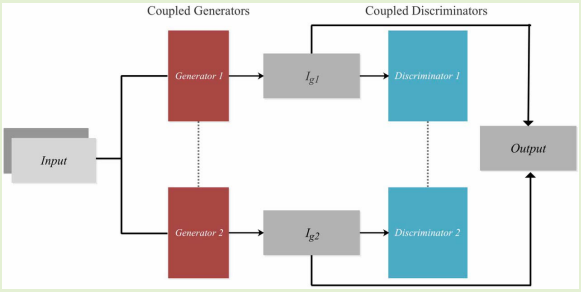
Coupled GAN With Relativistic Discriminators for Infrared and Visible Images Fusion

Qilei Li, *Student Member, IEEE*, Lu Lu, *Member, IEEE*, Zhen Li, *Student Member, IEEE*, Wei Wu, Zheng Liu, *Senior Member, IEEE*, Gwanggil Jeon, *Member, IEEE*, and Xiaomin Yang, *Member, IEEE*

***Abstract*—Infrared and visible images are a pair of multi-source multi-sensors images. However, the infrared images lack structural details and visible images are impressionable to the imaging environment. To fully utilize the meaningful information of the infrared and visible images, a practical fusion method, termed as RCGAN, is proposed in this paper. In RCGAN, we introduce a pioneering use of the coupled generative adversarial network to the field of image fusion. Moreover, the simple yet efficient relativistic discriminator is applied to our network. By doing so, the network converges faster. More importantly, different from the previous works in which the label for generator is either infrared image or visible image, we innovatively put forward a strategy to use a pre-fused image as the label. This is a technical innovation, which makes the process of generating fused images no longer out of thin air, but from "existence" to "excellent." The extensive experiments demonstrate the proposed RCGAN can produce a faithful fused image, which can efficiently persevere the rich texture from visible images and thermal radiation information from infrared images. Compared with traditional methods, it successfully avoids the complex manual designed fusion rules, and also shows a clear advantages over other deep learning-based fusion methods.**



***Index Terms*—Image fusion, infrared image, visible image, coupled generative adversarial network, relativistic discriminator, deep learning.**

I. Introduction

MANY cameras are equipped with both visible imaging sensor and infrared imaging sensor. These two types of sensors can capture the visible image and the infrared image respectively. Visible images contain rich texture information, while visible imaging sensors are susceptible to the environment. For instance, some important objects may be invisible in visible image the condition of darkness or thick fog. In contrast, infrared images are captured in accordance with the thermal radiation. Thus, they can work stably around the clock under all conditions. However, infrared images often lack texture information. To fully utilize the complementary information, infrared and visible images fusion technology [1] aims to integrate infrared image and visible image into a single image, which is rich in both texture and thermal radiation distribution.

Generally, existing infrared and visible images fusion methods can be roughly divided into two categories: traditional fusion methods and deep learning based methods. Traditional fusion methods process the source images either in the spatial domain or the transform domain. For spatial domain methods, the fused is obtained via analyzing the spatial relationship between infrared image and visible image. For instance, Li *et al.* [2] decomposed source images into two scales and calculated the weight maps for the two scales by using the guided filter [3]. Consequently, the fused image can be obtained by reconstructing the two scales via the weight maps. This is a representative spatial fusion method, which motivated other fusion methods [4], [5]. For the transform domain methods, the source images are usually transformed into coefficients by

Manuscript received January 27, 2019; revised May 21, 2019; accepted June 5, 2019. Date of publication June 10, 2019; date of current version February 17, 2021. This work was supported in part by the National Science Foundation of P.R. China under Grant 61711540303, Grant 61701327, and Grant 61601266, in part by the Science Foundation of Sichuan Science and Technology Department under Grant 2018GZ0178, and in part by the China Post-Doctoral Science Foundation Funded Project under Grant 2018M640916. The associate editor coordinating the review of this article and approving it for publication was Prof. Lyudmila Mihaylova. *(Corresponding authors: Lu Lu; Xiaomin Yang.)*

Q. Li, L. Lu, Z. Li, W. Wu, and X. Yang are with the College of Electronics and Information Engineering, Sichuan University, Chengdu 610064, China (e-mail: qilei.li@outlook.com; lulu19900303@126.com; zhenli0424@foxmail.com; weiwu@scu.edu.cn; arielyang@scu.edu.cn).

Z. Liu is with the Faculty of Applied Science, University of British Columbia, Kelowna, BC V1V 1V7, Canada (e-mail: zheng.liu@ieee.org).

G. Jeon is with the School of Electronic Engineering, Xidian University, Xi'an 710071, China, and also with the Department of Embedded Systems Engineering, Incheon National University, Incheon 22012, South Korea (e-mail: ggjeon@gmail.com).

Digital Object Identifier 10.1109/JSEN.2019.2921803

some mathematical tools. The next is to measure the activity level and to fuse the transform coefficients. Finally, the fused image is obtained via an inverse transformation. In this aspect, some typical transform methods, such as Laplacian pyramid (LP), discrete wavelet transform (DWT), nonsubsampled contourlet transform (NSCT), sparse representation (SR) were explored in [6]–[13]. However, the aforementioned traditional fusion methods may take a long time for decomposition. Besides, these traditional methods require complex manually designed fusion rules. Hence, these methods are hard to use in practical applications.

Recently, some deep learning based fusion methods have been proposed by virtue of the powerful representation ability of deep networks [14]. Liu *et al.* [15] combined Laplacian pyramid decomposition and shallow convolutional neural network to fused the infrared and visible images. Li *et al.* [16] decomposed the source images into a base part and a detail part, and used VGG-network [17] to fuse the detail layer. Nevertheless, the approaches above still require two complex handcrafted components as most traditional fusion methods do: activity level measurements and fusion rules. As a result, such methods occupy excessive computing and storage resources. To achieve an end-to-end mapping from source images to a fused image, Ma et al. introduced a generative adversarial network (GAN) [18] to infrared and visible images fusion task [19], termed as FusionGAN. In FusionGAN, a fused image with highlighted targets and abundant textures can be directly generated by the generator. The discriminator, which regards a corresponding visible image as a positive sample, can provide more texture information for the fused image.

Although FusionGAN is independent of complicated activity level measurements and fusion rules, we argue that it suffers from three drawbacks. First, single GAN only can exploit one relationship between different semantic levels, but highlighted targets in infrared images and abundant textures in visible images obviously should be treated in different ways. Second, there is no valid label in FusionGAN, making the process of generating fused images undirected. Third, the criterion used in the discriminator of FusionGAN determines whether a fused image is a visible image or not, but the absolute differences between a fused image and a visible image are difficult to fool the discriminator. Therefore, the training process of FusionGAN will wander off in the wrong direction.

