Hindawi
Advances in Multimedia
Volume 2023, Article ID 5588547, 13 pages
https://doi.org/10.1155/2023/5588547

Research Article

FFA-GAN: A Generative Adversarial Network Based on Feature Fusion Attention for Intelligent Safety Monitoring

R. Chang,[1,2] **B. Zhang**,[1] **Y. Zhang**,[1] **S. Gao**,[3] **S. Zhao**,[2,4] **Y. Rao**,[2,4] **X. Zhai**,[2,5] **T. Wang**,[2,4] **and Y. Yang**[2,4]

[1]*Yuxi Power Supply Bureau, Yunnan Power Grid Corporation, Yuxi 653100, China*
[2]*The Laboratory of Pattern Recognition and Artificial Intelligence, Yunnan Normal University, Kunming 650500, China*
[3]*Guangzhou Jiansoft Technology Co Ltd, Guangzhou 510600, China*
[4]*School of Information Science and Technology, Yunnan Normal University, Kunming 650500, China*
[5]*School of Physics and Electronic Information, Yunnan Normal University, Kunming 650500, China*

Correspondence should be addressed to S. Zhao; szhaoynnu@yeah.net

Received 18 January 2023; Revised 24 April 2023; Accepted 2 May 2023; Published 13 May 2023

Academic Editor: Yu-Chen Hu

Copyright © 2023 R. Chang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the rapid development of the national power grid, there is an increasing and strict demand for accurate intelligent management. However, the current detection algorithms have limited abilities under adverse conditions, especially in regions like Yunnan Province with complex terrain. To address this issue, we propose a method that utilizes infrared and visible images to make the images more informative, thereby improving the accuracy of the detection algorithm for electric power construction site safety. First, we design channel attention (CA) module and pixel attention (PA) module to focus on more important channels and resist thick haze pixels that focus on the thick haze pixels and more important channel information. Furthermore, we design a two-stage discriminator which imposes two restrictions on the fused results. Finally, we conduct a large number of comparison experiments with state-of-the-art methods, and the results show that our proposed fusion method achieves excellent performance in infrared and visible image fusion. This method has good prospects for application in the safety supervision of power construction sites and provides a line of defense for construction workers.

1. Introduction

The purpose of power grid construction management is to prevent and reduce power accidents, as well as to prevent serious impacts on society. This serves as the guarantee for power enterprises to fulfill social responsibilities and improve economic benefits. Throughout the process, monitoring and early warning to ensure the safety of construction workers have always been crucial but challenging [1]. Currently, there are two main approaches to power grid construction management: traditional methods and deep learning-based methods. The former relies on manual inspection and screening by security guards, while the latter employs deep learning algorithms to achieve automatic detection [2, 3]. Specifically, the traditional approach to managing construction safety is to establish safety policies, safety objectives, and safety culture based on safety theory, in order to enhance the safety awareness of construction workers. This approach mainly relies on manual monitoring, whereby workers are reminded to pay attention to safety and comply with safety regulations through broadcast alarms and remote calls to the person in charge, when violations are detected during patrols. This method aims to create a safety-conscious working environment and reduce the risk of accidents. Power grid safety supervision and management encompasses various aspects, such as historical event management, current situation management, and task assignment. Most of these methods are postmanagement measures, but they can effectively reduce the occurrence of accidents, and more importantly, enhance the real-time supervision and even early warning capabilities. Therefore, in recent years, developing a power grid security

supervision and management platform and promoting the informatization level of power grid security management have become a research hotspot.

Thanks to the breakthrough in deep learning for computer vision tasks, it is now possible to utilize deep learning based object detection technologies to enable intelligent monitoring of power grid construction sites [4–6], safety vests detection [7], unsafe behavior detection [8], and unauthorized intrusion detection [9]. These can be further analyzed to trigger security alarms. The abovementioned deep learning inspired detection algorithms are required to recognize, interpret, and comprehend images in surveillance video sequences. It is necessary to recognize complex scene systems based on the semantic model representing the monitored scene. This enables complex events to be identified from the surveillance video obtained from the work environment, which is critical in creating a safe work environment and tracking employees. To overcome these challenges, image fusion techniques have been proposed, which can combine complementary information sources from multiple images of the same scene. Its purpose is to enhance the information of a single generated image [10]. Image fusion can provide detailed and reliable images for high-level visual tasks. It plays a crucial role in computer vision, and have been applied in many aspects, including object detection [11, 12], pedestrian recognition [13], face recognition [14, 15], semantic segmentation [16], and other areas.

In power grid construction, safety management is of utmost importance, but the existing methods have several limitations. Traditional management strategies rely on manual inspection and screening by security personnel, which may not be able to prevent accidents immediately. In addition, these methods are labor-intensive and inefficient. This is partly due to the unique characteristics of power grids, including their extensive construction areas, complex site backgrounds, and a large workforce. As a result, implementing traditional methods on power grid construction sites can be challenging [17].

With the wide application of surveillance video technology, power grid enterprises have begun to use video recording to check for security risks. Although this strategy can alleviate some of the problems associated with traditional methods, long-term manual monitoring is prone to fatigue and can result in missed detection. Studies have shown that when an individual observes two monitoring screens simultaneously, they can miss 45% of useful information in 10 minutes and 95% in 22 minutes [18]. Therefore, artificial naked eye monitoring has limitations in accuracy and real-time performance.

Intelligent automatic image detection based on deep learning can greatly enhance detection accuracy and efficiency. However, the automatic detection algorithms used in power grid construction are all based on visible light. Visible light-based on-site operation video safety monitoring is often subject to environmental light, posture, expression, and ornaments, which lead to difficulties in accurately identifying and tracking specific targets, especially in complex working environments and harsh climate conditions.

Yunnan Province is located in the southwest of China. Due to the complexity of terrain and environment, Yunnan power grid is most difficult to monitor and manage in China and even in the world [19]. For instance, due to the foggy mountain area, the quality of the acquired visible image is unsatisfactory, which affects the subsequent high-level visual tasks. The use of infrared-visible image fusion algorithm before detection can effectively improve the accuracy of detection, but the application of infrared-visible image fusion in power grid is rarely studied [20]. This paper presents a novel and effective fusion method based on infrared and visible image fusion for power grid construction management. We summarize our major contributions as follows:

(i) This paper presents the first application of the infrared-visible image fusion method and explores a multifeature infrared-visible light multisource image enhancement technology. The proposed approach aims to improve the video monitoring effect for remote monitoring personnel.

(ii) In this work, we design a shared convolution group consisting of channel attention and pixel attention in a two-branch generator network, which is conducive to capturing the common modal features of infrared and visible images and generating stable and reliable fusion images. To address the issue of foggy environments caused by mountainous terrain in Yunnan, we have designed the two-stage discriminator for GAN-based method in the proposed method. This designment improves the perceptual and interpretive qualities of visible light images captured in foggy conditions, and enhances the effectiveness of subsequent high-level visual tasks.

(iii) Most power management algorithms currently in use are based on postmanagement, which means they can only detect and respond to security risks after they have occurred. In reality, early warning and real-time supervision are essential for truly preventing accidents from happening. The model proposed in this paper will be integrated into the power grid artificial intelligence such as intelligent video surveillance systems. This will allow for more proactive and effective safety management in power grid construction.

The remainder of this paper is organized as follows: Section 2 briefly describes the related works of existing deep learning based power grid operation safety management technology and multimodal image fusion algorithms. In Section 3, we introduce our method in detail, including the network architecture and function modules. Section 4 presents experiments to show the impressive performance of our method in comparison with state-of-the-art methods, followed by some concluding remarks in Section 5.

