Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 4209963, 9 pages
https://doi.org/10.1155/2021/4209963

Research Article

A Generative Adversarial Network with Dual Discriminators for Infrared and Visible Image Fusion Based on Saliency Detection

Dazhi Zhang,[1] **Jilei Hou**,[2,3] **Wei Wu**,[2,3] **Tao Lu**,[2,3] **and Huabing Zhou**[2,3]

[1]*China Nuclear Power Operation Technology Corporation, Wuhan 430000, China*
[2]*College of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China*
[3]*Hubei Key Laboratory of Intelligent Robot, Wuhan Institute of Technology, Wuhan 430205, China*

Correspondence should be addressed to Jilei Hou; houjilei455@gmail.com

Received 13 September 2021; Accepted 11 November 2021; Published 24 November 2021

Academic Editor: Sergey A. Suslov

Infrared and visible image fusion needs to preserve both the salient target of the infrared image and the texture details of the visible image. Therefore, an infrared and visible image fusion method based on saliency detection is proposed. Firstly, the saliency map of the infrared image is obtained by saliency detection. Then, the specific loss function and network architecture are designed based on the saliency map to improve the performance of the fusion algorithm. Specifically, the saliency map is normalized to [0, 1], used as a weight map to constrain the loss function. At the same time, the saliency map is binarized to extract salient regions and nonsalient regions. And, a generative adversarial network with dual discriminators is obtained. The two discriminators are used to distinguish the salient regions and the nonsalient regions, respectively, to promote the generator to generate better fusion results. The experimental results show that the fusion results of our method are better than those of the existing methods in both subjective and objective aspects.

1. Introduction

Image fusion aims to utilize complementary information of two source images to synthesize a fusion image with a more comprehensive understanding of the scene [1, 2]. The infrared image can identify the target according to thermal radiation contrast, and the visible image can provide a clear image in line with the human visual system [3, 4]. Due to the characteristics of the infrared and visible image, the fusion result of the infrared and visible image can preserve the significant target of the infrared image and the texture detail of the visible image simultaneously [5, 6]. Infrared and visible image fusion has been widely used in many fields, such as target recognition, video surveillance, and scene understanding [7–9].

The key of image fusion is to integrate the effective information and remove the redundant information of the source image to gain a better fusion image [10, 11]. For this purpose, a large number of infrared and visible image fusion methods have been proposed. These methods can be divided into two categories: (i) traditional methods, which usually complete the fusion task based on mathematical transformation and manual design; (ii) deep learning-based methods, which usually use the specific loss function to optimize the neural network to obtain a fusion result [12].

Although the abovementioned methods can complete the fusion task successfully, several aspects still need to be improved. Firstly, the manually designed fusion rules lead to the traditional methods being more complex and time-consuming. Secondly, some methods only apply deep learning in the part of fusion process, which is difficult to give full play to the advantages of deep learning [13]. Thirdly, due to the lack of ground truth, GAN-based methods are difficult to determine the input of the discriminator. The existing methods usually use two technical routes to solve this problem: (i) using one source image as the input of the discriminator, which will inevitably lead to the gradual loss of information of the other source image [14]; (ii) using the generative adversarial network with dual discriminators, which takes both two source images as the input. However,

this scheme is difficult to control the balance between the two discriminators [15].

To address these challenges, this paper proposes *a generative adversarial network with dual discriminators for infrared and visible image fusion based on saliency detection* (SDGAN). Firstly, the proposed method is based on deep learning, which optimizes the network through specific loss functions and overcomes the increasing complexity caused by manually designed fusion rules. Secondly, in order to solve the problem of lacking ground truth, we use the generative adversarial network with dual discriminators to deal with the fusion problem. At the same time, in order to maintain the balance between the two discriminators, we introduce saliency detection into image fusion. The two discriminators take significant targets and nonsignificant targets as inputs, respectively, to ensure that the two discriminators can work smoothly without conflict.

2. Related Work

2.1. Infrared and Visible Image Fusion

2.1.1. Traditional Methods. The traditional fusion method can be divided into three steps: feature extraction, feature fusion, and feature reconstruction [16]. As feature reconstruction is usually an inverse process of extraction, the key of traditional methods is feature extraction and feature fusion. By employing different strategies for feature extraction and feature fusion, a large number of fusion methods have been proposed.

For feature extraction, there are four categories: (i) multiscale transform, such as pyramid transform [17], wavelet transform, and edge-preserving filter; (ii) sparse representation [18]; (iii) subspace analysis, such as independent component analysis (ICA) [19], principal component analysis (PCA) [20], and nonnegative matrix factorization (NMF) [21]; (iv) hybrid methods, which can combine the advantages of other methods to obtain better fusion results [22]. After feature extraction, appropriate feature fusion methods must be selected to fuse these features. The commonly used rules include four categories: (i) maximum-operation; (ii) minimum-operation; (iii) addition-operation; (iv) L1-norm constraints.