To solve the issues mentioned above, a relativistic coupled GAN for infrared and visible images fusion, which is abbreviated as 'RCGAN' is proposed. We use coupled GAN [20] to play a two-team game (each team contains one generator and one discriminator) rather than a minimax two-player game in FusionGAN. As for generators, infrared and visible images share the same high-level concepts at the first layers, and utilize different low-level details to fuse meaningful information at the last layers. A pair of fused images, which are obtained by generators, are then fed into corresponding discriminators to distinguish the high-level representations among them. To solve the 'undirected' problem existed in the previous works [15], [16], [19], a pre-fused image is employed as the guide of generators. We employ relativistic discriminators [21] to measure the relative differences between the fused image and infrared/visible image. Compared with traditional fusion methods, our method successfully avoids complex manual designed fusion rules and can fuse source images in an end-to-end way. Meanwhile, the proposed RCGAN requires less computational computing and storage resources compared with [15], [16], it also overcomes the aforementioned inherent disadvantages in [19]. Compared with FuisonGAN, our method employs coupled GAN to handle multi-domain images in a different way. By using the pre-fused image, our network is able to purposefully synthesis a faithful fusion results. To summarize, our main contributions can be listed as follows:

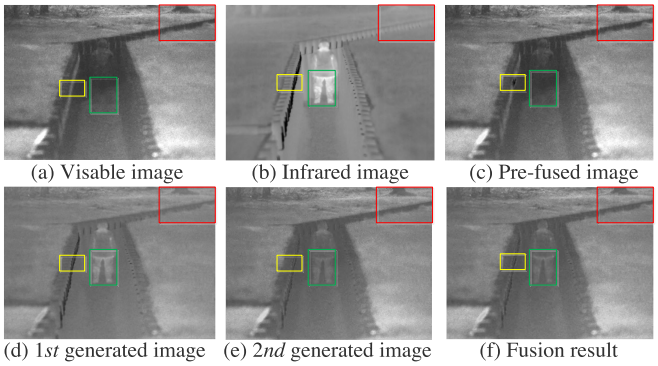


Fig. 1. Infrared and visible images fusion using RCGAN. (c) is the pre-fused image by GFF [2]. (d) and (e) are the generated results of the coupled generators, (f) is the final fusion result. The pre-fused image neglects the significant object (see the green box), and it introduces a noisy block (see the yellow box). The first generator optimizes the pre-fused image towards the infrared image, thus the person is visible, but the tree (see the red box) is fuzzy. The second generator optimize the pre-fused image towards the visible image. It addresses the noisy block in the pre-fused image, but the person is fuzzy. Thus, the (d) and (e) are averaged to take advantage of them, and the final fusion result (f) can be obtained. (f) is more clear and representative than the pre-fused image.

1) We pioneer the coupled GAN for multi-domain image fusion. A couple of generators and discriminators, which share common scenes from source images, are used to treat highlighted targets in infrared images and abundant textures in visible images in different ways.
2) We creatively use the pre-fused image as the guide for the coupled generators in the training phase.[1] By doing so, the objective of generators is to optimize the pre-fused image, rather than generate it out of thin air. In other words, it enables the fused image from 'existence' to 'excellent'.
3) We introduce the relativistic discriminators to evaluate the relative differences between the fused image and the infrared/visible image. Through using relativistic discriminators, the convergence process can be more stable during training and the fusion result can be more faithful.

The rest of the paper is organized as follows. Section II introduces some related works, such as the FusionGAN and coupled GAN. Section III describes the proposed RCGAN in detail. The experimental results and analyses are given in Section IV. Finally, the conclusion is drawn in Section V.

[1]It should be noted that the pre-fused image is not required during inference.

II. Related Works

A. FusionGAN

FusionGAN [19] first introduced the generative adversarial network (GAN) to image fusion task. In FusionGAN, the generator receives the concatenated infrared and visible images to generate an image with both infrared image thermal radiation information and visible image gradient detail. The discriminator will inject more visible image structure information by distinguishing the generated image and visible image.

The loss function for the generator is designed as

$$L_G = \Phi(G) + \alpha L_{content}, \tag{1}$$

where L_G represents the total loss for the generator, $\Phi(G)$ denotes the adversarial loss, $L_{content}$ represents the content loss between the generated image I_g and the infrared/visible image (I_{ir}/I_{vis}), and α is a constant to regulate the ratio between $\Phi(G)$ and $L_{content}$. In more detail, the two separate loss functions are formulated as

$$\begin{aligned} \Phi(G) &= \frac{1}{N}\sum_{n=1}^{N}\left(D(I_g^{(n)}) - a\right)^2 \\ L_{content} &= \frac{1}{WH}(\beta \left\|\nabla I_g - \nabla I_{vis}\right\|_2^2 + \left\|I_g - I_{ir}\right\|_2^2), \end{aligned} \tag{2}$$

where $I_g^{(n)}$ is the n-th generated image. There are N images totally. a is the label that discriminator thinks the generated data is real, W and H denote the width and height of the source images, respectively. The notation ∇ denotes the gradient operator, $\|\cdot\|_2$ denotes the L_2-norm, and β is a constant to control the ratio of gradient information.

The loss function for the discriminator is set to

$$L_D = \frac{1}{N}\sum_{n=1}^{N}(D(I_{vis}) - b)^2 + \frac{1}{N}\sum_{n=1}^{N}(D(I_g) - c)^2, \tag{3}$$

where L_D is the loss for discriminator, b and c are soft labels for the discriminator, and $D(\cdot)$ denotes the classification result of the discriminator. It can be learned from the above loss functions that the output of the initial generator is an image with infrared image thermal radiation information and visible image gradient detail. By the feedback of the discriminator, the structure information of visible image can be added to the generated image.

B. Coupled Generative Adversarial Network

Coupled GAN [20] aims to learn the joint distribution of multi-domain images. Except for a portion of the sample drawn from the marginal distribution, it does not require any external information. This is achieved by the weight-sharing mechanism of the generators.

It consists of two parts: the coupled generators and the coupled discriminators. They restrain each other follow the minimax criterion. Let $G_1(\cdot)$ and $G_2(\cdot)$ be the function of the first generator and the second generator, $D_1(\cdot)$ and $D_2(\cdot)$ be the function of the first discriminator and the second discriminator. The constrained minimax game in the coupled GAN can be formulated as

$$\begin{aligned} &\max_{G_1,G_2}\min_{D_1,D_2} L(G_1, G_2, D_1, D_2) \\ &\quad = \mathbb{E}_{I_1}[-\log D_1(I_1)] \\ &\qquad + \mathbb{E}_z[-\log(1 - D_1(G_1(z))] + \mathbb{E}_{I_2}[-\log D_2(I_2)] \\ &\qquad + \mathbb{E}_z[-\log(1 - D_2(G_2(z))], \end{aligned} \tag{4}$$

where z is a random vector, I_1 and I_2 represent images sampled from different domains.

