Grayscale dataset.

pixel, which can be introduced to measure the image clarity. A higher value represents a sharper image.

AG: It reflects the detail contrast and the clarity of texture features by the mean value of the horizontal and vertical gradients of the image. In general, a larger value of AG usually means that the image has a richer hierarchy and higher resolution.

IE: It measures the relationship between the fused image and the source images but focuses more on evaluating the pixel distribution of the former. If the SD value is large, image contrast quality is high.

SD: It measures the richness of image information, which describes the discrete gray-scale value instead of the mean gray value of the image. If the SD value is large, the quality of image contrast is high.

$Q^{AB/F}$*:* It measures the quality of the fused image by calculating the retention amount of the edge information of the source image in the fused image, usually ranging from 0 to 1. The higher the value of $Q^{AB/F}$, the sharper the edge and the higher the fusion quality.

VIF: It simulates the principle of the human visual system based on the fidelity of visual information to evaluate fusion quality [41]. A higher value means that the image is more consistent with the human visual system and has higher quality.

B. Experiment on Grayscale Dataset

To verify the generalization performance of Fusion-UDCGAN, the trained model is tested on the Grayscale dataset. Two groups of representative experimental results of proposed and compared methods are presented in Fig. 6 (specific regions in them are circled by the red boxes and enlarged and displayed in the form of a pseudocolor graph to show the performance of the methods more clearly).

Fig. 6 shows the subjective performance of different methods in the case of good and poor registration. Due to the introduction of window function filtering, DTCWT shows noticeable block effects. NSST can complete the fusion without block effect, but a certain contrast loss is at the focus–defocus regions' boundaries. FL and PCNN can preserve the information of the boundaries well, but the overall contrast of the fused images is greatly reduced. This may be caused by the weighted average strategy of low-frequency subbands, resulting in energy loss. UFA-FUSE has good fusion quality, but it produces artifacts at the edges when the images are not correctly registered. This may be related to the global rules of activity level measurement. In the top set of Fig. 6, IFCNN shows almost the same fusion performance as Fusion-UDCGAN. However, from the pseudocolor graphs in the bottom set, IFCNN shows wrong texture features, indicating that its fusion performance is not stable enough. Because FuseGAN is based on image segmentation, it has high fusion quality far from the boundaries of focus–defocus regions. However, affected by image registration failure, information errors may occur in the fused images at the boundaries, such as the edge around the student's hair in Fig. 6.

In terms of subjective performance, Fusion-UDCGAN has higher clarity, contrast, richer texture structure, stronger robustness, no block effects, and fewer artifacts when the image pairs are not correctly registered compared with other methods. In conclusion, the proposed method can obtain fused images with higher performance in subjective evaluation.

Considering that subjective evaluation may have some deviation due to different subjects, the objective evaluation results of eight image pairs and the average scores of all results in the Grayscale dataset are shown in Fig. 7 and Table II, respectively. Table II also shows the percentage improvement for the proposed and other best methods. The proposed method gets the highest score in almost all the indexes, which means that the proposed Fusion-UDCGAN is superior to other

TABLE II
AVERAGE SCORES OF SIX OBJECTIVE INDEXES ON THE GRAYSCALE DATASET

	DTCWT	NSST	FL	PCNN	UFA-FUSE	IFCNN	FuseGAN	UDFuse-GAN
SF	0.0837	0.0629	0.0590	0.0868	0.0869	0.0849	0.0854	**0.0886** (1.96%)
AG	0.0274	0.0223	0.0203	0.0285	0.0283	0.0287	0.0270	**0.0296** (3.25%)
IE	7.3146	7.3019	7.1448	7.3178	7.3569	7.3350	7.2879	**7.3946** (0.51%)
SD	0.2025	0.1988	0.1804	0.1905	0.2039	0.2042	0.1967	**0.2072** (1.46%)
$Q^{AB/F}$	0.6982	0.6865	0.6335	0.7413	0.7492	0.7526	0.7471	**0.7593** (0.89%)
VIF	0.5475	0.5826	0.5019	0.6748	0.7007	0.7064	0.7042	**0.7206** (2.00%)