Although the above three steps summarize most traditions, some methods are not suitable for the above-mentioned three steps, such as GTF, which is based on gradient transfer and total variation minimization [23].

2.1.2. Deep Learning-Based Methods. Although the traditional fusion methods can gain a satisfactory fusion image, the fusion methods are generally very complex and time-consuming due to the artificially designed fusion rules. With the rise of deep learning, more and more fusion methods based on deep learning are proposed.

Li and Wu [24] employed the encode/decode network architecture and introduced the densely connected convolution layer in the encoder to extract the features of the source image to avoid losing information in the convolution process. Yang et al. [25] proposed a fusion model based on visual saliency sparse representation and detail injection to avoid the loss of significant thermal radiation targets of infrared images. Zhang et al. [26] propose an image fusion network based on proportional maintenance of gradient and intensity, named PMGI, which can preserve source image information through the gradient and intensity path. With the rise of the generative adversarial networks, Feng et al. [11] tried to use GAN to solve the problem of image fusion, named FusionGAN. Subsequently, its variant [27] was proposed by introducing the target-enhancement loss to enhance edge details of the fused image. However, these methods force the fused image to obtain more details in visible images as the adversarial game proceeds. In contrast, the thermal information of infrared images is gradually lost. To address this issue, Ma et al. [15] introduced dual discriminators into GAN to avoid excessive loss of information in the source image.

2.2. Saliency Detection.

The human visual system will focus on important regions of the image, which helps humans easily obtain important information. Saliency detection aims to simulate the human visual system to extract the significant regions of the image and prioritize allocating computing resources for important regions in subsequent processing.

Itti et al. [28] first combined the multiscale features to get an initial saliency map and then used a neural network to optimize the initial saliency map to get the final result. Hou and Zhang [29] extracted the spectral residual of an image in the spectral domain by analyzing the log spectrum of an input image and proposed a fast method to construct the corresponding saliency map in the spatial domain. Cheng et al. [30] proposed a saliency detection method based on regional contrast, simultaneously evaluating global contrast differences and spatial coherence. Traditional saliency detection methods mainly rely on manual extraction of features and then combine these features to obtain a saliency map. Vig et al. [31] proposed an entirely automatic data-driven method that performs a large-scale search for optimal features to gain a saliency map. Kümmerer et al. [32] first used depth networks to solve saliency detection, which can reuse existing networks that have been pretrained on the task of object recognition in models of fixation prediction. Then, a large number of saliency detection methods based on the neural network have been proposed and achieved good results.

3. Proposed Method

3.1. Problem Formulation.

The infrared image can highlight the target by the difference of thermal radiation. Relatively, the visible image contains richer texture details. Infrared and visible image fusion can retain the highlighted target of the infrared image and the texture details of the visible image simultaneously. Saliency detection can extract highlighted targets of the image. Therefore, introducing saliency detection into infrared and visible image fusion can improve the performance of the image fusion algorithm.

For a given infrared image $\mathbf{I}_r$, the significance value $S(k)$ of pixel k is obtained by calculating the distance between pixel k and all other pixels i on the image, which can be defined as follows:

$$S(k) = \sum_{i \in I_r} |k - i|. \tag{1}$$

The significance map $\mathbf{S}$ of the infrared image $\mathbf{I}_r$ can be obtained by calculating the significance values of all pixels of the infrared image pixel by pixel.

Then, the weight map $\mathbf{w}$ can be obtained by normalizing all values on the saliency map $\mathbf{S}$ to the interval [0, 1], which can be used to constrain the fusion weights of different targets in the loss function. The calculation process of the weight is shown as follows:

$$\mathbf{w} = \frac{\mathbf{S}}{255}. \tag{2}$$

Finally, the saliency map $\mathbf{S}$ is binarized to extract the salient region of the image successfully. Specifically, the pixel value of the saliency map $\mathbf{S}$ is normalized pixel by pixel. If the pixel value is greater than b, the corresponding pixel value in the mask $\mathbf{m}$ takes the value 1; otherwise, it is 0. In this paper, the mask can be obtained when b is determined as 0.25. The mask $\mathbf{m}$ can be obtained when all pixels are calculated. The mask $\mathbf{m}$ calculation process is shown as follows:

$$\forall a \in \mathbf{S}, \quad a = \begin{cases} 1, & a > b, \\ 0, & a \le b. \end{cases} \tag{3}$$

As shown in Figure 1, the saliency map and mask of two typical infrared images are given. It can be seen that the saliency map can indeed detect the significant target of the infrared image and the mask can indeed represent the significant area.

Given an infrared image and a visible image, the goal of image fusion is to obtain a generator constrained by the source image. The fused image generated by the generator can retain the salient target of the infrared image and the texture details of the visible image at the same time.