The weights of the front layers in coupled generators which are used to extract the high-level features are shared. This is the key factor that enables coupled generators to learn the joint distribution. The coupled discriminators also employ the weight sharing mechanism in the last few layers. Although this does not help to learn the joint distribution, it can effectively reduce the parameters of the network.

Since coupled GAN can learn the joint distribution of multi-domain images, it is suitable for multi-domain image tasks such as image transformation [22] and domain adaptation [20]. This motivates us to explore coupled GAN for infrared and visible images fusion. Owing to the imaging principles of infrared images and visible images are different, they belong to two distinct domains. But they also have a strong internal correlation. Therefore, understanding their joint distribution will be very helpful for fusion.

Algorithm 1: The Training Process of RCGAN

for *number of epoch* **do**
  for *number of iterations* **do**
    // **update the coupled discriminators** ;
    Pick n images generated from first generator $(I_{g1}^{(1)}, \cdots, I_{g1}^{(n)})$;
    Pick n visible images $(I_{vis}^{(1)}, \cdots, I_{vis}^{(n)})$;
    Update the first discriminator by Adam optimizer;
    Pick n images generated from second generator $(I_{g2}^{(1)}, \cdots, I_{g2}^{(n)})$;
    Pick n infrared images $(I_{ir}^{(1)}, \cdots, I_{ir}^{(n)})$;
    Update the second discriminator by Adam optimizer;
    // **update the coupled generators** ;
    Pick n pre-fused images $(I_{pf}^{(1)}, \cdots, I_{pf}^{(n)})$;
    Pick n infrared images $(I_{ir}^{(1)}, \cdots, I_{ir}^{(n)})$;
    Update the first generator by Adam optimizer;
    Pick n visible images $(I_{vis}^{(1)}, \cdots, I_{vis}^{(n)})$;
    Update the second generator by Adam optimizer;
  end
end

III. Proposed Method

The framework of the proposed RCGAN is represented in Fig. 2. We design a sophisticated coupled GAN to fully exploit the representative information in infrared and visible images. The infrared and visible images are connected on the color channel and fed into parallel coupled generators.

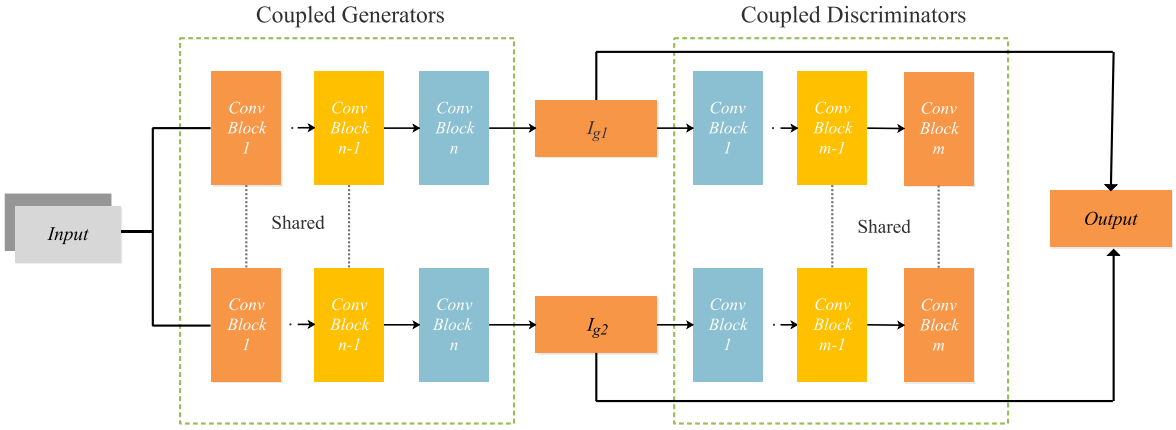


Fig. 2. Architecture of the proposed RCGAN. The infrared and visible images are connected in the color channel as the input for the coupled generators. Then, the two generated image I_{g1} and I_{g2} are fed to the coupled relativistic discriminators. With the game between coupled generators and the relativistic discriminators, both generated are faithful. Finally, the two generated images are averaged to obtain the fused image.

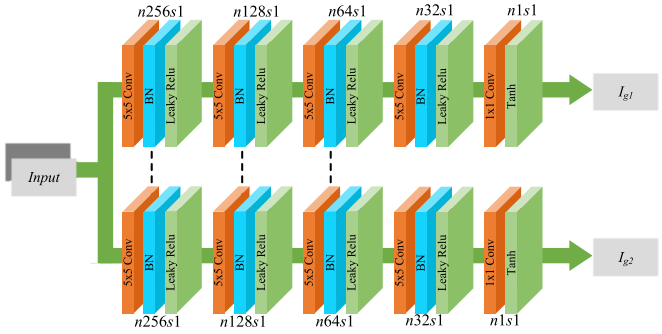


Fig. 3. Architecture of the coupled generators in RCGAN. BN denotes the batch normalization layer. n and s represent the number of the convolutional kernels and the stride, respectively.

The first generator attempts to generate an image with infrared image structure information based on the pre-fused image. The discriminator then measures the relative offset of the generated image to the visible image. Similarly, the second generator is dedicated to enhancing the gradient information of the visible image on the pre-fused image. Its discriminator is dedicated to measuring the offset of the second generated image relative to the infrared image. As the training iteration increases, both generators can obtain a corresponding faithful generated image which contains both structure information of the infrared image and the texture information of the visible image. However, due to the training strategy, the generated image will be relatively biased towards a certain source image. So the two generated images are averaged to get the final fused image. This can take advantage of the generated image and offset its own shortcomings. The training process is drawn in **algorithm** 1.

A. Architecture of Generators

The architecture of the coupled generators is shown in Fig. 3. The two generators have exactly the same structure. Each consists of five convolutional blocks. Each convolutional block contains a convolutional layer, a batch normalization (BN) layer and a leaky ReLU [23] activation layer (except the last convolutional block consists of a convolutional layer and a Tanh activation layer). The kernel size is set to 5×5 for the first and second convolutional blocks, 3×3 for the third and fourth convolutional blocks, and 1×1 for the last convolutional block. In the beginning, a larger convolution kernel, such as 5×5, can have a larger receptive field for extracting features. Then, a smaller convolution kernel, such as 3×3, can optimize the feature map efficiently. The 1×1 convolutional filter is mainly used to reduce the dimension so that the generated image can have the desired color channel. The stride for all convolution operations is set to 1, and no pooling is done. This setting can preserve useful information as much as possible. The number of convolutional kernels for the five convolutional blocks are set to 256, 128, 64, 32, and 1, respectively. Leaky ReLU is used as the activation function for the first four convolutional blocks, and Tanh is used for the last convolutional block. To couple the two generators together, the weights of the first three convolutional blocks are shared. By adopting weight sharing among the generators, on the one hand, the number of parameters can be greatly reduced, on the other hand, this helps to learn the joint distribution of multi-domain images. Thus the feature map in our network is more representative.