2. Related Work

2.1. Deep Learning-Based Management Technology. On the basis of artificial intelligence technology, the power grid is developing towards intelligentization, automation, digitization, and informatization. Existing research of power grid management mainly focused on detecting unsafe factors in construction based on deep learning algorithm. Faster R-CNN and deep ResNet were used to quickly and accurately detect workers against complex backgrounds in [21]. IFaster R-CNN approach is used to automatically detect workers and heavy equipment in real-time in [22]. YOLO v3 algorithm is used to detect whether the helmet is worn by the standard in [18]. An improved lightweight YOLOv4 model is used to detect the transmission line insulator defects in [23]. Immediately, Tang et al. [7] provide a method based on YOLOv5 to resolve the problems of low detection efficiency and poor accuracy caused by complex background and numerous personnel. Tan et al. [24] improved YOLOv5, the functional detection scale is increased to realize the detection of smaller targets. For the purpose of the lack of dataset in power grid scenarios, Peng et al. [3] proposed a contrastive Res-YOLOv5 network for intelligent safety monitoring on power grid construction sites.

2.2. Multimodal Image Fusion Algorithm. Image fusion (IF) is an emerging field for generating a robust and informative image through the integration of images obtained by different sensors, e.g., visible, infrared, computed tomography, and magnetic resonance imaging [25, 26]. Among them, infrared-visible image fusion has attracted much attention [14, 27], and it is more informative than that of single mode signal because visible image captures reflected light, while infrared image captures thermal radiation [28]. As mentioned above, the application of those fusion algorithms can be mainly divided into the following categories:

2.2.1. Face Recognition. Li et al.[29] proposed a GAs based infrared-visible image fusion to solve the problem of low face recognition sensitivity caused by glasses occlusion. Heo et al. provide two types of visual and thermal infrared images fusion methods to enhance the robustness of face recognition [10].

2.2.2. Object Detection. Han and Bhanu [30] proposed a search scheme based on the hierarchical genetic algorithm to achieve automatic registration color images and thermal image sequences, and then further used multiple fusion strategies to fuse registration and infrared images for human contour detection. Ulusoy and Yuruk [31] conducted background modeling and foreground detection for infrared, visual intensity, and visual color domains, respectively, so that the complementary regions from the domain were combined, the infrared foreground was covered by this fusion information, and the infrared foreground was fused with the covered visual foreground. Finally, active contour lines are applied to each connected part in the infrared domain to detect the object boundary. Gao et al. [32] proposed a flexible framework for visible and infrared video fusion moving target detection based on the low rank sparse decomposition method. Ma et al. [33] propose an end-to-end STDFusionNet to realize salient target detection. Zuo et al. [34] designed an attention fusion feature pyramid network for infrared small target detection. The network focuses on the important spatial position and channel information of small targets by acquiring and utilizing the global context information of images, and enhances the feature representation of small targets, thus improving the detection performance.

2.2.3. Pedestrian Recognition. Shopovska et al. [35] present a learning-based fusion method to enhance pedestrian visibility in variable conditions (day and night).

2.2.4. Semantic Segmentation. The cascade of the ResNet and improved CRFs are used to construct the semantic segmentation module for the aluminum electrolyte image in [36]. Hou et al. [37] proposed a semantic segmentation strategy using infrared and visible image fusion method based on GANs. Xu et al. present an AFNet based on deep learning, which effectively improves the accuracy of multispectral image semantic segmentation [38]. Recently, Zhao et al. [39] proposed a correlation-driven feature decomposition fusion network, which utilizes various modules to extract the high-frequency and low-frequency features of an image.

In summary, in terms of current fusion performance, image fusion methods based on deep learning generally outperform traditional methods. In practical applications, different model architectures should be designed in combination with specific image fusion task drivers to improve the advanced visual application of fusion images in real scenes.

3. Proposed Method

In this section, we proposed a fusion algorithm to enrich image information and make detection more accurate for intelligent safety monitoring on power grid construction sites. Firstly, we introduce an overview of the proposed fusion model, a feature fusion attention network based on generative adversarial network (FFA-GAN). Then, we introduce the shared convolution group (SCG) module, the channel attention (CA) module, and the pixel attention (PA) module in order, which are designed in sequence for the generator to deal with multimodal features flexibly. Finally, we describe the discriminator dehaze (DDE) and discriminator fusion (DFU) designed for the two-stage discriminator, in which they jointly guarantee the FFA-GAN to achieve good performance on the infrared and visible image fusion task.

3.1. Overview. In the field of image fusion, the generative adversarial network (GAN)-based methods are usually used as representative baselines, especially for infrared and visible image fusion tasks. The characteristic of GAN-based fusion methods is the fused image containing abundant information, while retaining structural similarity between fused image and source images. Thus, the GAN has great potential to achieve success in the area of image fusion. Inspired by the framework of the existing GAN-based image fusion algorithms, we have designed a feature fusion attention network based on the generative adversarial network, as shown in Figure 1, which has been abbreviated as FFA-GAN hereinafter.

As illustrated in Figure 1, the FFA-GAN consists of a generator network and two discriminator networks. In this work, the generator is designed to constantly explore the feature fusion mapping function between infrared and visible images to obtain fused images. Moreover, the two-stage discriminator is designed in the FFA-GAN to provide two constraints for the generator to get clean fused images with both infrared and visible information. Specially, the DDE is used to recognize whether images are clean while the DFU is used to identify the proportion of infrared and visible information in fused images. During the testing stage, since the generator has learned to fuse images, the discriminator is not required to provide constrains and only the generator is needed to get the fusion image. Then, we describe the architecture of the generator and discriminators in detail.

The fused images generated by the designed generator network are used to fool the discriminator. The discriminator network consists of three convolutional layers and three fully connected layers to classify the input image. In this paper, two discriminators are designed in the FFA-GAN to identify haze images and normal images, respectively. The reason lies on to avoid information loss caused by single countermeasure architecture when dealing with fog images and normal images. At the same time, it forces the generated image to retain more meaningful information from the source image.

3.1.1. Generator. Firstly, the infrared and visible features are extracted through two convolution layers. Secondly, these features are fed into three SCG modules to capture modality-common features. Then, the output of each group by channel connection (CC) is integrated. In order to select and reweight significant infrared and visible features, we adopt two CA modules in dual branches of the generator. After that, a PA module will be used to achieve fine-grained modality fusion which aimed to blend dual-branch features. Finally, a 3×3 convolution layer and a 1×1 convolution layer are used to map the fusion features to the two-dimensional plane, and the fusion results are obtained. Note that, merging together CA modules and PA module constitute a special structure, which can be helpful to handle source images with complex distribution, like uneven haze distribution.

3.1.2. Discriminator. The DDE which is designed in the FFA-GAN is used to identify whether the input image is a fuzzy image or a clean image to avoid the loss of information caused by a single game architecture while processing fog images and clean images. At the same time, the DFU forces the generated results to retain more meaningful information from the source image by balancing the proportion of information between infrared and visible images. The structure of the DDE and DFU is similar, both consist of three convolution layers and three fully connected layers. The DDE aims to obtain a one-dimensional class vector, while the DFU tries to obtain a two-dimensional proportion vector.

3.2. Designments in Generator

3.2.1. SCG Module. The detailed structure of the SCG module is shown in Figure 2. There are N contiguous convolutional blocks (CBs), represented by grey squares, which help to increase depth and expressiveness of the FFA-GAN. The detailed structure of the CB is shown at the bottom of Figure 2. The CB consists of skip feature residual connections and cascaded CA and PA modules. These skip feature residual connections are designed to reduce information loss and get around training difficulty. And the cascaded CA and PA modules are used to select more significant features. The key of the generator to obtain effective fused images is to select and reweight multimodal features is CA and PA modules introduced next.

