remote sensing

MDPI

Article

GF-Detection: Fusion with GAN of Infrared and Visible Images for Vehicle Detection at Nighttime

Peng Gao, Tian Tian *, Tianming Zhao, Linfeng Li, Nan Zhang and Jinwen Tian

School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China; gaopengde@hust.edu.cn (P.G.); tming@hust.edu.cn (T.Z.); d201780635@hust.edu.cn (L.L.); d202180998@hust.edu.cn (N.Z.); jwtian@mail.hust.edu.cn (J.T.)

* Correspondence: ttian@hust.edu.cn

Abstract: Vehicles are important targets in the remote sensing applications and nighttime vehicle detection has been a hot study topic in recent years. Vehicles in the visible images at nighttime have inadequate features for object detection. Infrared images retain the contours of vehicles while they lose the color information. Thus, it is valuable to fuse infrared and visible images to improve the vehicle detection performance at nighttime. However, it is still a challenge to design effective fusion models due to the complexity of visible and infrared images. In order to improve vehicle detection performance at nighttime, this paper proposes a fusion model of infrared and visible images with Generative Adversarial Networks (GAN) for vehicle detection named GF-detection. GAN is utilized in the image reconstruction and introduced in the image fusion recently. To be specific, to exploit more features for the fusion, GAN is utilized to fuse the infrared and visible images via the image reconstruction. The generator fuses the image features and detection features, and then generates the reconstructed images for the discriminator to classify. Two branches, visible and infrared branches, are designed in the GF-detection model. Different feature extraction strategies are conducted according to the variance of the visible and infrared images. Detection features and self-attention mechanism are added to the fusion model aiming to build a detection task-driven fusion model of infrared and visible images. Extensive experiments based on nighttime images are conducted to demonstrate the effectiveness of the proposed fusion model in night vehicle detection.

Keywords: vehicle detection; fusion of visible and infrared images; Generative Adversarial Networks

Citation: Gao, P.; Tian, T.; Zhao, T.; Li, L.; Zhang, N.; Tian, J. GF-Detection: Fusion with GAN of Infrared and Visible Images for Vehicle Detection at Nighttime. *Remote Sens.* **2022**, *14*, 2771. https://doi.org/10.3390/rs14122771

Academic Editors: Angel D. Sappa and Gemine Vivone

Received: 13 April 2022
Accepted: 3 June 2022
Published: 9 June 2022

1. Introduction

Vehicles are important targets in remote sensing applications and nighttime vehicle detection has been a hot topic in recent years. Vehicle detection tasks at nighttime could provide valuable information such as the size, category and location distribution, which is utilized in Smart Parking [1] and Battlefield Situational Awareness [2]. However, nighttime visible images suffer from poor image quality due to the poor illumination. The contour and color information are lost, which make the vehicle targets difficult to distinguish from the background. Infrared images reflect the thermal characters of targets and preserve significant contrast to the background. In this way, infrared images could retain the complementary features of the visible images. Nevertheless, the color features are absent in the infrared images which are exploited in the vehicle classification and the discrimination of the targets and interference. Therefore, it is valuable to fuse infrared and visible images to improve the nighttime vehicle detection performance.

A variety of visible and infrared vehicle datasets have recently been publicly released, such as DLR 3K [3], VEDAI [4], VIVID [5] and NPU_CS_UAV_IR_DATA [6]. Some datasets containing paired infrared and visible images for image fusion are proposed. However, there is still a lack of annotated vehicle datasets with paired images of the infrared and visible images, especially at nighttime.

Different images contain various illuminations at nighttime. Figure 1 shows that visible images vary in color and context according to the different illumination conditions. Figure 2 shows

that vehicles in the visible images and infrared images vary in color and context according to the different illumination conditions. There is a large amount of interference to the vehicle targets in the infrared images due to the lack of color information.