This paper proposes a generative adversarial network with dual discriminators for infrared and visible image fusion based on saliency detection, named SDGAN. The entire procedure of our proposed SDGAN is shown in Figure 2. The infrared image and visible image are input to the generator G to gain an initial fused image. However, it is difficult to obtain a satisfactory fusion image only by the generator G; therefore, two discriminators D_r and D_v are introduced in our network to establish the adversarial games with the generator. The generator G can generate a better fused image through adversarial games. The discriminator D_r is used to distinguish the salient regions of the fused image and the infrared image. The discriminator D_v is used to distinguish the nonsalient regions of the fused image and the visible image. The significant region of the image can be obtained by multiplying the mask $\mathbf{m}$ and the image pixel by pixel, and the nonsalient region can be obtained by multiplying the mask $(1 - \mathbf{m})$ and the image pixel by pixel. Since the two discriminators are used to distinguish the complementary regions of the source image and the fused image, the two discriminators can complete their own tasks independently without conflict.

The goal of generator G is to synthesize a fused image, which can make it difficult for both discriminators to distinguish whether the input image is from the fused image or the source image at the same time. Mathematically, the training goal of generator G is minimization:

$$\begin{aligned} &\min_G \max_{D_r, D_v} E\left[\log D_r\left(\mathbf{m} \odot \mathbf{I}_r\right)\right] \\ &+ E\left[\log\left(1 - D_r\left(\mathbf{m} \odot \mathbf{I}_f\right)\right)\right] \\ &+ E\left[\log D_v\left((1 - \mathbf{m}) \odot \mathbf{I}_v\right)\right] \\ &+ E\left[\log\left(1 - D_v\left((1 - \mathbf{m}) \odot \mathbf{I}_f\right)\right)\right], \end{aligned} \tag{4}$$

where $\odot$ represents the *Hadamard* product, G represents the generator, D_r and D_v represent the two discriminators, $\mathbf{m}$ represents the mask, which is used to extract the salient area of the image, and $(1 - \mathbf{m})$ is used to extract the nonsalient area of the image. The training goal of D_r and D_v is to maximize equation (4).

3.2. Loss Function. The original GAN is prone to lead to artifacts and noisy or other incomprehensible results in the generated image due to the instability of its training process. In order to make the training process more stable, a common solution is to introduce content loss. To improve the quality of fusion image, in addition to adversarial loss $\mathscr{L}_{\text{con}}$, this paper also introduces an enhancement loss $\mathscr{L}_{\text{enh}}$. Therefore, the loss function of generator $\mathscr{L}_G$ mainly includes three parts: content loss $\mathscr{L}_{\text{con}}$, enhancement loss $\mathscr{L}_{\text{enh}}$, and adversarial loss $\mathscr{L}_{\text{adv}}$, as shown in

$$\mathscr{L}_G = \mathscr{L}_{\text{con}} + \lambda_1 \mathscr{L}_{\text{enh}} + \lambda_2 \mathscr{L}_{\text{adv}}, \tag{5}$$

where λ_1 and λ_2 are introduced to control the tradeoff.

The content loss $\mathscr{L}_{\text{con}}$ is used to constrain the similarity between the fused image and the source image in content, which mainly consists of two parts, gradient loss $\mathscr{L}_{\text{grad}}$ and intensity loss $\mathscr{L}_{\text{int}}$, as shown in

$$\mathscr{L}_{\text{con}} = \mathscr{L}_{\text{grad}} + \gamma \mathscr{L}_{\text{int}}, \tag{6}$$

where γ is obtained to maintain the balance between the gradient loss $\mathscr{L}_{\text{grad}}$ and intensity loss $\mathscr{L}_{\text{int}}$.

The gradient loss $\mathscr{L}_{\text{grad}}$ is committed to preserving the texture details of the source image in the fused image, which is defined as follows:

$$\mathscr{L}_{\text{grad}} = \mathbb{E}\left[\left\|\nabla \mathbf{I}_f - \nabla \mathbf{I}_r\right\|_2^2 + \xi_1 \left\|\nabla \mathbf{I}_f - \nabla \mathbf{I}_v\right\|_2^2\right], \tag{7}$$

where ∇ represents the gradient operator, which is used to extract the gradient of the image, $\|\cdot\|_2$ represents the Euclidean norm, and ξ_1 is used to balance two items.

The intensity loss $\mathscr{L}_{\text{int}}$ is used to constrain that the fused image and the source image have similar intensity distribution, which is defined as follows:

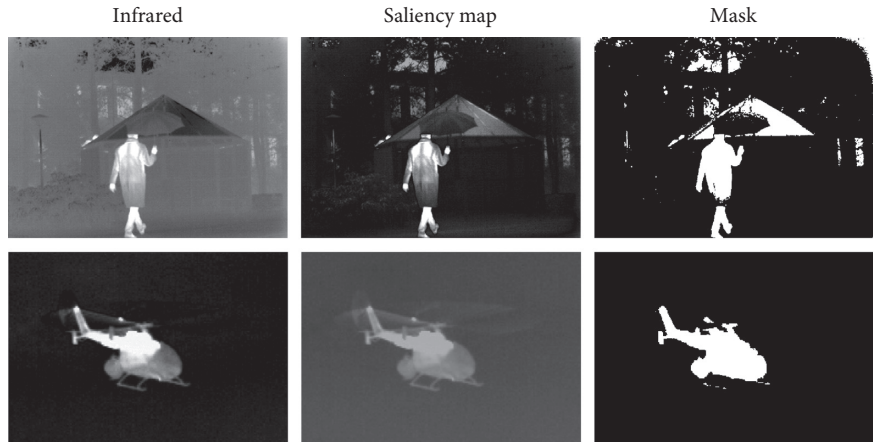

Figure 1: The saliency map and mask of two typical infrared images.