3.2.2. CA Module. Most existing deep learning-based fusion strategies for combining features simply integrate them equally through channel connections, without considering the varying importance of different feature channels. However, as the fusion network becomes deeper, it is likely that only a small subset of features will respond meaningfully. To address this issue, we propose using CA modules to assign appropriate weights to different features, based on their similarity relationships across channels. The structure of the CA module is illustrated in Figure 3. To obtain channel-level global information of input feature map which be denoted as X, we first apply global average pooling. This operation calculates the average value of each channel feature. Specifically, it can be expressed as follows:

$$\mathrm{GAP}\left(X^c\right) = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{i=1}^{W} X_{i,j}^c, \tag{1}$$

where X^c denotes the c_{th} channel feature, and i and j represent the coordinate information of the feature value. H and W are height and width of feature maps, respectively. Then, the compressed global feature weights obtained by the global average pooling operation are transmitted to a 3×3 convolution layer, followed by a ReLU activation layer, and another 3×3 convolution layer. These operations help to refine and enhance the global features. Finally, a sigmoid

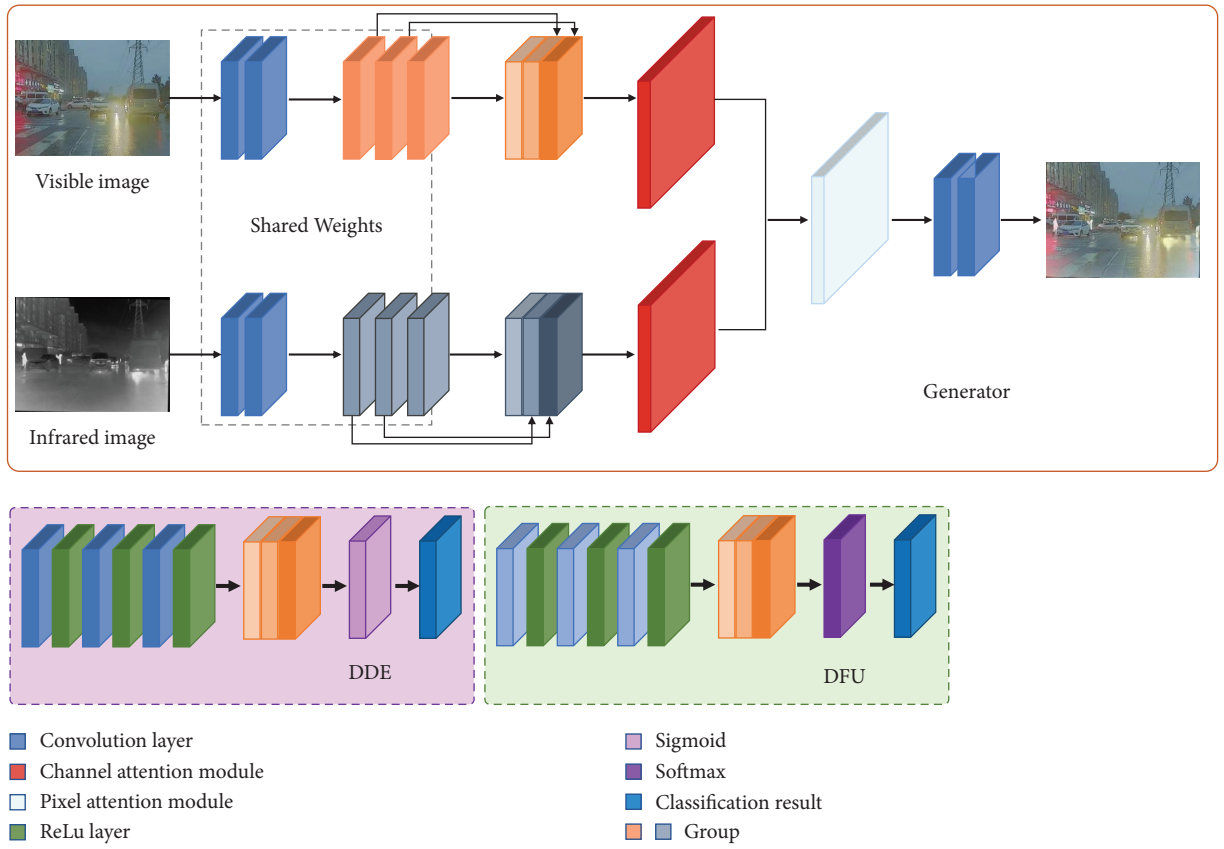

Figure 1: The architecture of the proposed FFA-GAN contains a generator with the shared convolution groups (SCGs) and the pixel attention module (PA), channel attention module (CA), and the two-category discriminator (i.e., discriminator dehaze and discriminator fusion).

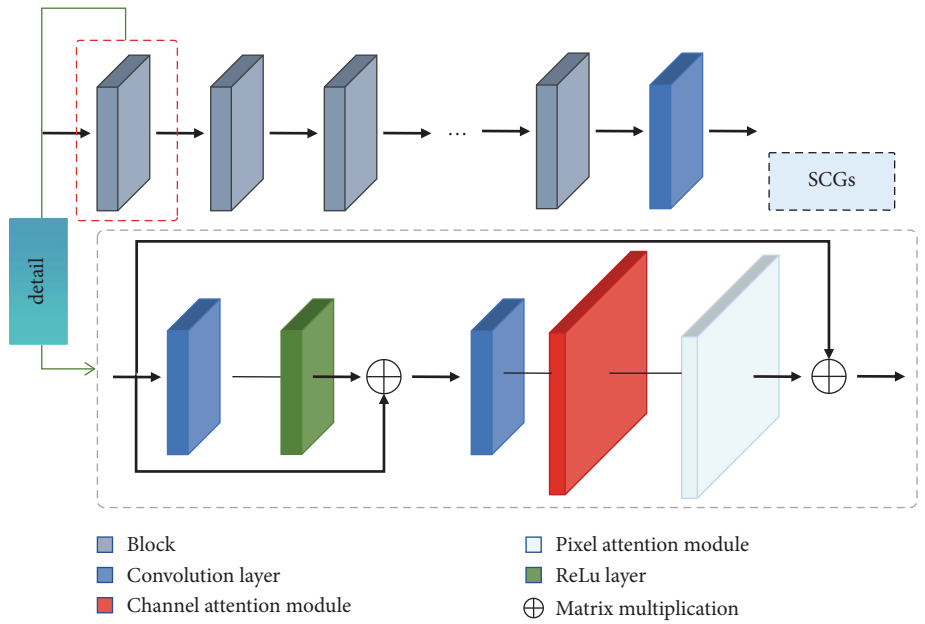

Figure 2: The structure of the shared convolution groups (SCGs).

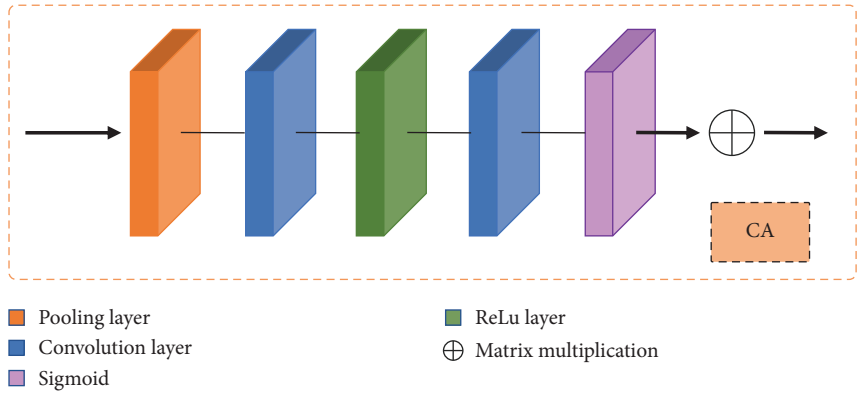

Figure 3: The structure of the channel attention module (CA).

activation layer is used to obtain the channel weight, which is then applied to reweight the input source features X. This enables the FFA-GAN to focus on the most meaningful and relevant features, which helps to improve its fusion performance. We can express this whole process as follows:

$$\mathrm{CA}(X) = \mathrm{Sigmoid}\left(\mathrm{Conv}\left(\mathrm{ReLU}\left(\mathrm{Conv}\left(\sum_{c=1}^{C}\mathrm{GAP}(X^c)\right)\right)\right)\right)\cdot X, \tag{2}$$

where Conv(·) denotes a 3×3 convolution operation. ReLU(·) and Sigmoid(·) denote the ReLU and sigmoid activation functions.