We want to make the first generator G_1 learn the thermal radiation information of the infrared image based on the pre-fusion image. So the loss function is set to

$$L_{G_1} = \Phi(G_1) + \alpha L_{content1}, \tag{5}$$

where $\Phi(G_1)$ and $L_{content1}$ denote the GAN loss and content loss for the first generator, which can be formulated as

$$\Phi(G_1) = \frac{1}{N}\sum_{n=1}^{N}\left(D_{Ra1}(I_{g1}^{(n)}, I_{vis}^{(n)}) - a\right)^2$$
$$L_{content1} = \frac{1}{WH}(\beta \left\| I_{g1} - I_{ir} \right\|_2^2 + \left\| I_{g1} - I_{pf} \right\|_2^2), \tag{6}$$

where $D_{Ra1}(I_{real}, I_{fake})$ is the function of the first relativistic discriminator in which the data I_{real} tends to be real is labeled as a. Correspondingly, I_{fake} is the data tends to be fake, I_{g1} stands for the first generated image obtained by G_1, $L_{content1}$ denotes the content loss for the first generator. α and

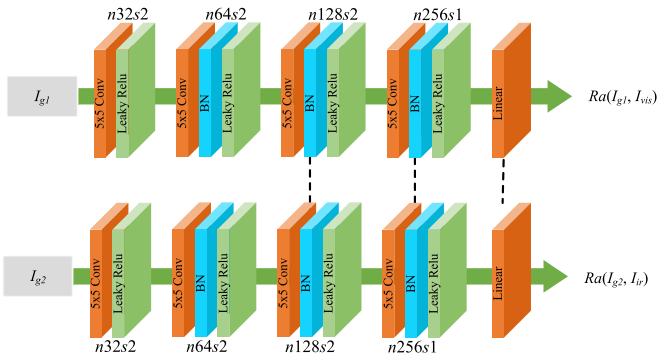


Fig. 4. Architecture of the coupled relativistic discriminators in RCGAN. BN denotes the batch normalization layer. *n* and *s* represent the number of the convolutional kernels and the stride, respectively.

β (in 5) are two factors two control the ration of the content and infrared image I_{ir}.

With the help of this loss function, the first generated image I_{g1} can simultaneously persevere details from the pre-fused image I_{pf} and learn radiation information of the infrared image I_{ir}.

Similarly, we want to inject the gradient information of the visible image I_{vis} into the second generated image based on the pre-fusion image I_{pf}. Therefore, the loss function of the second generator is set to

$$L_{G_2} = \Phi(G_2) + \alpha L_{content2}, \tag{7}$$

where $\Phi(G_2)$ and $L_{content2}$ are formulated as

$$\Phi(G2) = \frac{1}{N}\sum_{n=1}^{N}(D_{Ra2}(I_{g2}^{(n)}, I_{ir}^{(n)}) - a)^2$$

$$L_{content2} = \frac{1}{WH}(\beta \left\|\nabla I_{g2} - \nabla I_{vis}\right\|_2^2 + \left\|I_{g2} - I_{pf}\right\|_2^2), \tag{8}$$

where $D_{Ra2}(I_{real}, I_{fake})$ is the function of the second relativistic discriminator in which the real data is labeled as a. I_{g2} is the generated image of the second generator, and ∇ is the gradient operation. Thus, I_{g2} can learn the gradient details of the visible image based on the pre-fusion image.

Therefore, the role of the coupled generators can be thought of as optimizing the pre-fusion image along with different directions. But each generated image is biased toward a particular source image (see Fig. 1). This biased issue will be alleviated in the subsequently coupled discriminator.

B. Architecture of Relativistic Discriminators

The architecture of the coupled relativistic discriminators is shown in Fig. 4. The two generated image I_{g1} and I_{g2} are input for the coupled relativistic discriminators. Each relativistic discriminator consists of four convolutional blocks and a linear layer. The first block is composed of a convolutional layer and a Leaky Relu activation layer. The next three blocks are made up of a convolutional layer, a batch normalization layer and a Leaky Relu activation layer. The kernel size of all the convolutional layer is set to 3, and the stride is set to 2. Thus the width and height of the feature map will shrink rapidly. The numbers of the kernels for the four convolutional layers are set to 32, 64, 128, and 256, respectively. Only the first block is valid padded. The last linear layer will convert the flatten feature map into one output, which represents the relative distance between I_g and the corresponding image. To reduce the amount of the parameter, the weight for the third and fourth convolutional block and the liner layer are shared as shown in Fig. 4.

The role of the coupled relativistic discriminators is to calculate how the generated image relatively close to another image. In this way, by backpropagation, the generated image can simultaneously contain the information of the corresponding opposite image. In detail, for the first generator, the infrared image is used as part of the loss function to optimize the results, so in the corresponding first discriminator, we calculate how close the first generated image is to the visible image. Thus, the loss function for the first relativistic discriminator D_{Ra1} is set to

$$L_{Ra1} = D_{Ra1}(I_{vis}, I_{g1}), \tag{9}$$

Similarly, the second relativistic discriminator D_{Ra2} aims to measure how the second generated image I_{g2} is relatively close to the infrared image. Thus, its loss function can be formulated as

$$L_{Ra2} = D_{Ra2}(I_{ir}, I_{g2}), \tag{10}$$

where L_{Ra1} and L_{Ra2} denote the loss function for the first relativistic discriminator and the second relativistic discriminator, respectively. The function of two relativistic discriminators can be formulated as

$$\begin{aligned} D_{Ra1}(I_{real}, I_{fake}) &= C_1(I_{real}) - \mathbb{E}_{I_{fake}}[C_1(I_{fake})] \\ D_{Ra2}(I_{real}, I_{fake}) &= C_2(I_{real}) - \mathbb{E}_{I_{fake}}[C_2(I_{fake})], \end{aligned} \tag{11}$$

where $C_1(\cdot)$ and $C_2(\cdot)$ denote the non-linear transformation of two discriminators. The coupled relativistic discriminators allow the single generated image to have information of the opposite image. But the resulting image is still biased to some extent (See Fig. 1), so the two generated images are averaged to obtain the final fusion result F as

$$F = 0.5 \times (I_{g1} + I_{g2}). \tag{12}$$

Since the two images, I_{g1} and I_{g2}, are both generated on the basis of the pre-fusion image, the averaging operation can make the final resulting image reserve both infrared thermal radiation information and the visible image texture information based on the pre-fused image.

There is no doubt that using a more complex network model, such as ResNet [24] and DenseNet [25], can significantly improve the capabilities of the network. But for a fair comparison, the design of the single branch of RCGAN is similar to FusionGAN.