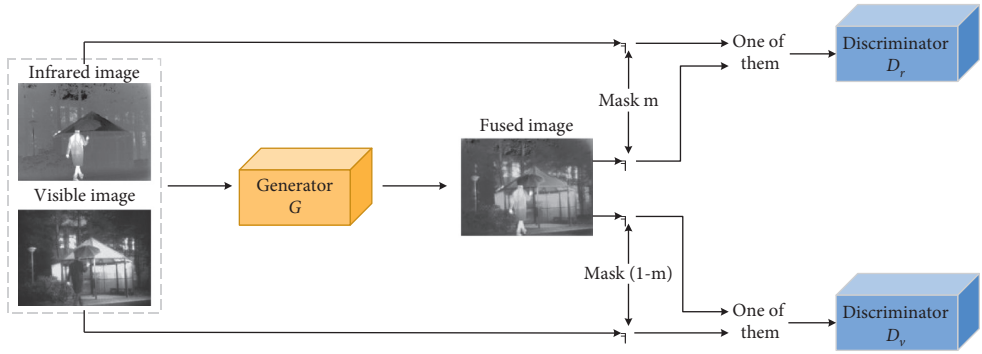

Figure 2: The entire procedure of our SDGAN for infrared and visible image fusion.

$$\mathscr{L}_{\text{int}} = \mathbb{E}\left[\left\|\mathbf{I}_f - \mathbf{I}_r\right\|_2^2 + \xi_2\left\|\mathbf{I}_f - \mathbf{I}_v\right\|_2^2\right], \tag{8}$$

where ξ_2 is introduced to control the tradeoff.

The enhancement loss $\mathscr{L}_{\text{enh}}$ is mainly used to enhance the highlighted targets and the texture details, which is defined as follows:

$$\mathscr{L}_{\text{enh}} = \mathbb{E}\left[\left\|\mathbf{w}\odot\left(\mathbf{I}_f - \mathbf{I}_r\right)\right\|_2^2 + \xi_3\left\|(1-\mathbf{w})\odot\left(\nabla\mathbf{I}_f - \nabla\mathbf{I}_v\right)\right\|_2^2\right], \tag{9}$$

where $\mathbf{w}$ represents the weight map, which is used to control the retention degree of significant targets in the fused image, $(1-\mathbf{w})$ is used to control the retention degree of nonsignificant targets in the fused image, and ξ_3 is used to balance two items.

The adversarial loss $\mathscr{L}_{\text{adv}}$ comes from the game between the generator and the discriminators, as shown in

$$\begin{aligned}\mathscr{L}_{\text{adv}} = {} & \mathbb{E}\left[\log\left(1 - D_r\left(\mathbf{m}\odot\mathbf{I}_f\right)\right)\right] \\ & + \mathbb{E}\left[\log\left(1 - D_v\left((1-\mathbf{m})\odot\mathbf{I}_f\right)\right)\right],\end{aligned} \tag{10}$$

where $\mathbf{m}$ represents the mask, which is used to extract the significant region of the fused image, and $(1-\mathbf{m})$ is used to extract the nonsignificant region of the fused image.

In order to make the generator converge smoothly, two discriminators D_r and D_v are obtained to construct the adversarial relationship between the generator and the discriminators. The loss functions of the two discriminators D_r and D_v are defined as follows:

$$\begin{aligned}\mathscr{L}_{D_r} = {} & \mathbb{E}\left[-\log\left(D_r\left(\mathbf{m}\odot\mathbf{I}_r\right)\right)\right] \\ & + \mathbb{E}\left[-\log\left(D_r\left(\mathbf{m}\odot\mathbf{I}_f\right)\right)\right], \\ \mathscr{L}_{D_v} = {} & \mathbb{E}\left[-\log\left(D_v\left((1-\mathbf{m})\odot\mathbf{I}_v\right)\right)\right] \\ & + \mathbb{E}\left[-\log\left(D_v\left((1-\mathbf{m})\odot\mathbf{I}_f\right)\right)\right].\end{aligned} \tag{11}$$

3.3. Network Architecture

3.3.1. Generator Architecture. As shown in Figure 3, a dual-encoder-single-decoder structure is introduced in the generator. Two encoders are used to extract the features of two source images, respectively. Each path of the encoder adopts four-layer network architecture for feature extraction. All convolution kernel sizes are set to 3×3. All steps are set to 1, and batch normalization and ReLU activation function are used to avoid the vanishing gradient and speed up network convergence. Moreover, dense connections are employed in each encoder path to realize feature reuse [33]. For the decoder, the output of the dual encoder is connected as the input to reconstruct the fused image. The decoder also adopts a four-layer network architecture, with the