3.2.3. PA Module. Due to the uneven distribution of haze on different pixels in hazed images, it is necessary to use pixel attention to focus on the features of each individual pixel. To refine pixel feature fusion and reduce interference affected by the haze, we consider employing the PA module, which is illustrated in Figure 4. Unlike the channel CA module, the PA module includes a 3×3 convolution layer, a ReLU activation layer, and another 3×3 convolution layer that work together to refine and enhance the features for each pixel not channel. Then, a sigmoid activation layer is used to weight the feature weight of each pixel based on its importance. These weights are then applied to the input features to obtain the final output. We can express the complete PA module as follows:

$$\mathrm{PA}(X)\mathrm{Sigmoid}\left(\mathrm{Conv}\left(\mathrm{ReLU}\left(\mathrm{Conv}\left(\sum_{c=1}^{C}\mathrm{GAP}(X^c)\right)\right)\right)\right)\cdot X. \tag{3}$$

3.3. Two-Stage Discriminator. The proposed FFA-GAN incorporates a two-stage discriminator composed of a discriminator dehaze (DDE) network and a discriminator fusion (DDF) network, as illustrated in Figure 1.

3.3.1. DDE Network. The DDE network serves as a simple image classifier distinguishing whether input images are hazed or clean. The DDE network can take the fusion result

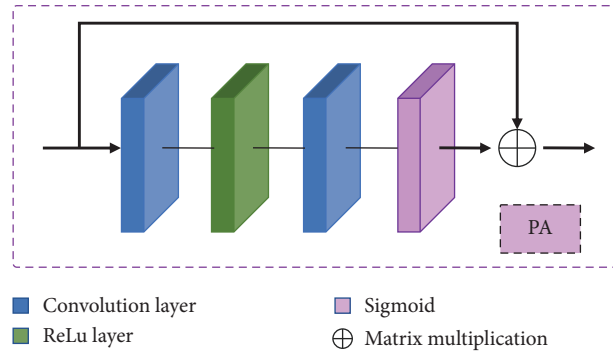

Figure 4: The structure of the pixel attention module (PA).

of the proposed generator or the source image as input. As shown in Figure 1, the input of the DDF network is subjected to three 3×3 convolution operations and three ReLU activations before being processed through three fully connected layers. Finally, a sigmoid activation layer is used to obtain the probability that the image is a clean image, which produces a one-dimensional class vector.

3.3.2. DFU Network. The structure of the DFU network is similar to the DDE network. There are also three 3×3 convolution layers, three ReLU combinations, and three fully connected layers. However, unlike the DDE network, the DFU network uses a Softmax activation layer to obtain the proportion of infrared and visible image features, producing a two-dimensional class vector.

3.4. Loss Function. In order to obtain desired fusion results for our proposed FFA-GAN, we will describe the loss function in detail from two parts: generator loss and discriminator loss in the next.

3.4.1. Generator Loss. The generator loss is defined as the distance between the fused results and the desired results. This can be measured using image pixel loss, image gradient loss, perceptual loss, and adversary loss. The formula is as follows:

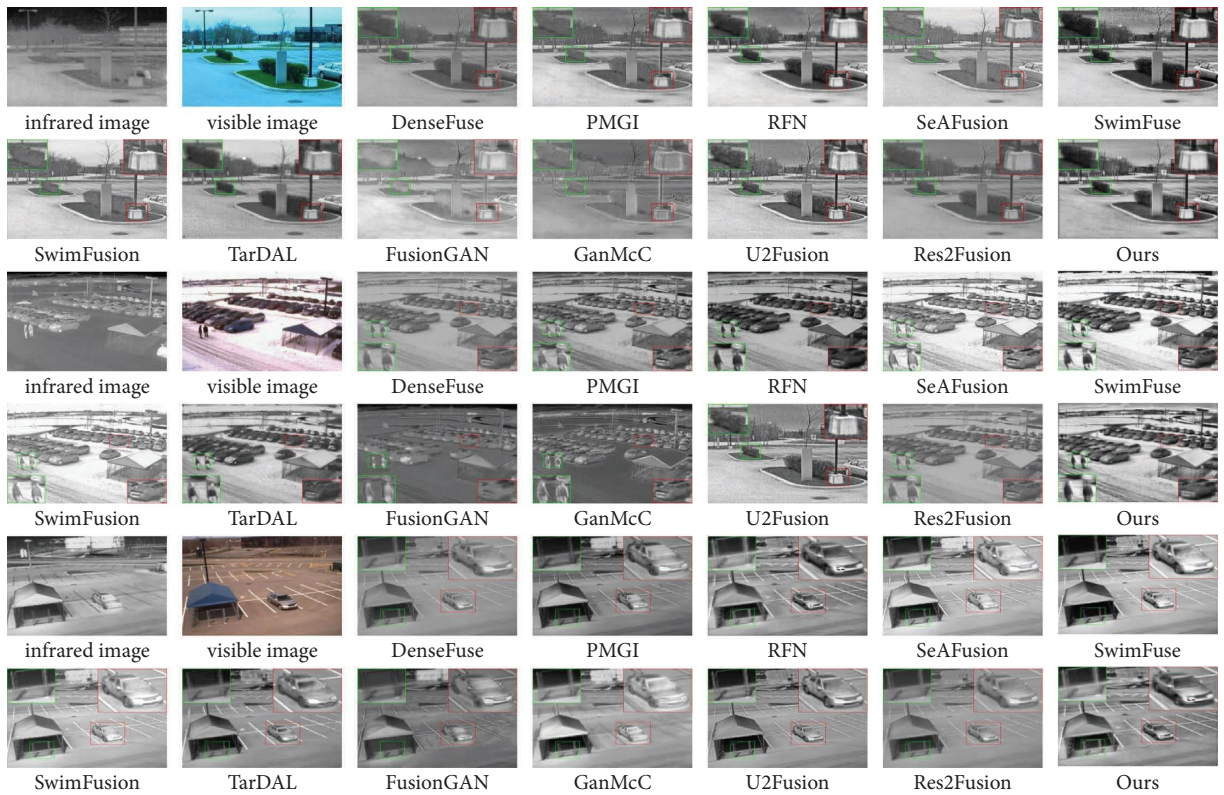

FIGURE 5: Comparison experiment on the INO dataset. Areas with large differences are highlighted by red and green boxes, and enlarged images of red boxes are in the lower right or left corner.

$$L_{\text{Generator}} = L_{\text{pixel}} + L_{\text{gradient}} + L_{\text{perceptual}} + L_{\text{adversary}}. \tag{4}$$

The image pixel loss is designed to maintain suitable pixels from infrared and visible images through Euclidean distance. It can be represented as follows:

$$L_{\text{pixel}} = \frac{1}{H \cdot W}\left(\left\|I_{\text{fused}} - I_{\text{vi}}\right\| + \left\|I_{\text{fused}} - I_{\text{ir}}\right\|\right), \tag{5}$$

where I_{fused}, I_{vi}, and I_{ir} are fused image, visible image, and infrared image; H and W are height and width of images; and $\|\cdot\|$ indicates the adoption of L2 normal.