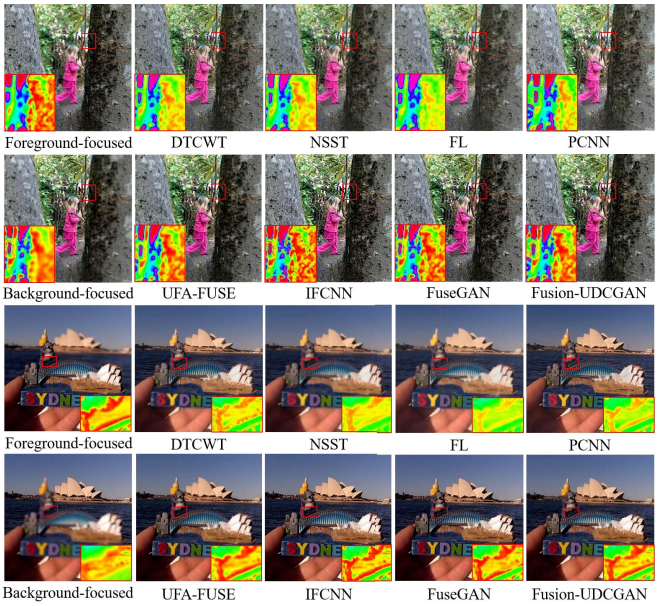

Fig. 8. Experimental results of the Lytro dataset and pseudocolor graphs of the enlarged regions of DTCWT [42], NSST [43], FL [44], PCNN [45], UFA-FUSE [35], IFCNN [36], FuseGAN [9], and Fusion-UDCGAN.

comparison methods in terms of clarity, information richness, contrast, edge retention, and visual fidelity and generates higher quality fused images.

C. Experiment on Lytro Dataset

Fusion-UDCGAN is tested on the Lytro dataset, and two sets of results in Fig. 8 are selected to prove the superiority of the proposed method in subjective evaluation. The Lytro dataset consists of RGB three-channel images, so, to retain the chroma information, the RGB images are converted to the YCbCr color space during training and testing. This is because the Y channel can represent the texture and energy information of the images, which are the main goals of multifocus image fusion. The Y image channel is input into the generator to get the fused Y channel image, while the Cb and Cr channels are fused using the weighted average method, and the fused images of the three channels are combined and converted to the RGB color space. The detailed three-channel color recovery algorithm is shown in Algorithm 1.

As can be seen from the results, all the methods can complete image fusion on the Lytro dataset, among which Fusion-UDCGAN fusion quality is the best. Due to the introduction of the filtering window, the fusion result of DTCWT has apparent block effects, which affects the image quality. NSST uses a weighted average in the fusion of low-frequency subbands, resulting in the loss of image contrast. FL directly fuses the pixels of the images through the fuzzy logic system, resulting in a decrease in the sharpness. Although PCNN has high fusion quality far from the boundaries of focus–defocus regions, a certain degree of artifact appears at the boundaries, which may be because the blurring phenomenon of natural focusing boundary is not taken into account when training the network. UFA-FUSE uses max–min and average strategies to fuse the source images, but the contrast and the chroma of fused images suffer a loss. IFCNN can process the fusion at the boundaries well, but the sharpening at some high-frequency features of the fused images is different from that of the source images. FuseGAN uses GAN to generate the focusing region mask, which may cause classification errors (the boundaries of tree trunks and leaves, the sea, and toys in the red boxes in Fig. 8 at the focus–defocus boundaries).

Compared with other methods, Fusion-UDCGAN has a stronger contrast, fuller chroma information, higher resolution, and complete retention of details in the overall look of fused images. At the same time, more importantly, the proposed method can better preserve the texture at the boundaries of focus–defocus regions, minimize the information loss at these places, and obtain better fusion results.

As shown in Fig. 9, to further evaluate the performance of the proposed method on objective indicators, ten pairs of multifocus images are randomly selected from the Lytro dataset to calculate the performance of the six indexes. The results show that the proposed method achieves the highest scores in almost every index, which means that Fusion-UDCGAN has higher contrast, richer hierarchy levels, more feature information, higher clarity, better visual fidelity, fewer artifacts, and no block effect. Table III calculates the average scores of the six indexes of all methods and the percentage improvement compared with the other best method in the Lytro dataset, and the proposed method achieves the highest score, which further proves the above conclusion. It is worth noting that, although FuseGAN is close to the proposed method in terms of almost indexes, Fusion-UDCGAN can obtain higher quality fusion results on the whole because the former may cause classification errors at the boundaries of focus–defocus regions.