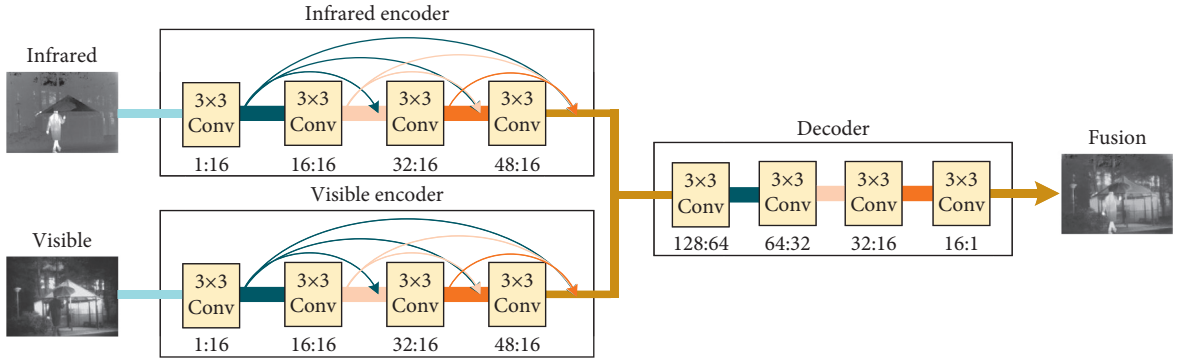

Figure 3: The overall architecture of the generator.

convolution kernel size of 3×3, and contains batch normalization and LReLU activation functions.

3.3.2. Discriminator Architecture. Two discriminators D_r and D_v are used to establish the adversarial game with the generator and promote the generator to generate more realistic and detailed images by distinguishing the input. The discriminator D_r is used to distinguish the true and false aspects of significant targets between the fused image and infrared image, and the discriminator D_v is used to distinguish the true and false aspects of nonsignificant targets between the fused image and the visible image. The architecture of two discriminators D_r and D_v is the same, but they does not share parameters. The network architecture is shown in Figure 4. The first four layers are convolution kernels with a size of 3×3, and the activation function is LReLU. The last layer is the linear layer, and the activation function is tanh, which is used to generate a scalar to estimate the probability that the input is from real data. The step of all convolutions is set to 2.

4. Experimental Results

4.1. Dataset and Training Details. The training dataset comes from the public infrared and visible dataset TNO, which is the most commonly used dataset in infrared and visible image fusion tasks. 28 images are selected from the dataset TNO to train the model in this paper; however, only 28 images are not enough to train a good model. Therefore, the clipping strategy is carried to expand the training dataset, and each image is cropped into image patches with the size of 120×120. Eventually, 23364 image patches can be used to train the model.

In the training process, the generator selects 32 pairs of infrared and visible image patches as input at one time. Next, we used 32 pairs of the salient areas of the infrared and fused image patches as the input of the discriminator D_r. Simultaneously, 32 pairs of the nonsalient areas of the visible and fused image patches are used to input into the discriminator D_v. We first train the discriminator 1 time and then train the generator until reaching the maximum number of training iterations. All the parameters of our model are updated by the Adam optimizer [34] at a learning rate of 10^{-4}.

4.2. Compared Methods and Objective Indexes. As we all know, we need to make qualitative and quantitative comparisons with the existing advanced methods in order to evaluate the performance of our method. For qualitative comparisons, we compare our method with five existing methods, including three traditional methods, i.e., LP [35], DTCWT [36], and FPDE [37], and two deep learning-based methods, i.e., FusionGAN [14], and DenseFuse [24]. All traditional methods run on the same CPU i7-7700k, while deep learning-based methods run on the same GPU GTX 1080ti. All comparison methods are implemented based on public code, and the parameters are default.

Although qualitative comparison can measure the performance of the method to a certain extent, it is easy to be affected by people's subjectivity. In this paper, qualitative comparisons are used to evaluate our method more comprehensively. Three qualitative metrics are adopted to evaluate our SDGAN and other comparison methods, i.e., EN [38], SD [39], and SSIM [40].

Entropy (EN) is a common parameter for statistical image features, reflecting the amount of information obtained from infrared and visible images. Mathematically, entropy can be defined as follows:

$$\mathrm{EN} = -\sum_{l=0}^{L-1} p_l \log\, p_l, \tag{12}$$

where L denotes the gray level of the image and p_l is the normalized histogram with the gray-scale value of l in the fused image.

Standard deviation (SD) represents the dispersion of image gray-scale value relative to the average gray-scale value, defined as in

$$\mathrm{SD} = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N} \left(F(i,j) - \mu_F\right)^2}, \tag{13}$$

where $F(i, j)$ is the pixel value of the fused image F in the i-th row and the j-th column, $M \times N$ denotes the size of the fused image F, and μ_F is the average pixel value of the fused images.