The image gradient loss is proposed to calculate image gradient to preserve texture information. It can be expressed as follows:

$$L_{\text{gredient}} = \frac{1}{H \cdot W}\left(\left\|\nabla I_{\text{fused}} - \nabla I_{\text{vi}}\right\| + \left\|\nabla I_{\text{fused}} - \nabla I_{\text{ir}}\right\|\right), \tag{6}$$

where ∇ denotes the function to calculate image gradient through Laplace operator. The image perceptual loss is proposed to calculate the distance of image features through VGG. It can be expressed as follows:

$$L_{\text{perceptual}} = \frac{1}{H \cdot W}\left(\left\|\text{VGG}\left(I_{\text{fused}}\right) - \text{VGG}\left(I_{\text{vi}}\right)\right\| + \left\|\text{VGG}\left(I_{\text{fused}}\right) - \text{VGG}\left(I_{\text{ir}}\right)\right\|\right), \tag{7}$$

where VGG(·) denotes the function to get image feature maps through VGG. The adversary loss is the key for GAN-based fusion methods, which can be denoted as follows:

$$\begin{aligned} L_{\text{eqadversary}} = \alpha\Big[&\left(\text{DDE}\left(I_{\text{fused}}\right) - 1\right)^2 - \text{DDE}\left(I_{\text{vi}}\right)\Big)^2 + \left(\text{DDE}\left(I_{\overline{vi}}\right) - 1\right)^2\Big)\Big] \\ &+ \Big(\left|DFU\left(I_{\text{fused}}\right) - V_{\text{fused}}\right| + \left|\text{DFU}\left(I_{\text{vi}}\right) - V_{\text{vi}}\right| + \left|\text{DFU}\left(I_{\text{ir}}\right) - V_{\text{ir}}\right|\Big), \end{aligned} \tag{8}$$

Figure 6: Comparison experiment on the M3FD dataset. Areas with large differences are highlighted by red and green boxes, and enlarged images of red boxes are in the lower right or left corner.

where $|\cdot|$ denotes the function to calculate the vector length. In addition, $\bar{I}_{vi}$ are clean visible images.

3.4.2. Discriminator Loss. The discriminator loss consists of DDE loss and DFU loss, which can be represented as follows:

$$L_{\text{Discriminator}} = L_{\text{DDE}} + L_{\text{DFU}}. \tag{9}$$

The DDE loss is designed to guide the DDE network in identifying whether input images are clean or hazy. Hence, there is a label of 0 or 1 to represent the hazy image or the clean image. It can be expressed as follows:

$$L_{\text{DDE}} = \left(\text{DDE}\left(I_{\text{fused}}\right) - 0\right)^2 + \text{DDE}\left(I_{\text{vi}}\right)^2 + \left(\text{DDE}\left(\bar{I}_{\text{vi}}\right) - 1\right)^2. \tag{10}$$

The DFU loss is introduced to measure the proportion of infrared and visible information. It can be expressed as follows:

$$L_{\text{DFU}} = \left|\text{DFU}\left(I_{\text{fused}}\right) - V_{\text{fused}}\right| + \left|\left(\text{DFU}\left(I_{\text{vi}}\right) - V_{\text{vi}}\right|\right| + \left|\left(\text{DFU}\left(I_{\text{ir}}\right) - V_{\text{ir}}\right|\right|. \tag{11}$$

4. Experiment

In this section, we first provide an overview of the dataset used in our training and testing process. Then, we briefly introduce the fusion metrics used in our experiments and compare them with 11 state-of-the-art fusion methods. Then, we present extensive experiments to demonstrate the rationality and superiority of our method. Finally, we analyze the results of our method from both qualitative and quantitative perspectives. It is worth noting that only partial results are given due to the page limits.

4.1. Dataset and the Implementation Details

4.1.1. Dataset. We use two datasets to conduct all experiments. First, the INO dataset is the largest center of expertise in optics and photonics in Canada. They collected many mixed long-wave infrared videos and color visible videos as part of the INO dataset (these data can be found in https://www.ino.ca/en/technologies/video-analyticsdataset/videos/). In addition, there are ground truths of corresponding objects. Second, the M3FD dataset is designed by Liu et al. [40] to consider the inconsistency of image properties and features

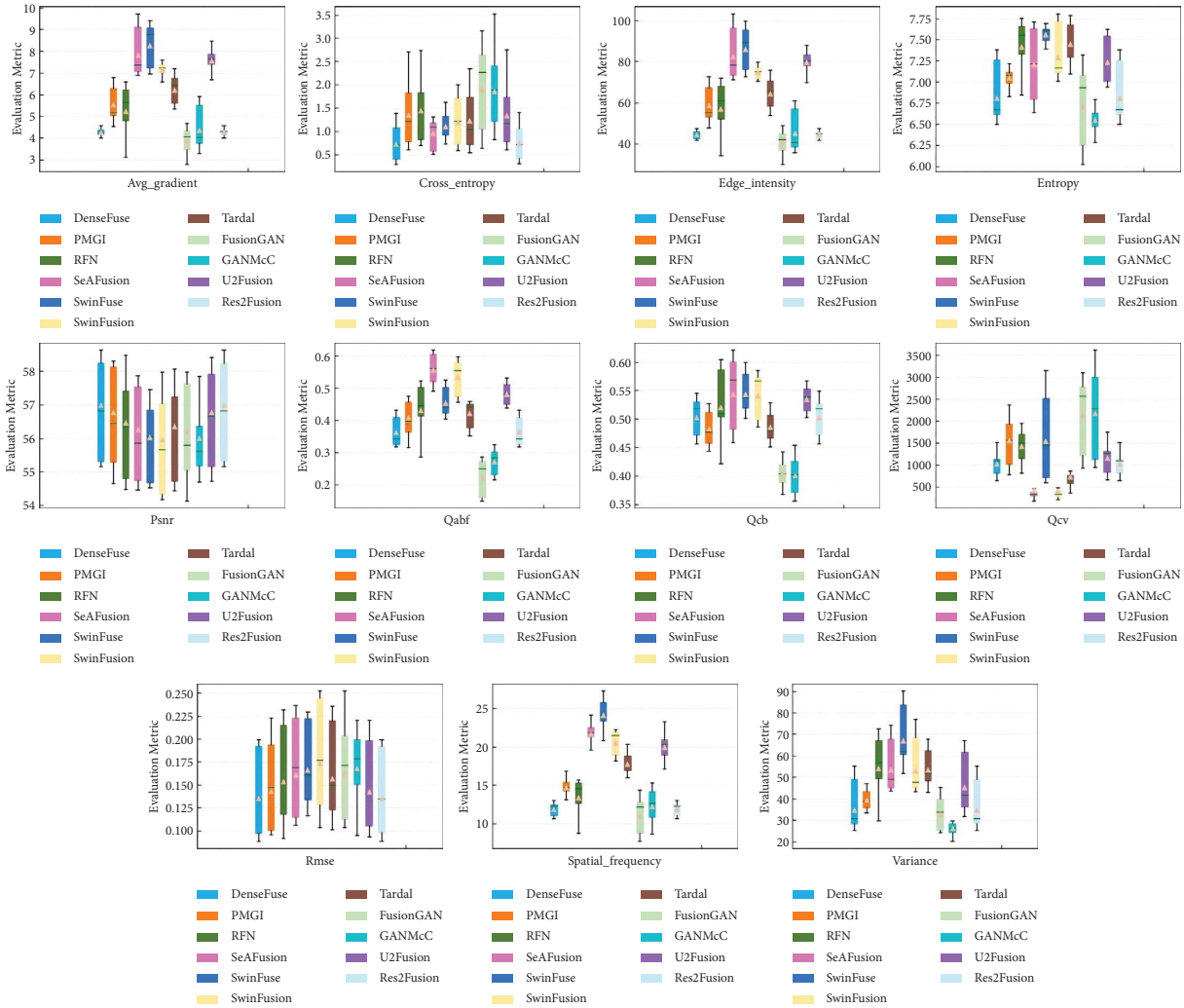

FIGURE 7: Quantitative analysis of 13 evaluation indexes on 133 image pairs from the INO dataset. High indexes of AG, EI, EN, MI, PSNR, QAB/F, Qcb, SF, SSIM, and SD represent better fusion performance; low indexes of CE, Qcv, and RMSE represent better fusion performance.

presented by infrared and visible images under different scenes. They collected high-quality infrared and visible image pairs of campus, tourist resort, urban road, and other scenes. Furthermore, there are six tagged targets including people, cars, buses, motorcycles, lights, and trucks (these data can be found at https://github.com/JinyuanLiu-CV/TarDAL).