Fusion-UDCGAN can also perform well in sequence multifocus image fusion with common defocus regions. Sequential images in the Lytro dataset are fused, and two groups of the

TABLE III
AVERAGE SCORES OF SIX OBJECTIVE INDEXES ON THE LYTRO DATASET

INDEXES	DTCWT	NSST	LD	PCNN	UFA-FUSE	IFCNN	FuseGAN	UDFuse-GAN
SF	0.0640	0.0452	0.0452	0.0648	0.0649	0.0632	0.0632	**0.0666** (2.56%)
AG	0.0243	0.0166	0.0166	0.0241	0.0239	0.0240	0.0242	**0.0249** (3.08%)
IE	7.3656	7.2987	7.3020	7.3679	7.4297	7.3657	7.4502	**7.5350** (1.14%)
SD	0.1838	0.1779	0.1791	0.1890	0.1933	0.1873	0.1959	**0.2004** (2.32%)
$Q^{AB/F}$	0.5036	0.5581	0.5112	0.6862	0.6916	0.6930	0.6864	**0.6983** (0.76%)
VIF	0.5749	0.5984	0.5970	0.6911	0.6940	0.6967	0.6695	**0.7060** (1.33%)

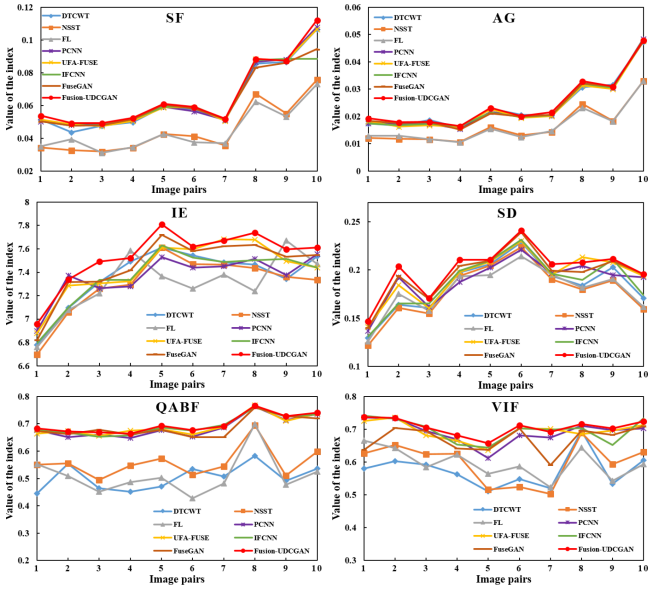

Fig. 9. Experimental results of six objective indexes of ten sets of image pairs on the Lytro dataset.

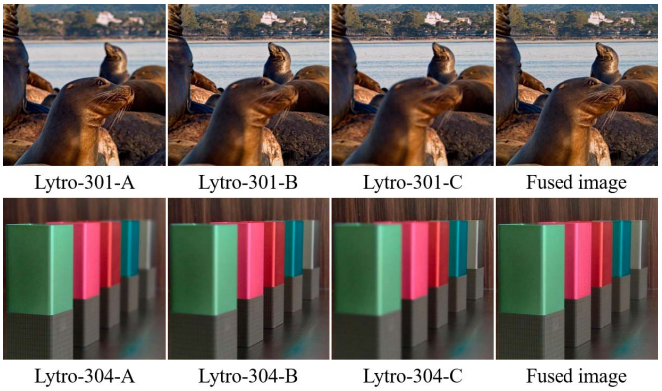

Fig. 10. Two sets of experimental results of sequential images fusion on the Lytro dataset.

results are shown in Fig. 10. As can be seen from the figure, the fused image generated by Fusion-UDCGAN fuses the clear features of different focusing regions of sequence images without interference from the overlapping fuzzy regions. The fusion of sequential images is similar to that of image pairs. Specifically, two images are randomly selected to be input into the network for fusion, and then, the fused image and the third image are input into the network to generate the final fused image. Experimental results show that the order of input images does not affect the quality of fusion.

D. Ablation Experiments

In this section, ablation experiments are set up to prove the effectiveness of each module in the proposed method. The experimental control groups are set as follows.

No Dense Long Connections: The dense long connections between the low-level features of the encoder and the high-level features of the decoder are removed as a control group.