Structural similarity (SSIM) mainly simulates image loss and distortion from three aspects: loss of correlation,

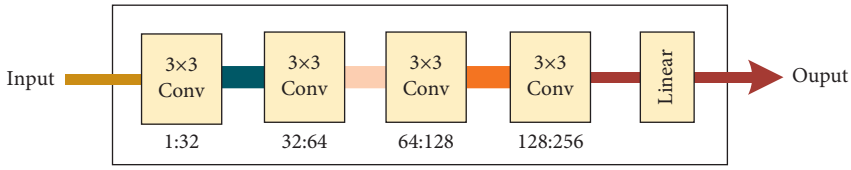

Figure 4: The architecture of our discriminators.

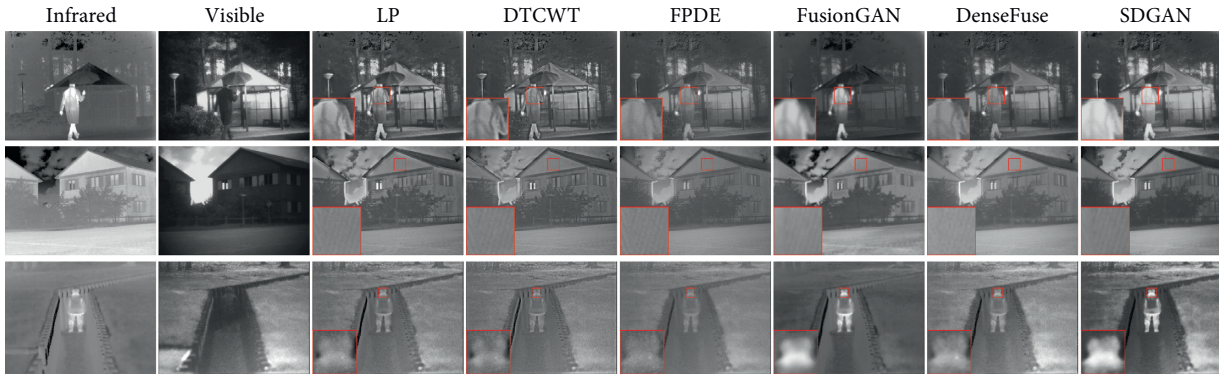

Figure 5: Qualitative fusion results on three typical infrared and visible image pairs.

Table 1: Mean of three metrics on fusion results.

Metrics	LP	DTCWT	FPDE	FusionGAN	DenseFuse	*W/o* discriminators	SDGAN
EN	6.5542	6.4287	6.3093	6.4072	6.9590	7.0055	**7.0370**
SD	27.5637	25.0506	23.8361	26.4110	35.4724	38.7233	**40.3962**
SSIM	0.7240	0.7277	0.7072	0.6015	0.6767	0.6662	**0.7443**

luminance, and contrast distortion. The product of the three components is the evaluation result of the fused image, which can be defined as follows:

$$\mathrm{SSIM}_{x,y} = \sum_{x_i,y_i} \frac{2\mu_{x_i}\mu_{y_i} + \mathrm{C}_1}{\mu_{x_i}^2 + \mu_{y_i}^2 + \mathrm{C}_1} \cdot \frac{2\sigma_{x_i}\sigma_{y_i} + \mathrm{C}_2}{\sigma_{x_i}^2 + \sigma_{y_i}^2 + \mathrm{C}_2} \cdot \frac{\sigma_{x_i y_i} + \mathrm{C}_3}{\sigma_{x_i}\sigma_{y_i} + \mathrm{C}_3}, \tag{14}$$

where x and y represent two source images. μ denotes the mean value, σ represents the standard deviation/covariance, and C_1, C_2, and C_3 are the parameters to make the metric stable.

4.3. Qualitative Comparisons. Qualitative comparisons mainly evaluate the performance of the method according to the human visual system. In this paper, three typical infrared and visible images are used to evaluate the method. The experimental results are shown in Figure 5.

From top to bottom in Figure 5, the three lines are the fusion result of Kaptein_1654, Marne_02, and soldier_in_trench_2. From left to right, the first two columns in Figure 5 present the original infrared images and visible images. The last column is the fusion results of the proposed SDGAN, and the remaining columns correspond to the fusion results of LP, DTCWT, FPDE, FusionGAN, and DenseFuse.

As shown in Figure 5, all methods can complete the fusion task, but all comparison methods can only better retain the information of a certain source image. For example, the fusion result of FusionGAN better retains the significant target of the infrared image but loses a lot of texture details of the visible image. Relatively, although the fusion results of LP, DTCWT, FPDE, and DenseFuse retain the texture details of the visible image, the significant targets of the infrared image are not so prominent. In contrast, the fusion results of our SDGAN can highlight significant targets and retain texture details of visible images at the same time, such as the human in Kaptein_1654 and soldier_in_trench_2, significantly brighter than other comparison methods. The texture details on the wall in Marne_02 are clearer than in other methods.