4.1.2. Experimental Settings. Our work exploits the mapping function between source infrared and visible images through the FFA-GAN. During the training phase, we use the Adam algorithm to guide minimizing the generator loss and discriminator. The learning rate is set to 0.0001. As we employ data augmentation by cutting source images into patches, we select 30 strictly aligned infrared and visible image pairs in the M3FD dataset. While more image pairs could be used, there are enough infrared and visible image patch pairs to make the proposed algorithm effective. All experiments are conducted on a laptop with a 3.60 GHz 11th i7-11700K CPU and GeForce RTX 3090. The code is implemented with Python and MATLAB. We compare the proposed method with 11 state-of-the-art fusion methods, including DenseFuse [41], PMGI, RFN [42], SeAFusion [43], SwinFuse, SwinFusion [44], TarDAL [40], FusionGAN [20], GANMcC [45], U2Fusion [46], and Rse2Fusion. These are implemented based on available codes.

4.1.3. Fusion Metrics. We choose average gradient (AG), cross entropy (CE), edge intensity (EI), entropy (EN), mutual information (MI), peak signal-to-noise ratio (PSNR), QAB/F, Qcb, Qcv, root mean squared error (RMSE), spatial frequency (SF), structural similarity index measure (SSIM), and standard deviation (SD) as fusion metrics based on [47–49]. In all experiments, we use an up arrow or a down arrow to indicate that the higher or lower the indicator, the better the fusion.

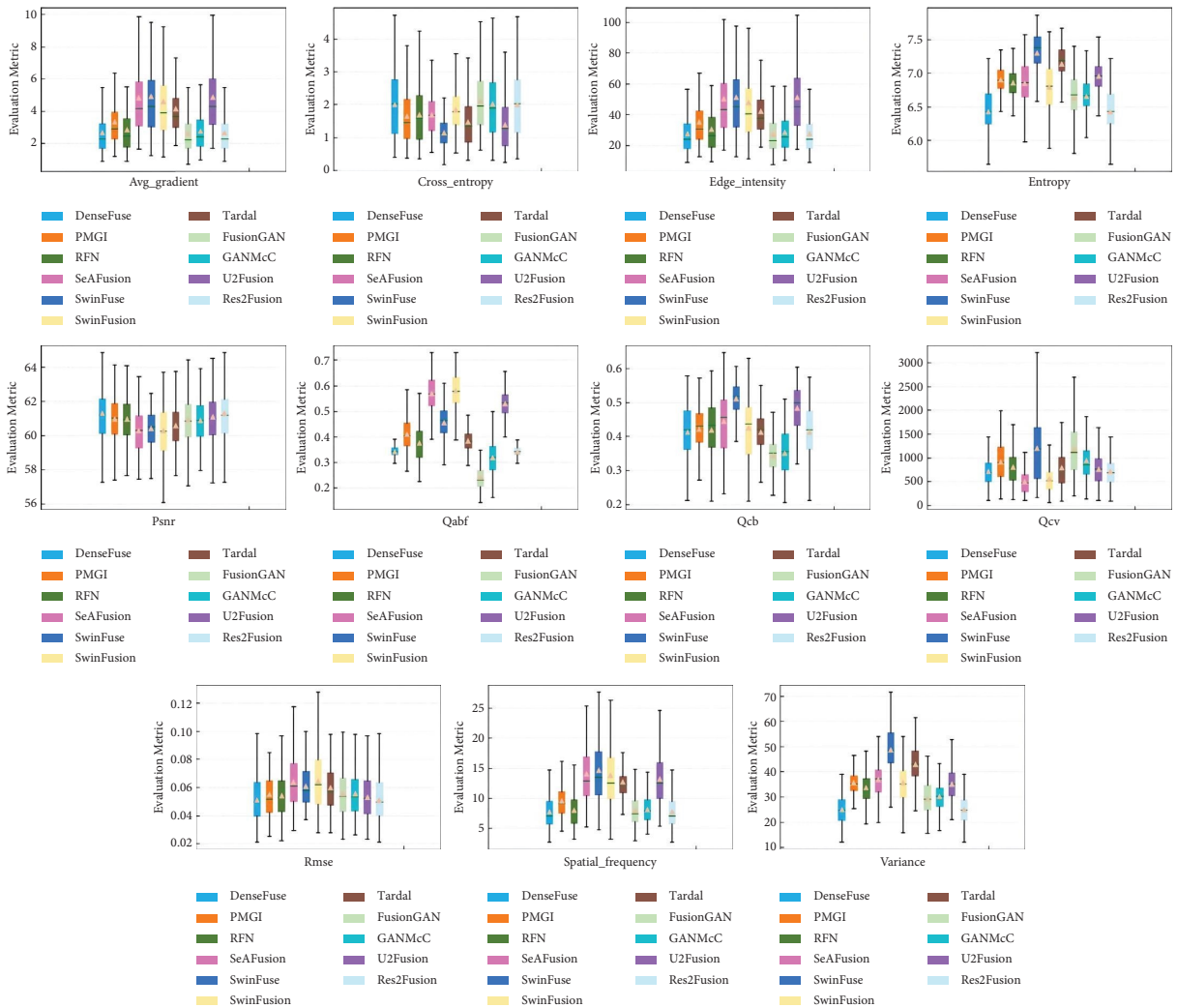

FIGURE 8: Quantitative analysis of 13 evaluation indexes on 300 image pairs from the M3FD dataset. High indexes of AG, EI, EN, MI, PSNR, QAB/F, Qcb, SF, SSIM, and SD represent better fusion performance; low indexes of CE, Qcv, and RMSE represent better fusion performance.

4.2. Comparison with State-of-the-Art Methods. In order to convincingly verify the performance of our fusion method, we compare the FFA-GAN with other 11 fusion methods on the public benchmarks INO dataset and the M3FD dataset. The qualitative and quantitative analysis of the fusion results of different methods are presented below.

4.2.1. Qualitative Comparisons. In Figure 5, we first present the qualitative visualization of our FFA-GAN on the INO dataset. We provide 3 sets of representative infrared and visible images and their experimental results are shown in Figure 5. In each set, the first two subfigures present the infrared-visible image pair, the third to thirteenth subfigures show the results of the advanced fusion models mentioned above, and the last subfigure presents the fused image of our method. The meaningful region is enlarged and marked with red boxes in each fused result. We can see that, the results of our method is more clear and contain more texture and contour information from the source images, which will be advantageous for advanced visual tasks, such as object detection in power grid.

Another qualitative analysis in this part is carried out on the 3 groups of M3FD dataset in Figure 6. Similar to Figure 5, the first two subfigures, the third to thirteenth subfigures, and the last one, respectively, present the infrared-visible image pair, the results of the comparison models, and the fused image of our method in each set. It can be clearly observed in Figure 6, our method is able to preserve rich texture information, scene information, and unique contrast information. In contrast, the target in the fusion result lacks clarity and the background is blurred, indicating that the target region in the infrared image and the typical features in the visible image, such as license plate information and human information, are not well preserved. It is worth emphasizing that our method is highly robust in the presence of strong light interference during the night.