No Dense Blocks: The dense blocks in the downsampling blocks of the encoder and the upsampling blocks of the decoder are removed as a control group.

General Training Set: The training set is constructed using a method similar to [6] as a control group.

L_{adv}*:* It only uses adversarial loss as the overall optimization objective of the network.

$L_{adv} + L_{cont}$*:* It introduces adversarial loss and content loss as overall network optimization goals.

$L_{adv} + L_{clar}$*:* It is similar to the previous group but replaced content loss with clarity loss.

The ablation experiments are performed on ten pairs of multifocus images, and Fig. 11 shows one group of the results. It is worth mentioning that the experimental image pairs are selected from the Grayscale and Lytro datasets. As can be seen from the figure, the group of no dense long connections, L_{adv}, and $L_{\text{adv}} + L_{\text{clar}}$ all appeared with different degrees of block effect, which reduced the image quality. This indicates that dense long connections, content loss, and clarity loss can effectively reduce the loss of information and, thus, promote the quality improvement of fusion images. Compared with the proposed methods, local blurring and contrast reduction occurred in both groups of no dense blocks and $L_{\text{adv}} + L_{\text{cont}}$, indicating that dense blocks and clarity loss can promote sharp images. In the general training set group, significant feature loss occurred at the focus–defocus boundaries of the image, possibly due to the simulation of the natural focusing situation in the dataset. This shows that the proposed method is beneficial to eliminate the loss at the boundaries and optimize the fusion image.

A more accurate qualitative comparison is made for each control group to eliminate subjective errors. The experimental results of control groups and the proposed method on six indexes are compared, as shown in Fig. 12. As the figure shows, Fusion-UDCGAN achieves the best performance in all indexes on every image pair. The method's performance decreases after removing some modules, indicating that all modules positively affect the proposed method. In general,

TABLE IV
AVERAGE TIME COMPLEXITY OF EACH METHOD

Method	DTCWT	NSST	FL	PCNN	UFA-FUSE	IFCNN	FuseGAN	Fusion-UDCGAN
Time (s)	3.1320	8.2367	7.6381	4.3697	**0.4976**	0.5467	0.6325	0.7025

Fig. 11. Group of results from the ablation experiments.

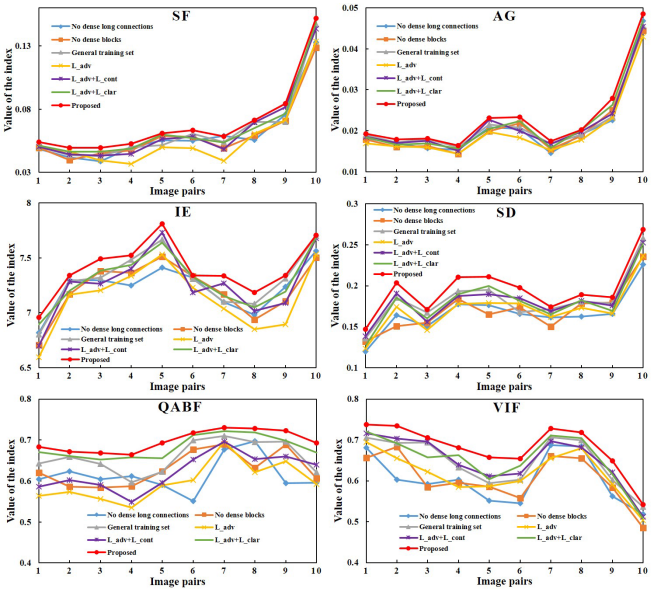

Fig. 12. Results of the ablation experiments on six objective evaluation indexes for ten pairs of images.

the introduction of dense long connections and clarity loss contributes the most to improving method performance, while the improvement of the dataset has the most negligible impact. The ablation experiment results show that each module proposed in Fusion-UDCGAN has a beneficial effect on the model, helping the method to generate high-quality fused images with higher definition, higher contrast, richer detail texture, and more consistent with human visual characteristics.