4.4. Quantitative Comparisons. Quantitative comparisons show difficulty in avoiding the influence of people's subjective emotions. In order to evaluate our SDGAN more comprehensively, quantitative comparisons are also employed to evaluate our fusion methods. It is not comprehensive to use only one objective metric to evaluate the fusion method. Therefore, entropy (EN), standard deviation (SD), and structural similarity (SSIM) are used to evaluate the methods in this paper. Quantitative comparisons are performed on 32 image pairs. The results are shown in Table 1.

Table 2: Mean of running time of different methods.

Method	LP	DTCWT	FPDE	FusionGAN	DenseFuse	SDGAN
Time (s)	0.0060	1.2625	0.3937	0.1719	0.1444	0.1311

Figure 6: Ablation experiments related to the discriminators.

From the experimental results, we can find that the proposed SDGAN achieves the optimal results on three metrics. The optimal entropy shows that the fusion result of our SDGAN obtains the most information from source images, which shows that the fusion method in this paper is indeed effective and can retain rich source image information. The largest standard deviation shows that the fused image of our method has higher contrast, which proves that the fused image retains more intensity information of the infrared image in the significant area and more texture details of the visible image in the nonsignificant area by significance detection. The optimal structural similarity shows a strong correlation between the fusion results of SDGAN and the source images, and the fused image does not have serious distortion, which shows that the fusion method proposed in this paper can retain more information from the source image.

The average running time of LP, DTCWT, FPDE, FusionGAN, DenseFuse, and the proposed SDGAN is presented in Table 2. It can be seen that the average running time of this method is second only to LP, indicating that this method does not lose the efficiency of the algorithm on the basis of improving the quality of the fused image.

4.5. Ablation Experiment. In order to generate high-quality fusion images, two discriminators are employed in our network. We have conducted ablation experiments to verify the role of two discriminators by removing all the discriminators. The comparison results are given in Figure 6. We can find that our GIDGAN can better preserve the texture details of the visible image while preserving the significant targets of the infrared image, such as the result of our GIDGAN can better preserve details of shrubs in the first row.

5. Conclusions

In this paper, an infrared and visible image fusion method based on saliency detection is proposed. The saliency map of the infrared image is extracted through saliency detection, which is employed not only in the loss function to train the model but also in network architecture. We obtain a generative adversarial network with dual discriminators. The saliency map can divide the image into significant regions and nonsignificant regions, and the dual discriminators can be used to identify them, respectively. By the adversarial game, the generator can generate more realistic fusion images with highlighted target and rich texture details. Qualitative and quantitative experiments show that the proposed SDGAN can achieve the promised effect that the fused image retains both the salient target and rich texture details.

Data Availability

The dataset used to support the findings of this study are included within the open data collection in https://figshare.com/articles/TNO20Image20Fusion20Dataset/%201008029.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to thank Jiayi Ma and Hui Li for providing their codes. This work was supported in part by the National Natural Science Foundation of China under Grants 62171327, 62171328, and 62072350, the Hubei Technology Innovation Project under Grant 2019AAA045, the Key Scientific and Technological Research Project of Hubei Provincial Education Department under Grant D20201507, the first batch of application basic technology and science research foundation in Hubei Nuclear Power Operation Engineering Technology Research Center under Grant B210610, and the Nuclear Energy Development Project (Sub-Project: Artificial Intelligence in Nuclear Reactors) in State Administration of Science, Technology and Industry for National Defence, PRC under Grant ZX200302.

References

[1] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: a survey," *Information Fusion*, vol. 45, pp. 153–178, 2019.

[2] Y. Zhang, D. Li, and W. Zhu, "Infrared and visible image fusion with hybrid image filtering," *Mathematical Problems in Engineering*, vol. 2020, Article ID 1757214, 2020.

[3] J. Hou, D. Zhang, W. Wu, J. Ma, and H. Zhou, "A generative adversarial network for infrared and visible image fusion based on semantic segmentation," *Entropy*, vol. 23, no. 3, 2021.

[4] X. Yun, Y. Sun, X. Yang, and N. Lu, "Discriminative fusion correlation learning for visible and infrared tracking," *Mathematical Problems in Engineering*, vol. 2019, Article ID 2437521, 11 pages, 2019.

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

[6] Y. Zhao, G. Fu, H. Wang, and S. Zhang, "The fusion of unmatched infrared and visible images based on generative adversarial networks," *Mathematical Problems in Engineering*, vol. 2020, Article ID 3739040, 12 pages, 2020.

[7] J. Ma, H. Zhang, Z. Shao, P. Liang, and H. Xu, "GANMcC: a generative adversarial network with multiclassification constraints for infrared and visible image fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–14, 2020.

[8] H. Xu, H. Zhang, and J. Ma, "Classification saliency-based rule for visible and infrared image fusion," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 824–836, 2021.