Table 1: The ranking of thirteen evaluation indexes by different methods on INO dataset and the sort after a sorting rule.

Model	AG↑	CE↓	EI↑	EN↑	MI↑	PSNR↑	QAB/F↑	Qcb↑	Qcv↓	RMSE↓	SF↑	SSIM↑	SD↑
Dense	10	**1**	11	10	7	**1**	10	9	6	**1**	10	**1**	10
PMGI	7	9	7	8	5	4	6	7	9	4	7	***3***	8
RFN_Nest	8	10	8	4	6	5	7	6	8	5	8	6	6
SeAFusion	***3***	9	***3***	4	***3***	9	**1**	*2*	*2*	8	***3***	7	***3***
SwinFuse	*2*	5	*2*	**1**	4	11	4	**1**	***3***	11	*2*	10	**1**
SwinFusion	5	7	5	5	*2*	12	*2*	***3***	**1**	12	4	4	5
Tardal	6	6	6	***3***	**1**	6	5	10	10	***3***	6	8	4
FusionGAN	12	12	12	11	11	10	12	11	11	10	12	11	11
GANmcc	9	11	9	12	10	8	11	12	12	9	9	9	12
U2Fusion	4	8	4	7	9	***3***	***3***	4	7	***3***	5	5	7
Res2Fusion	11	*2*	10	9	8	*2*	9	8	5	*2*	11	*2*	9
Ours	**1**	***3***	**1**	*2*	12	7	8	5	4	7	**1**	12	*2*

Bold values indicate the best result, italics values represent the second best result, and bold with italics values represent the third best result.

Table 2: The ranking of thirteen evaluation indexes by different methods on M3FD dataset and the sort after a sorting rule.

Model	AG↑	CE↓	EI↑	EN↑	MI↑	PSNR↑	QAB/F↑	Qcb↑	Qcv↓	RMSE↓	SF↑	SSIM↑	SD↑
Dense	11	4	11	9	***3***	**1**	10	6	7	**1**	11	**2**	9
PMGI	5	10	6	4	7	11	4	9	***3***	11	6	7	7
RFN_Nest	8	6	8	***3***	6	4	5	4	6	4	8	6	6
SeAFusion	*2*	7	*2*	5	*2*	6	**1**	5	11	5	*2*	5	5
SwinFuse	6	8	5	10	8	9	8	**1**	10	8	***3***	12	*2*
SwinFusion	***3***	5	***3***	6	**1**	7	*2*	8	12	7	4	4	4
Tardal	4	*11*	4	*2*	5	10	6	10	9	9	5	8	***3***
FusionGAN	12	12	12	12	10	5	12	12	*2*	6	12	9	12
GANmcc	9	9	9	7	11	8	11	11	**1**	10	9	10	11
U2Fusion	7	*2*	7	11	9	***3***	***3***	***3***	5	***3***	7	***3***	10
Res2Fusion	10	***3***	10	8	4	*2*	9	7	8	*2*	10	**1**	8
Ours	**1**	**1**	**1**	**1**	12	12	7	*2*	4	12	**1**	11	**1**

Bold values indicate the best result, italics values represent the second best result, and bold with italics values represent the third best result.

4.2.2. Quantitative Analysis. Afterward, 13 metrics mentioned above on the INO and M3FD dataset (100 image pairs) are employed to quantitatively compare the abovementioned results, which are displayed in Figures 7 and 8. Due to space constraints, we have not provided a detailed introduction of each evaluation indicator in this paper. Interested readers can refer to [47, 50], and [51] for more information about these indicators. To quantify the comparison results, we introduce a ranking rule of ranking the average of 13 evaluation indexes, as shown in Tables 1 and 2.

In a comprehensive perspective, our method obtains the satisfactory performance in AG, CE, EI, EN, SF, and SD on the INO dataset (as shown in Table 1). Besides, we obtain the best performance in AG, CE, EI, EN, SF, and SD on the M3FD dataset (as shown in Table 2). These are indicators based on the feature of the image, indicating that our fused images are informative and more consistent with the human visual system. In addition, we have also found that the Dense, U2Fusion, and Res2Fusion models have achieved good results in the PSNR, RMSE, and SSIM indicators. The likely reason is that the authors of these algorithms were more focused on specific pieces of information during their design. This phenomenon further suggests that an image fusion algorithm should be evaluated using a variety of indicators for comprehensive comparison, which demonstrates the benefits of our FFA-GAN. Unfortunately, our method did not perform well on the abovementioned PSNR, SSIM, and RMSE indices. The primary reason for this outcome is the lack of a ground truth for fusing infrared and visible light images. In particular, due to noise and other factors, the fusion of infrared and visible light may result in inaccuracies in the values of these three indicators, as they are compared to reference images. Overall, our FFA-GAN stably retains rich useful information from source images,

and can describe the scene information of the whole image, especially visible images are contaminated (such as strong light and haze).

5. Conclusions

In this work, we have proposed a new generative adversarial network called the FFA-GAN, which is based on feature fusion attention. We have applied this network to power grid security management. The key design in our approach is a shared convolution group (SCG) in the dual-branch generator network. These are designed to extract modality-common features from source images. To handle multimodality information flexibly, each SCG contains both channel attention modules and pixel attention modules.

We have also incorporated infrared and visible image features into our network, using a CA and PA combination structure to fuse these features. Our two-stage discriminator includes both DDE and DFU to ensure that the proposed FFA-GAN achieves good performance in infrared and visible image fusion tasks. Experimental results demonstrate that our fusion network performs well. It can be embedded in the grid AI platform to provide services for related applications and provides a strong guarantee for power grid safety.

However, there are some limitations to our approach due to the lack of aligned infrared and visible data with haze. In our experiments using the M3FD dataset, we used dark channel prior to remove image haze. Although this approach can effectively remove haze, the image may still be affected by the distribution of the haze. Therefore, improving the performance of our proposed method further is subject to overcome these limitations.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the science and technology project of China Southern Power Grid Co., Ltd. under Grant YNKJXM20220142.

References

[1] Y. Guo, H. Yuan, Y. Zhuang, J. Xu, and Y. Zhang, "Research on informatization construction of digital substation in smart grid," in *Proceedings of the IOP Conference Series: Earth and Environmental Science*, p. 42089, Kamakura City, Japan, June 2021.

[2] H. Jiangtao, "Discussion on the construction of substation security video surveillance system," *IOP Conference Series: Materials Science and Engineering*, vol. 563, no. 3, p. 32004, 2019.

[3] G. Peng, Y. Lei, H. Li, D. Wu, J. Wang, and F. Liu, "CORY-Net: Contrastive res-YOLOv5 network for intelligent safety monitoring on power grid construction sites," *IEEE Access*, vol. 9, pp. 160461–160470, 2021.

[4] S. Ren, K. He, R. Girshick, J. Sun, and R.-C. N. N. Faster, "Towards realtime object detection with region proposal networks," 2015, https://arxiv.org/abs/1506.01497.

[5] J. Redmon, S. Divvala, R. Girshick, and F. Ali, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, Honolulu, HI, USA, June 2016.

[6] Q. Fang, H. Li, X. Luo et al., "Detecting non-hardhat-use by a deep learning method from far-field surveillance videos," *Automation in Construction*, vol. 85, pp. 1–9, 2018.

[7] S. Tang, D. Roberts, and M. Golparvar-Fard, "Human-object interaction recognition for automatic construction site safety inspection," *Automation in Construction*, vol. 120, Article ID 103356, 2020.

[8] L. Ding, W. Fang, H. Luo, P. E. D. Love, B. Zhong, and X. Ouyang, "A deep hybrid learning model to detect unsafe behavior: integrating convolution neural networks and long short-term memory," *Automation in Construction*, vol. 86, pp. 118–124, 2018.