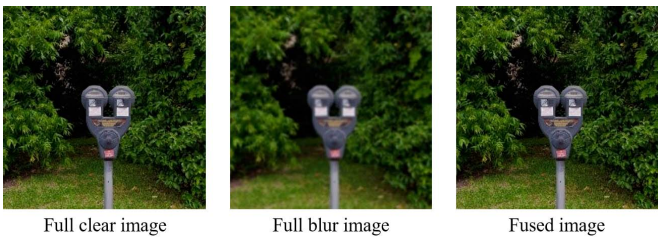

Fig. 13. Fusion experiments of full clear and full blur images.

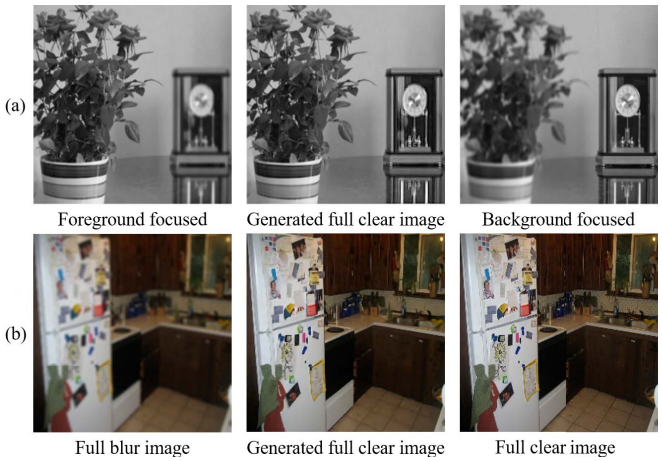

Fig. 14. Image deblurring experiment using single-branch Fusion-UDCGAN network. (a) From left to right: foreground-focused image, generated full clear image, and background-focused image for reference. (b) From left to right: full blur image, generated deblurred image, and full clear image for reference.

E. Computational Efficiency

To evaluate the computational efficiency, we calculate the average algorithm running time of different methods on the Grayscale and Lytro datasets in Table IV. Deep learning-based methods are more efficient than traditional methods, and the time complexity of the proposed method is equal to the other three deep learning-based methods.

V. DISCUSSION

For discussion, this article tries to fuse full blur and full clear images on the trained model, and the fusion results are observed in Fig. 13. The figure shows that Fusion-UDCGAN trained with CDMFI-SHU is not affected by the size of the fuzzy regions of the source images and has a better stability.

Meanwhile, Fusion-UDCGAN has fairly good model extensibility. To prove this feature, this article tries to input partial blur or full blur images into the single-branch network to

generate full clear images. In particular, this experiment turned the encoder of Fusion-UDCGAN into a single branch and then fed half of the multifocus dataset into this branch to train the network to generate full clear images. A partial blur or full blur image is input into the network to obtain a full clear image in the test phase. The experimental results are shown in Fig. 14. From the results, in the generated full clear images, the clear regions of the input image are completely retained, while the blur regions become clearer, which indicates that the proposed method can identify and make the blur regions clearer without changing the clear regions and improve the overall quality of the source images. It is worth mentioning that, different from the multifocus image fusion experiment above, the images in the third column in Fig. 14 are not input into the network during the training and testing phases.

VI. Conclusion

In this article, a novel generative adversarial network with dense Z-type connection and U-type structure called Fusion-UDCGAN is proposed for multifocus image fusion. The encoder and the decoder of the method are designed to extract and fuse clear regions of multifocus images based on downsampling blocks and upsampling blocks, which are inspired by dense connections. Simultaneously, to combine low- and high-level features, dense long connections are established between them, and dense feature reuse is utilized to reduce information loss. Based on LSGAN, the content loss and clarity loss using the $L1$ norm and NSML are designed to extract and fuse the clear regions of multifocus images and further enhance fused images' texture details. In addition, due to the lack of a large-scale multifocus image fusion dataset, the CDMFI-SHU dataset based on histogram adaptive threshold and guided filtering is also made for model training. Experimental results of subjective and objective evaluation on two commonly employed datasets indicate that Fusion-UDCGAN generated multifocus fused images have higher overall clarity, contrast, and chroma and fully express the detailed texture features of the source image, especially at the boundaries of the focus–defocus regions.

To verify the performance of Fusion-UDCGAN in various situations, some additional experiments are added. These experimental results show that the fusion results of multifocus images are not affected by blur conditions' differences and have better robustness in all cases. Moreover, at the end of this article, we try to use the method of a single-branch network for image deblurring and also achieve a satisfactory effect, which suggests that Fusion-UDCGAN may in image denoising and super-resolution imaging fields to make some beneficial attempts; this will be the focus of our future work.