IV. Experimental Results and Analysis

A. Training Details

To train the RCGAN, 39 pairs of infrared and visible images from TNO dataset are used. We first convert the training images to gray, and crop patches from them with the size

of 120×120 and stride of 14. Then, the training patches are centralized to $[-1, 1]$. Since the coupled generators employ no padding operation, the width and height of the feature map will shrink. To solve this issue, all the patches are zero-padded to 132×132. The infrared image patch and visible image patch are connected in the color channel and fed to the coupled generators. Thus the shape of the generated image is 120×120. The two parameters λ and ϵ which are used to adjust the ratio are set to 100 and 6, respectively. The label a for the coupled discriminator is set to 1. We use Adam [26] as the optimizer, and the batch size is 32. A complete epoch consists of a coupled discriminators training session at first and a coupled generators training session at second, and 100 epoch is trained. We employ the fusion results of GFF [2] as the pre-fused image. The implementation of RCGAN is available at https://github.com/liqilei/RCGAN.

B. Comparison With State-of-the-Art Methods

To demonstrate the superiority of the proposed RCGAN, it is compared with seven state-of-the-art methods, which are guided filter based fusion (GFF) [2], adaptive sparse representation (ASR) [11], sparse representation (SR) [27], Laplacian pyramid with sparse representation (LP-SR) [28], convolutional neural network based fusion (CNN) [29], fusion with convolutional neural network (FCNN) [15], and FusionGAN [19]. The implements or models of these methods are publicly available, and all the parameters are set in accordance with the corresponding publications. The experiments are performed on two public datasets including TNO dataset [30] and INO dataset [31]. The INO dataset consists of a number of infrared and visible video pairs, and we capture some frames from them to test our method. Besides the subjective visual effect, the structural similarity index (SSIM) [32], which measure the structural similarity between the source images and the fused image, is employed to objectivity measure the proposed method. Since the fused image should persevere the representative structures of both infrared image and visible image. Thus, a higher SSIM can reflect a better fusion result.

For TNO dataset, four representative image pairs (*bench, house, fennek01, and jeep*) is used to evaluate our methods. The *bench* image pair depicts a man sitting on a bench with a reflection in front of the river. But the light is too dark to see the man in the visible image. The *fennek01* pair shows a vehicle that is traveling on the road. The third image pair, called *house*, shows a house with light, but the light source is only visible in the visible image. The last image, called *jeep*, shows a jeep at night. The sky and details of the car can only be seen in the infrared image, while the surrounding environment is more clear in the visible image. The images and the fusion results of the RCGAN and other seven contrast methods are shown in Fig. 5, and the corresponding quantitative evaluation of SSIM is given in Table I. In general, SR easily introduces a large amount of noise and spots into the fused image. For example, in the *bench* image, white spots appear in the river, and the shade of the man is difficult to recognize. Besides, a lot of noise is introduced on the ground

TABLE I
QUANTITATIVE EVALUATION OF SSIM ON FOUR PAIRS OF INFRARED AND VISIBLE IMAGES IN TNO DATASET

images	GFF	ASR	SR	LP-SR	CNN	FCNN	FusionGAN	RCGAN
bench	0.564	0.544	0.550	0.549	0.558	0.558	0.487	**0.586**
fennek01	0.735	0.758	0.731	0.760	0.736	0.760	0.653	**0.800**
house	0.748	0.768	0.699	0.745	0.743	0.735	0.715	**0.775**
jeep	0.694	0.708	0.675	0.676	0.682	0.677	0.660	**0.711**

TABLE II
QUANTITATIVE EVALUATION OF SSIM ON THREE PAIRS OF INFRARED AND VISIBLE IMAGES IN INO DATASET

images	GFF	ASR	SR	LP-SR	CNN	FCNN	FusionGAN	RCGAN
CoatDeposit	0.757	0.772	0.712	0.750	0.748	0.750	0.690	**0.770**
MulitpleDeposit	0.737	0.744	0.696	0.735	0.732	0.735	0.677	**0.748**
ParkingSnow	0.604	0.618	0.528	0.593	0.606	0.590	0.560	**0.619**

in the *house* image. To make matters worse, in the *jeep* pair, SR cannot fuse the information from the infrared image and causes the body of the jeep is hard to identify. Since the fusion rules are designed based on sparse coefficients in these SR based methods, a slight change in the sparse coefficient may cause a huge disaster in spatial the domain. Besides, a single dictionary in SR [27] is difficult to represent the structural features of infrared and visible images simultaneously. Therefore, the SSIM for the fused image is relatively low. ASR, which employs multiple dictionaries, can alleviate this problem. However, ASR's performance for some details is still poor. For example, in the *jeep* pair, it can be seen from the enlarged visible image that the wheel has a small circle in the innermost. But these contrast methods, including ASR, fail to fuse this detail. Besides, ASR requires a lot of running time and computational costs. The multi-scale methods, including GFF and LP-SR, rely too much on the decomposition level and the manual activity measurement. Thus their fusion results may neglect some important information. For instance, in the *bench* image, the used image of GFF is dark overall. For LP-SR, the fusion result of *house* neglects the light source, and the fusion result for *jeep* ignores the innermost circuit of the wheel in the enlarged image. The CNN introduces a white block in the river on the *bench* image and makes the light source in the *house* image invisible. Similarly, FCNN ignores some details in the fused images, such as the innermost circles inside the wheel. This is because the two methods are not accurate enough on detecting the activity level. As analyzed earlier, a single GAN network cannot exploit the representative information of the multi-domain images. Thus, FusionGAN's results are often too smooth and have a severe halo effect, which is not conducive to visual perception. The fusion results of our RCAN are clearer and most of the representative details are preserved.

We also evaluation RCGAN and other contrast methods on three image pairs. The three image pairs are captured in different scenes, and they are named *CoatDeposit, MultipleDeposit, and ParkingSnow* in INO dataset. The visual effect of the fusion results is shown in Fig. 6, and the objective measurement is drawn in Table II. It can be seen that the fusion results of SR suffer from heavy spatial distortion. Especially for the *ParkingSnow*, it is difficult to recognize the appearance of the snow in the fused image. ASR and