[9] H. Zhang and J. Ma, "SDNet: a versatile squeeze-and-decomposition network for real-time image fusion," *International Journal of Computer Vision*, pp. 1–25, 2021.

[10] X. Fan, P. Shi, J. Ni, and M. Li, "A thermal infrared and visible images fusion based approach for multitarget detection under complex environment," *Mathematical Problems in Engineering*, vol. 2015, Article ID 750708, 11 pages, 2015.

[11] Z. Feng, X. Zhang, L. Yuan, and J.-N. Wang, "Infrared target detection and location for visual surveillance using fusion scheme of visible and infrared images," *Mathematical Problems in Engineering*, vol. 2013, Article ID 720979, 7 pages, 2013.

[12] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: a survey and perspective," *Information Fusion*, vol. 76, 2021.

[13] H. Li, X. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 2705–2710, Beijing, China, August 2018.

[14] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: a generative adversarial network for infrared and visible image fusion," *Information Fusion*, vol. 48, pp. 11–26, 2018.

[15] 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 Transactions on Image Processing*, vol. 29, pp. 4980–4995, 2020.

[16] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: a unified unsupervised image fusion network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[17] H. Jin and Y. Wang, "A fusion method for visible and infrared images based on contrast pyramid with teaching learning based optimization," *Infrared Physics & Technology*, vol. 64, pp. 134–142, 2014.

[18] Q. Zhang, Y. Liu, R. S. Blum, J. Han, and D. Tao, "Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: a review," *Information Fusion*, vol. 40, pp. 57–75, 2018.

[19] N. Mitianoudis and T. Stathaki, "Pixel-based and region-based image fusion schemes using ica bases," *Information Fusion*, vol. 8, no. 2, pp. 131–142, 2007.

[20] U. Patil and U. Mudengudi, "Image fusion using hierarchical PCA," in *Proceedings of the 2011 International Conference on Image Information Processing*, pp. 1–6, Shimla, India, November 2011.

[21] W. Kong, Y. Lei, Y. Lei, and J. Zhang, "Technique for image fusion based on non-subsampled contourlet transform domain improved nmf," *Science China Information Sciences*, vol. 53, no. 12, pp. 2429–2440, 2010.

[22] Y. Liu, X. Chen, R. K. Ward, and Z. Jane Wang, "Image fusion with convolutional sparse representation," *IEEE Signal Processing Letters*, vol. 23, no. 12, pp. 1882–1886, 2016.

[23] J. Ma, C. Chen, C. Li, and J. Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Information Fusion*, vol. 31, pp. 100–109, 2016.

[24] H. Li and X.-J. Wu, "DenseFuse: a fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2614–2623, 2019.

[25] Y. Yang, Y. Zhang, S. Huang, Y. Zuo, and J. Sun, "Infrared and visible image fusion using visual saliency sparse representation and detail injection model," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–15, 2020.

[26] H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma, "Rethinking the image fusion: a fast unified image fusion network based on proportional maintenance of gradient and intensity," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 12797–12804, New York, NY, USA, February 2020.

[27] J. Ma, P. Liang, W. Yu et al., "Infrared and visible image fusion via detail preserving adversarial learning," *Information Fusion*, vol. 54, pp. 85–98, 2020.

[28] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[29] X. Hou and L. Zhang, "Saliency detection: a spectral residual approach," in *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Minneapolis, MN, USA, June 2007.

[30] M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 37, no. 3, pp. 569–582, 2014.

[31] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proceedings of the 2014 IEEE Conference On Computer Vision And Pattern Recognition*, pp. 2798–2805, Columbus, OH, USA, June 2014.

[32] M. Kümmerer, L. Theis, and M. Bethge, "Deep gaze I: boosting saliency prediction with feature maps trained on imagenet," 2014, https://arxiv.org/abs/1411.1045.

[33] G. Huang, Z. Liu, L. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, Honolulu, HI, USA, July 2017.

[34] D. P. Kingma and B. J. Adam, "A method for stochastic optimization," 2014, https://arxiv.org/abs/1412.6980.

[35] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a com-pact image code," *IEEE Transactions on Communications*, vol. 31, no. 4, 1987.

[36] J. J. Lewis, R. J. O'Callaghan, S. G. Nikolov, D. R. Bull, and N. Canagarajah, "Pixel-and region-based image fusion with complex wavelets," *Information Fusion*, vol. 8, no. 2, pp. 119–130, 2007.

[37] D. P. Bavirisetti, G. Xiao, and G. Liu, "Multi-sensor image fusion based on fourth order partial differential equations," in *Proceedings of the International Conference on Information Fusion*, pp. 1–9, Xi'an, China, July 2017.

[38] J. W. Roberts, J. A. Van Aardt, and F. B. Ahmed, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *Journal of Applied Remote Sensing*, vol. 2, no. 1, 2008.

[39] Y.-J. Rao, "In-fibre bragg grating sensors," *Measurement Science and Technology*, vol. 8, no. 4, 1997.

[40] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.