References

[1] F. Zhou, X. Li, J. Li, R. Wang, and H. Tan, "Multifocus image fusion based on fast guided filter and focus pixels detection," *IEEE Access*, vol. 7, pp. 50780–50796, 2019.

[2] M. A. Rahman, S. Liu, C. Y. Wong, S. C. F. Lin, S. C. Liu, and N. M. Kwok, "Multi-focal image fusion using degree of focus and fuzzy logic," *Digit. Signal Process.*, vol. 60, pp. 1–19, Jan. 2017.

[3] Y. Yang, Y. Que, S.-Y. Huang, and P. Lin, "Technique for multi-focus image fusion based on fuzzy-adaptive pulse-coupled neural network," *Signal, Image Video Process.*, vol. 11, no. 3, pp. 439–446, 2017.

[4] Y. Yang, S. Tong, S. Huang, and P. Lin, "Multifocus image fusion based on NSCT and focused area detection," *IEEE Sensors J.*, vol. 15, no. 5, pp. 2824–2838, May 2015.

[5] A. Vishwakarma and M. K. Bhuyan, "Image fusion using adjustable non-subsampled shearlet transform," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 9, pp. 3367–3378, Sep. 2018.

[6] Y. Yang, M. Yang, S. Huang, M. Ding, and J. Sun, "Robust sparse representation combined with adaptive PCNN for multifocus image fusion," *IEEE Access*, vol. 6, pp. 20138–20151, 2018.

[7] C. Zhang, "Multifocus image fusion using multiscale transform and convolutional sparse representation," *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 19, no. 1, Jan. 2021, Art. no. 2050061.

[8] Y. Liu, X. Chen, H. Peng, and Z. F. Wang, "Multi-focus image fusion with a deep convolutional neural network," *Inf. Fusion*, vol. 36 pp. 191–207, Jul. 2017.

[9] X. Guo, R. Nie, J. Cao, D. Zhou, L. Mei, and K. He, "FuseGAN: Learning to fuse multi-focus image via conditional generative adversarial network," *IEEE Trans. Multimedia*, vol. 21, no. 8, pp. 1982–1996, Aug. 2019.

[10] H. Li, X.-J. Wu, and T. Durrani, "NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9645–9656, Dec. 2020.

[11] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2019.

[12] H. Zhang, Z. Le, Z. Shao, H. Xu, and J. Ma, "MFF-GAN: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion," *Inf. Fusion*, vol. 66, pp. 40–53, Feb. 2021.

[13] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4980–4995, 2020.

[14] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, Aug. 2019.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[16] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[17] X. Jin *et al.*, "Brain medical image fusion using L2-norm-based features and fuzzy-weighted measurements in 2-D Littlewood–Paley EWT domain," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 8, pp. 5900–5913, Aug. 2020.

[18] H. Xu, J. Ma, and X.-P. Zhang, "MEF-GAN: Multi-exposure image fusion via generative adversarial networks," *IEEE Trans. Image Process.*, vol. 29, pp. 7203–7216, 2020.

[19] C. Wang *et al.*, "DiCyc: GAN-based deformation invariant cross-domain information fusion for medical image synthesis," *Inf. Fusion*, vol. 67, pp. 147–160, Mar. 2021.

[20] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.

[21] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2794–2802.

[22] D. Peng, Y. Zhang, and H. Guan, "End-to-end change detection for high resolution satellite images using improved UNet++," *Remote Sens.*, vol. 11, no. 11, p. 1382, Jun. 2019.

[23] K. Zhao, J. Cheng, T. Liu, and H. Deng, "A generative adversarial network for fusion of infrared and visible images based on UNet++," *Proc. SPIE*, vol. 11584, Nov. 2020, Art. no. 1158405.

[24] W. Zhang, J. Li, and Z. Hua, "Attention-based tri-UNet for remote sensing image pan-sharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3719–3732, 2021.

[25] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.

[28] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[29] W. Guo, Z. Xu, and H. Zhang, "Interstitial lung disease classification using improved DenseNet," *Multimedia Tools Appl.*, vol. 78, no. 21, pp. 30615–30626, Nov. 2019.