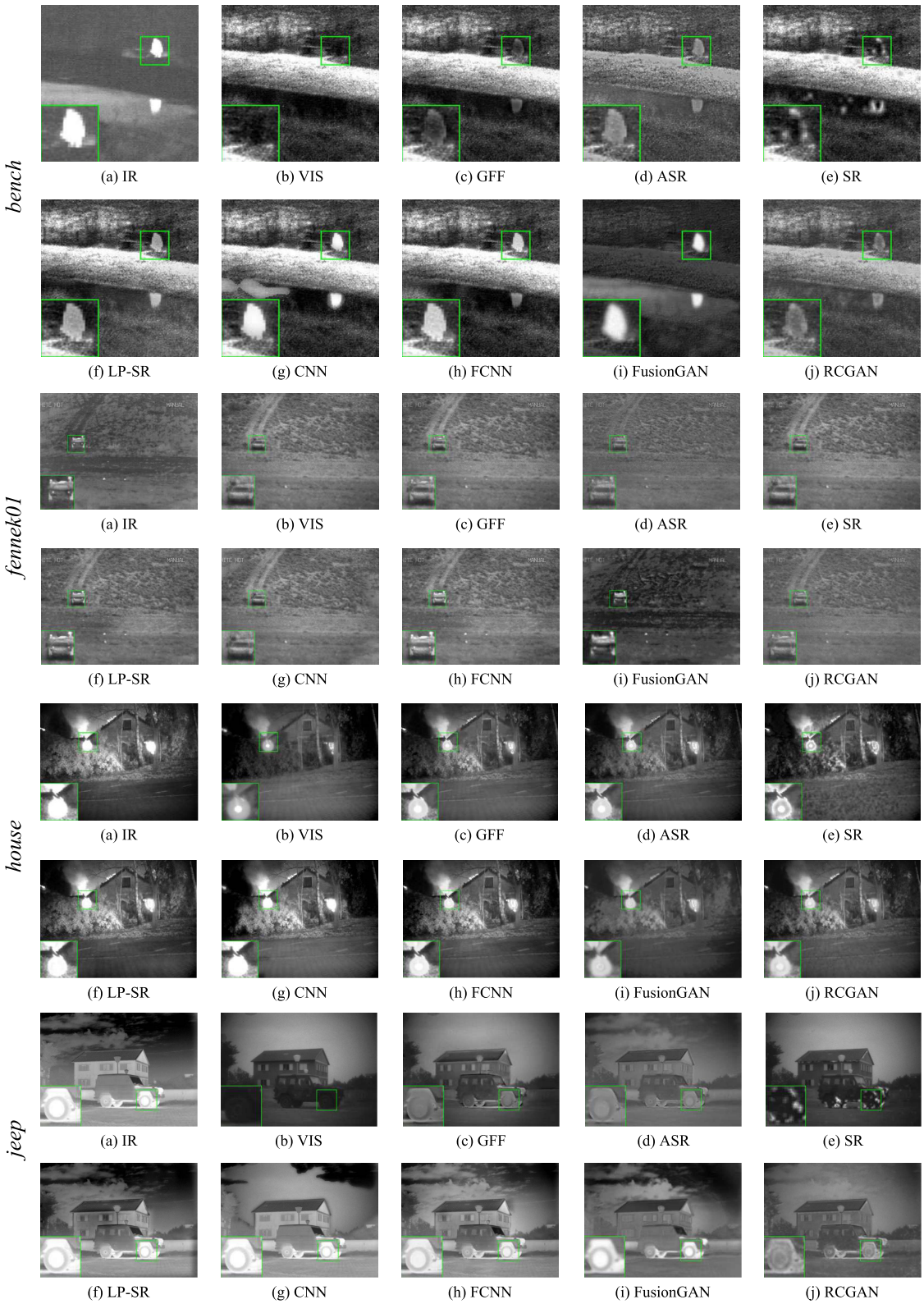


Fig. 5. Fusion results for four representative infrared and visible images from the TNO dataset. The frist two rows are the fusion results for *bench* image pair; the third and fourth rows are the results for *fennek01* image pair; the fifth and sixth rows are the results for *house* image pair; the last two rows are the fusion results for *jeep* image pair. For a better comparison, a small region is enlarge in the green box.

FusionGAN introduce holo around the objects. Due to the inappropriate activity level measurement, the fused images of LP-SR and FCNN are often unsatisfactory. In the fusion results of LP-SR and FCNN on *ParkingSnow* image, the body of the man is gray and the head of the man is white. This is unfriendly for perception and analysis. Compared with other methods, RCGAN can make full use of information from infrared and visible images, and the fused images are rich