[9] Q. Fang, H. Li, X. Luo et al., "A deep learning-based method for detecting non-certified work on construction sites," *Advanced Engineering Informatics*, vol. 35, pp. 56–68, 2018.

[10] J. Heo, S. Kong, B. Abidi, and M. Abid, "Fusion of visual and thermal signatures with eyeglass removal for robust face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 366–373, Washington, DC, USA, June 2004.

[11] X. Zhang, P. Ye, H. Leung, K. Gong, and G. Xiao, "Object fusion tracking based on visible and infrared images: a comprehensive review," *Information Fusion*, vol. 63, pp. 166–187, 2020.

[12] L. Wang, Y. Shoulin, H. Alyami et al., "A novel deep learning-based single shot multibox detector model for object detection in optical remote sensing images," *Geoscience Data Journal*, vol. 18, 2022.

[13] Z. Duan, J. Lan, Y. Xu, B. Ni, L. Zhuang, and X. Yang, "Pedestrian detection via bi-directional multi-scale analysis," in *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 1023–1031, Mountain View, CA, USA, March 2017.

[14] 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.

[15] G. Li and W. Liu, "Multimedia data processing technology and application based on deep learning," *Advances in Multimedia*, vol. 2023, Article ID 4184425, 15 pages, 2023.

[16] M. Pu, Y. Huang, Q. Guan, and Q. Zou, "Graphnet: learning image pseudo annotations for weakly-supervised semantic segmentation," in *Proceedings of the 26th ACM International Conference on Multimedia*, pp. 834–848, Virtual Event China, October 2018.

[17] J. Wu, N. Cai, W. Chen, H. Wang, and G. Wang, "Automatic detection of hardhats worn by construction personnel: a deep learning approach and benchmark dataset," *Automation in Construction*, vol. 106, Article ID 102894, 2019.

[18] L. Huang, Q. Fu, M. He, D. Jiang, and Z. Hao, "Detection algorithm of safety helmet wearing based on deep learning," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 13, 2021.

[19] M. Cao, K. Cao, B. Wu, and M. Tan, "Intelligent condition monitoring and management for power transmission and distribution equipments in Yunnan Power Grid," in *Proceedings of the International Conference on High Voltage Engineering and Application*, pp. 8–11, Beijing, China, September 2012.

[20] 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, 2019.

[21] H. Son, H. Choi, H. Seong, and C. Kim, "Detection of construction workers under varying poses and changing background in image sequences via very deep residual networks," *Automation in Construction*, vol. 99, pp. 27–38, 2019.

[22] W. Fang, L. Ding, B. Zhong, P. E. D. Love, and H. Luo, "Automated detection of workers and heavy equipment on construction sites: a convolutional neural network approach," *Advanced Engineering Informatics*, vol. 37, pp. 139–149, 2018.

[23] Z. Qiu, X. Zhu, C. Liao, D. Shi, and W. Qu, "Detection of transmission line insulator defects based on an improved lightweight YOLOv4 model," *Applied Sciences*, vol. 12, no. 3, p. 1207, 2022.

[24] S. Tan, G. Lu, Z. Jiang, and L. Huang, "Improved YOLOv5 network model and application in safety helmet detection," in *Proceedings of the IEEE International Conference on Intelligence and Safety for Robotics*, pp. 330–333, Kyoto, Japan, October 2021.

[25] K. Shahid, A. Qadir, U. Farooq, M. Shakir, and A. A. Laghari, "Hyperspectral imaging: a review and trends towards medical imaging," *Current Medical Imaging*, vol. 45, Article ID 35598236, 2022.

[26] S. Karim, G. Tong, J. Li, A. Qadir, U. Farooq, and Y. Yu, "Current advances and future perspectives of image fusion: a comprehensive review," *Information Fusion*, vol. 90, pp. 185–217, 2023.

[27] A. A. Laghari, V. E. Vania, and S. Yin, "How to collect and interpret medical pictures captured in highly challenging environments that range from nanoscale to hyperspectral imaging," *Current Medical Imaging*, vol. 54, Article ID 36582065, 2022.

[28] M. Han, K. Yu, J. Qiu et al., "Boosting target-level infrared and visible image fusion with regional information coordination," *Information Fusion*, vol. 92, pp. 268–288, 2023.

[29] S. Singh, A. Gyaourova, G. Bebis, and I. Pavlidis, *SPIE Proceedings*, vol. 5404, 2004.

[30] J. Han and B. Bhanu, "Fusion of color and infrared video for moving human detection," *Pattern Recognition*, vol. 40, no. 6, pp. 1771–1784, 2007.

[31] I. Ulusoy and H. Yuruk, "New method for the fusion of complementary information from infrared and visual images for object detection," *IET Image Processing*, vol. 5, no. 1, pp. 36–48, 2011.

[32] S. Gao, Y. Cheng, and Y. Zhao, "Method of visual and infrared fusion for moving object detection," *Optics Letters*, vol. 38, no. 11, pp. 1981–1983, 2013.

[33] J. Ma, L. Tang, M. Xu, H. Zhang, and G. Xiao, "STDFusionNet: an infrared and visible image fusion network based on salient target detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–13, 2021.

[34] Z. Zuo, X. Tong, J. Wei et al., "AFFPN: attention fusion feature pyramid network for small infrared target detection," *Remote Sensing*, vol. 14, no. 14, p. 3412, 2022.

[35] I. Shopovska, L. Jovanov, and W. Philips, "Deep visible and thermal image fusion for enhanced pedestrian visibility," *Sensors*, vol. 19, no. 17, p. 3727, 2019.

[36] Z. Xu, J. Wang, and L. Wang, "Infrared image semantic segmentation based on improved deepLab and residual network," in *Proceedings of the 10th International Conference on Modelling, Identification and Control*, pp. 1–9, Guiyang, China, July 2018.

[37] 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, p. 376, 2021.

[38] J. Xu, K. Lu, H. Wang, and H. Wang, "Attention fusion network for multi-spectral semantic segmentation," *Pattern Recognition Letters*, vol. 146, pp. 179–184, 2021.

[39] Z. Zhao, H. Bai, J. Zhang et al., "CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion," 2023, https://arxiv.org/abs/2211.14461.

[40] J. Liu, X. Fan, Z. Huang et al., "Target-aware dual adversarial learning and a multi-scenario multimodality benchmark to fuse infrared and visible for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5802–5811, New Orleans, LU, USA, June 2022.

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

[42] H. Li, X. Wu, and J. Kittler, "RFN-Nest: an end-to-end residual fusion network for infrared and visible images," *Information Fusion*, vol. 73, pp. 72–86, 2021.

[43] L. Tang, J. Yuan, and J. Ma, "Image fusion in the loop of high-level vision tasks: a semantic-aware real-time infrared and visible image fusion network," *Information Fusion*, vol. 82, pp. 28–42, 2022.

[44] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "Swinfusion: cross-domain long-range learning for general image fusion via swin transformer," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 7, pp. 1200–1217, 2022.

[45] 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, 2021.

[46] 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*, vol. 44, no. 1, pp. 502–518, 2022.

[47] X. Hang, P. Ye, and G. Xiao, "VIFB: a visible and infrared image fusion benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 104-105, Virtual Conference, June 2020.

[48] A. A. Laghari, H. He, M. Shafiq, and A. Khan, "Assessment of quality of experience (QoE) of image compression in social cloud computing," *Multiagent and Grid Systems*, vol. 14, no. 2, pp. 125–143, 2018.

[49] K. Shahid, H. He, A. H. Magsi, and R. A. Laghari, "Quality of service (QoS): measurements of image formats in social cloud computing," *Multimedia Tools and Applications*, vol. 80, pp. 4507–4532, 2019.

[50] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 94–109, 2012.

[51] 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.