[30] T. Li, W. Jiao, L.-N. Wang, and G. Zhong, "Automatic DenseNet sparsification," *IEEE Access*, vol. 8, pp. 62561–62571, 2020.

[31] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani, and G. Jamalipour Soufi, "Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101794.

[32] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[33] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1125–1134.

[34] H. Ullah, B. Ullah, L. Wu, F. Y. O. Abdalla, G. Ren, and Y. Zhao, "Multi-modality medical images fusion based on local-features fuzzy sets and novel sum-modified-Laplacian in non-subsampled shearlet transform domain," *Biomed. Signal Process. Control*, vol. 57, Mar. 2020, Art. no. 101724.

[35] Y. Zang, D. Zhou, C. Wang, R. Nie, and Y. Guo, "UFA-FUSE: A novel deep supervised and hybrid model for multifocus image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–17, 2021.

[36] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "IFCNN: A general image fusion framework based on convolutional neural network," *Inf. Fusion*, vol. 54, pp. 99–118, Feb. 2020.

[37] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 746–760.

[38] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397–1409, Jun. 2013.

[39] M. Nejati, S. Samavi, and S. Shirani, "Multi-focus image fusion using dictionary-based sparse representation," *Inf. Fusion*, vol. 25, pp. 72–84, Sep. 2015.

[40] J. Saeedi and K. Faez, "A classification and fuzzy-based approach for digital multi-focus image fusion," *Pattern Anal. Appl.*, vol. 16, no. 3, pp. 365–379, 2013.

[41] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[42] Y. Tian, J. Luo, W. Zhang, T. Jia, A. Wang, and L. Li, "Multifocus image fusion in Q-shift DTCWT domain using various fusion rules," *Math. Problems Eng.*, vol. 2016, pp. 1–12, Sep. 2016.

[43] S. Liu, J. Wang, Y. Lu, S. Hu, X. Ma, and Y. Wu, "Multi-focus image fusion based on residual network in non-subsampled shearlet domain," *IEEE Access*, vol. 7, pp. 152043–152063, 2019.

[44] B. Wang, J. Zeng, S. Lin, and G. Bai, "Multi-band images synchronous fusion based on NSST and fuzzy logical inference," *Infr. Phys. Technol.*, vol. 98, pp. 94–107, May 2019.

[45] Z. Wang, S. Wang, and L. Guo, "Novel multi-focus image fusion based on PCNN and random walks," *Neural Comput. Appl.*, vol. 29, no. 11, pp. 1101–1114, Jun. 2018.

Yuan Gao was born in Zhengzhou, Henan, China, in 1993. He received the M.A.Sc. degree from the College of Computer Science and Technology, Shanghai University of Electric Power, Shanghai, China, in 2018. He is currently pursuing the Ph.D. degree with the School of Mechatronic Engineering and Automation, Shanghai University, Shanghai.

His current research interests include image processing, machine vision, and pattern recognition.

Shiwei Ma received the bachelor's and master's degrees in electronics from Lanzhou University, Lanzhou, in 1986 and 1991, respectively, and the Ph.D. degree in control science and engineering from Shanghai University, Shanghai, China, in 2000.

He was engaged in post-doctoral research at the National Institute of Industrial Safety, Tokyo, Japan, as a JST Fellow, from February 2001 to February 2003. He is currently a Professor with the School of Mechatronics Engineering and Automation, Shanghai University. His current research fields include image processing, machine learning, and pattern recognition.

Jingjing Liu received the bachelor's degree in electrical engineering from Zhengzhou University, Zhengzhou, China, in 2002, and the M.Sc. and Ph.D. degrees in control science and engineering from Shanghai University, Shanghai, China, in 2013 and 2018, respectively.

She was a Joint Ph.D. Student with the Computation Department, Curtin University, Bentley, WA, Australia, from December 2015 to December 2016. She currently holds a post-doctoral position at the School of Electronic Science and Technology, Fudan University, Shanghai, with interests in pattern recognition, computer vision, machine learning, and image processing.

Xianchao Xiu (Member, IEEE) received the Ph.D. degree in operations research from Beijing Jiaotong University, Beijing, China, in 2019.

From June 2019 to May 2021, he was a Post-Doctoral Researcher with Peking University, Beijing. He is currently a Faculty Member with the School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China. His current research interests include large-scale sparse optimization, signal processing, deep learning, and data-driven fault detection.