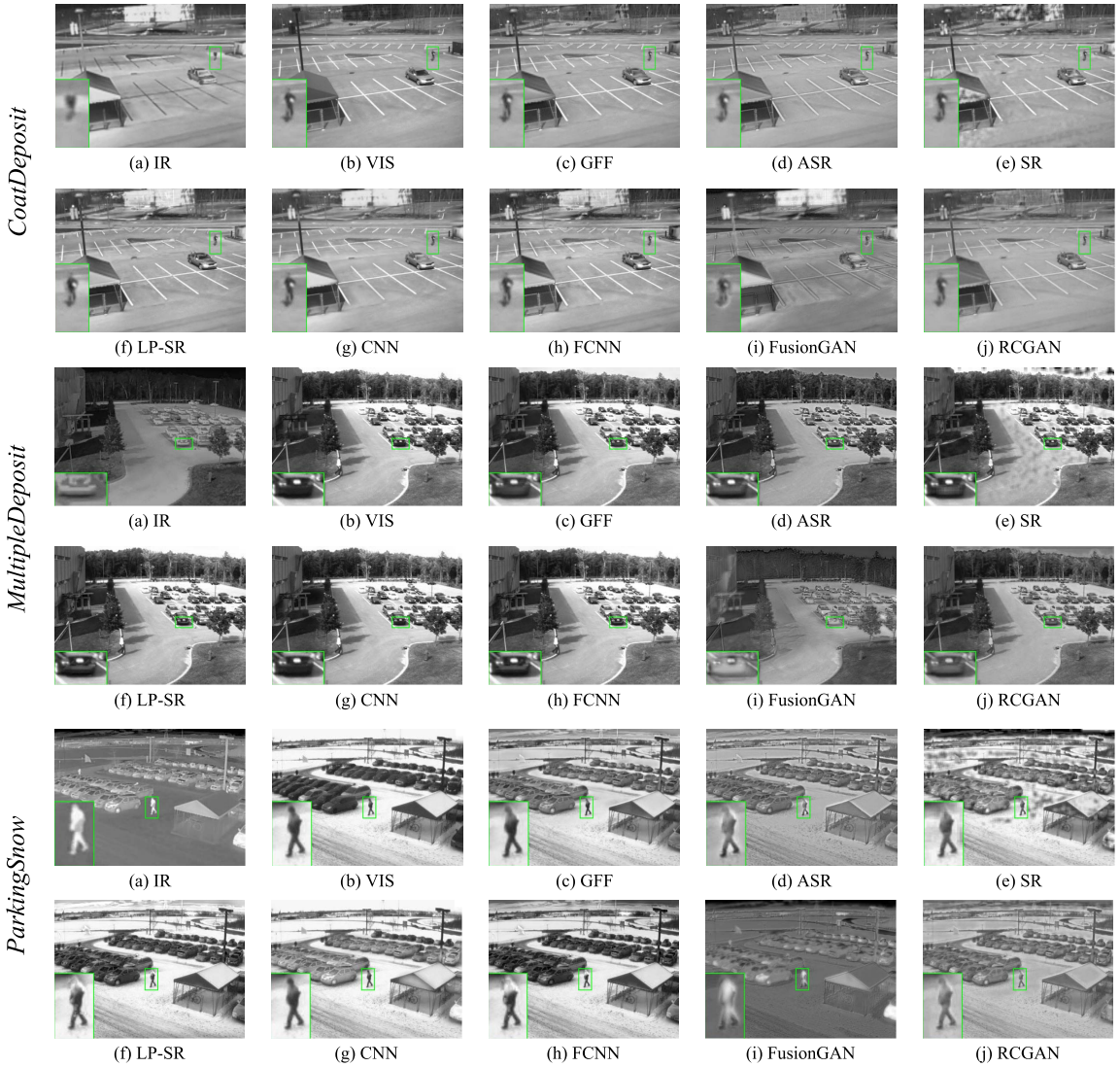


Fig. 6. Fusion results for three representative infrared and visible images from the INO dataset. The frist two rows are the fusion results for *CoatDeposit* image pair; the third and fourth rows are the results for *MulitpleDeposit* image pair; the last two rows are the fusion results for *ParkingSnow* image pair. For a better comparison, a small region is enlarge in the green box.

in detail and can highlight key objectives, such as the man in *CoatDeposit* and *ParkingSnow*. The quantitative evaluation results in Table II also demonstrate our method can produce a satisfying fusion result which can persevere most structure information of infrared and visible images. The visual effects and the objective evaluation results of both TNO dataset and INO dataset demonstrate that our RCGAN has obvious advantages over other state-of-the-art methods.

C. Ablation Study

1) Ablation Study on Relativistic Discriminator: To demonstrate the importance of the relativistic discriminator, we design CGAN, which employs the coupled standard absolute discriminators instead of the coupled relativistic discriminators. The other settings are exactly the same as those of RCGAN except for the coupled discriminators. The loss functions for the coupled discriminators in CGAN are formulated as

$$\begin{aligned} D_1(x) &= C_1(x) \\ D_2(x) &= C_2(x), \end{aligned} \tag{13}$$

where $C_1(x)$ and $C_2(x)$ are the non-transformed output of the two standard discriminators.

We visualize the loss of discriminators in RCGAN and CGAN during training as shown in Fig. 7. Since the different categories of discriminators, the magnitude of value is different. But it is still obvious that the loss of RCGAN equipped with relativistic discriminator can reduce stably. The CGAN with standard discriminator is accompanied by severe jitter, which is not conducive to the training progress. In addition, relativistic discriminator helps the network converge faster. With the help of relativistic discriminator, both discriminators in RCGAN can converge around 20 epochs. Nevertheless, CGAN requires about 30 epochs to reach the same level.

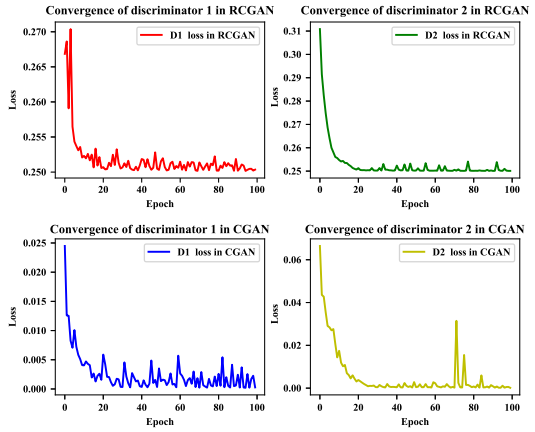


Fig. 7. Ablation study on relativistic discriminator. The first row shows the loss of RCGAN equiped with relativistic discriminator, the second row shows the CGAN with standard discriminator.

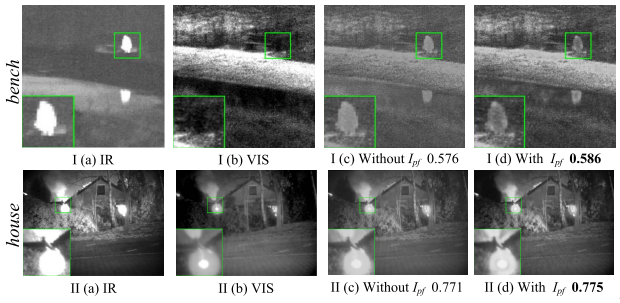


Fig. 8. Ablation study on the pre-fused image I_{pf}. The first row shows the fusion results for *bench* image pair; the second row shows the results for *house* image pair. The values in (c) and (d) indicate the SSIM evaluation metrics.

2) Ablation Study on Pre-Fused Image: To demonstrate the advantages of using pre-fused images, a comparative experiment in which no pre-fused image used is designed. In detail, when training the coupled generator, only use the infrared/visible image is used as its label. The coupled discriminator part is not redesigned since it does not involve pre-fusion images. Except for the content loss of coupled generators, all other training details are exactly the same as before. The content loss in the comparative experiment is

$$L_{content1} = \frac{1}{WH}(\beta \left\| I_{g1} - I_{ir} \right\|_2^2)$$
$$L_{content2} = \frac{1}{WH}(\beta \left\| \nabla I_{g2} - \nabla I_{vis} \right\|_2^2). \quad (14)$$

We compared the fused images obtained from this network with those obtained by RCGAN in both subjective aspect and objective aspect in Fig. 8. The fused images without I_{pf} are dim. For example, for the *bench* image pair, I_{pf} helps the forest and the bank become more contrasting, which helps the perception of the human visual system. For the *house* image pair, by using I_{pf}, the bushes illuminated by the lights are more vivid. The outline of the lighting source is also sharper. From the quantitative indicators, by using I_{pf}, a higher SSIM value can be obtained, which means that the fused image can retain more structure information of infrared and visible images. In general, I_{pf} allows the fusion of infrared and visible images to have a direction.

V. Conclusion

In this paper, an efficient infrared and visible image fusion method named RCGAN was proposed. The RCGAN exploited the coupled GAN structure to achieve an end to end fusion operation. We creatively transformed the issue of generating an image to optimize the pre-fused image. Nevertheless, due to the limitation of training data, RCGAN can only handle specific kinds of infrared images. Besides, the coupled GAN still requires a fair amount of parameters. In the future, more in-depth work could be carried out on fusing multiple kinds of infrared images and reducing the parameters of the network.

References

[1] Z. Liu, E. Blasch, Z. Xue, J. Zhao, R. Laganiere, and W. Wu, "Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: A comparative study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 94–109, Jan. 2012.

[2] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2864–2875, Jul. 2013.

[3] 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.

[4] L. Jian, X. Yang, Z. Zhou, K. Zhou, and K. Liu, "Multi-scale image fusion through rolling guidance filter," *Future Gener. Comput. Syst.*, vol. 83, pp. 310–325, Jun. 2018.

[5] Q. Li, X. Yang, W. Wu, K. Liu, and G. Jeon, "Multi-focus image fusion method for vision sensor systems via dictionary learning with guided filter," *Sensors*, vol. 18, no. 7, p. 2143, 2018.

[6] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. COMM-31, no. 4, pp. 532–540, Apr. 1983.

[7] H. Li, B. S. Manjunath, and S. K. Mitra, "Multisensor image fusion using the wavelet transform," *Graph. Models Image Process.*, vol. 57, no. 3, pp. 235–245, 1995.

[8] Q. Zhang and B.-L. Guo, "Multifocus image fusion using the nonsubsampled contourlet transform," *Signal Process.*, vol. 89, no. 7, pp. 1334–1346, 2009.

[9] X. Lu, B. Zhang, Y. Zhao, H. Liu, and H. Pei, "The infrared and visible image fusion algorithm based on target separation and sparse representation," *Infr. Phys. Technol.*, vol. 67, pp. 397–407, Nov. 2014.

[10] Q. Li, X. Yang, W. Wu, K. Liu, and G. Jeon, "Pansharpening multispectral remote-sensing images with guided filter for monitoring impact of human behavior on environment," *Concurrency Comput., Pract. Exper.*, p. e5074, 2018. doi: 10.1002/cpe.5074.

[11] Y. Liu and Z. Wang, "Simultaneous image fusion and denoising with adaptive sparse representation," *IET Image Process.*, vol. 9, no. 5, pp. 347–357, 2015.

[12] C.-I. Chen, "Fusion of PET and MR brain images based on IHS and log-Gabor transforms," *IEEE Sensors J.*, vol. 17, no. 21, pp. 6995–7010, Nov. 2017.

[13] Z. Liu and W. Wu, "Fusion with infrared images for an improved performance and perception," in *Pattern Recognition, Machine Intelligence and Biometrics*. Berlin, Germany: Springer, 2011, pp. 81–108.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[15] Y. Liu, X. Chen, J. Cheng, H. Peng, and Z. Wang, "Infrared and visible image fusion with convolutional neural networks," *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 16, no. 3, 2018, Art. no. 1850018.

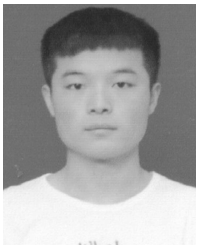
[16] H. Li, X.-J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 2705–2710.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[18] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

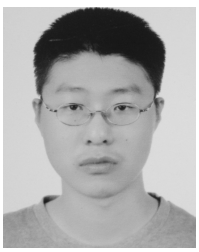
[19] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, Aug. 2019.

[20] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 469–477.
[21] A. Jolicoeur-Martineau, "The relativistic discriminator: A key element missing from standard GAN," 2018, *arXiv:1807.00734*. [Online]. Available: https://arxiv.org/abs/1807.00734
[22] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 700–708.
[23] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, Jun. 2013, vol. 30, no. 1, p. 3.
[24] 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.
[25] 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.
[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980
[27] B. Yang and S. Li, "Multifocus image fusion and restoration with sparse representation," *IEEE Trans. Instrum. Meas.*, vol. 59, no. 4, pp. 884–892, Apr. 2010.
[28] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Inf. Fusion*, vol. 24, pp. 147–164, Jul. 2015.
[29] 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.
[30] A. Toet. (Apr. 2014). *TNO Image Fusion Dataset*. [Online]. Available: https://figshare.com/articles/TNO_Image_Fusion_Dataset/1008029
[31] O. I. of Canada, *Video Analytics Dataset*. [Online]. Available: https://www.ino.ca/en/technologies/video-analytics-dataset/
[32] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.



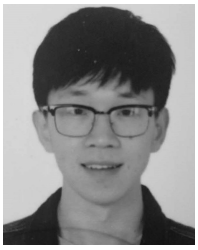
Qilei Li (S'19) received the B.S. degree in electronic information engineering from the Shandong University of Technology. He is currently pursuing the M.S. degree with the College of Electronics and Information Engineering, Sichuan University, Chengdu, China.

His research interests are image processing and deep learning.



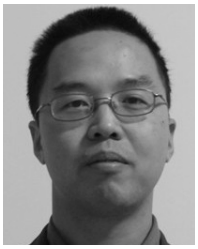
Lu Lu (M'18) was born in Chengdu, China, in 1990. He received the Ph.D. degree from the School of Electrical Engineering, Southwest Jiaotong University, Chengdu, in 2018.

From 2017 to 2018, he was a Visiting Ph.D. Student with the Electrical and Computer Engineering, McGill University, Montreal, QC, Canada. He is currently a Post-Doctoral Doctorate with the College of Electronics and Information Engineering, Sichuan University. His research interests include adaptive signal processing, kernel methods, and distributed estimation.



Zhen Li (S'19) is currently pursuing the M.S. degree with the College of Electronics and Information Engineering, Sichuan University, Chengdu, China.

His research interests are image restoration and deep learning.



Wei Wu received the B.S. degree from Tianjin University in 1998 and the M.S. and Ph.D. degrees in communication and information system from Sichuan University in 2003 and 2008, respectively. He is currently a Professor with the College of Electronics and Information Engineering, Sichuan University. His research interests fall under the umbrella of image processing, particularly image enlargement, super-resolution, image enhancement, as well as computational intelligence.

Zheng Liu (SM'05) received the Ph.D. degree in engineering from Kyoto University, Kyoto, Japan, in 2000, and the Ph.D. degree from the University of Ottawa in 2007.

In 2015, Dr. Liu started his laboratory on Intelligent Sensing, Diagnostics, and Prognostics with the University of British Columbia at Okanagan, Kelowna, BC, Canada.

Gwanggil Jeon (M'12) received the B.S., M.S., and Ph.D. degrees from Hanyang University in 2003, 2005, and 2008, respectively.

From 2009 to 2011, he was a Post-Doctoral Fellow with the University of Ottawa, Ottawa, ON, Canada. From 2011 to 2012, he was an Assistant Professor with Niigata University. He is currently a Professor with Xidian University, Xi'an, China, and Incheon National University, Incheon, South Korea. His research interests fall under the umbrella of image processing.

Xiaomin Yang (M'19) received the B.S. and Ph.D. degrees in communication and information system from Sichuan University in 2002 and 2007, respectively.

She is currently a Professor with the College of Electronics and Information Engineering, Sichuan University. Her research interests fall under the umbrella of image processing, particularly image enlargement, super-resolution, image enhancement, as well as computational intelligence.