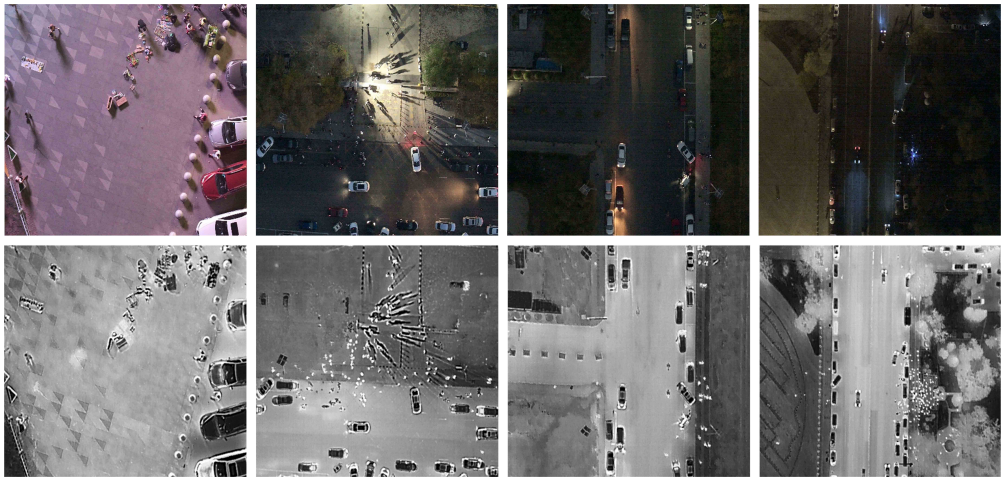

Figure 1. Some examples of paired infrared and visible images. The first row is the visible image and the second row is the infrared image.

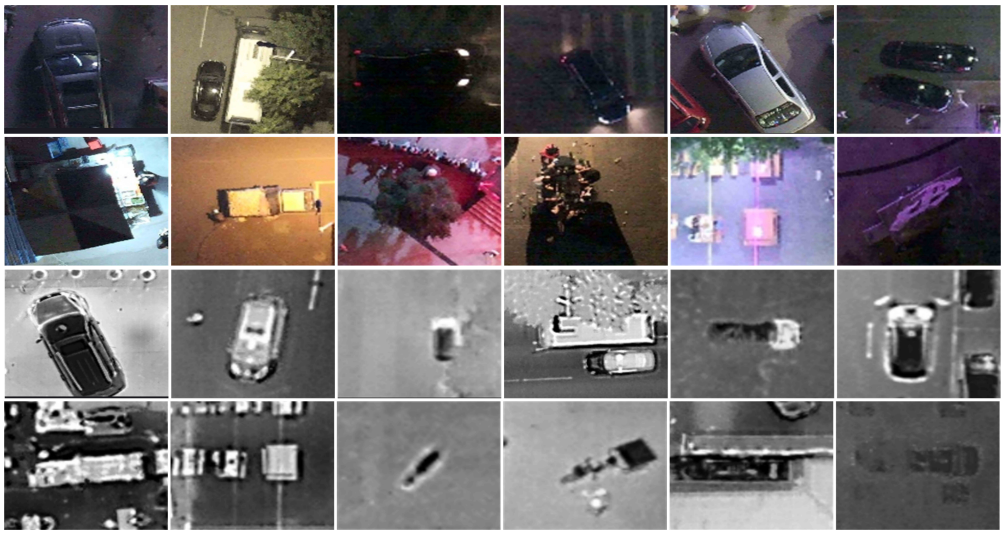

Figure 2. Some examples of paired vehicles and interferences. The first row is the vehicle in the visible image, the second row is the interference in the visible image, the third row is the vehicle in the infrared image and the fourth row is the interference in the infrared images.

There is a variance of visible and infrared images at nighttime. Figure 1 visualizes some examples of paired infrared and visible images at nighttime. Figure 2 presents the vehicle targets and inferences in the visible and infrared images. These images are collected from Drone RGBT Crowding counting datasets [7]. Images in the visible and infrared images vary in color, context and interference. Visible images are RGB images and contain colorful vehicles. Vehicles in the visible images lose the color or contour information in low illumination. Infrared images are gray images and contain limited color information of vehicles. Vehicles in the infrared images retain the contour information in low illumination while suffering more interference. The variances of the visible and infrared images raise a possibility of image fusion for vehicle detection at nighttime. Many fusion works are put forward for utilizing the complementary features of the visible and infrared images to improve the detection performance.

Traditional fusion models based on features for detection contain visible and infrared branches. Each branch extracts features via the convolution layers and feeds them into the fusion model. The fusion model fuses the two features and feeds them into the detection branch. Detection-guided branch plays a vital role in the fusion model for vehicle detection. A detection branch is placed behind the fusion models in the traditional fusion work. However, the detection features and the detection model learned are not sufficient and it is still necessary to add detection supervision to the infrared and visible branch, respectively.

The variance of the visible and infrared images should be exploited in the fusion works. The existing fusion works try to extract the complementary features from the visible and infrared images for vehicle detection. However, the complementary features are difficult to distinguish from all features extracted by the convolution layers. Many decomposition manners such as DRF [8], Decomposition [9] and TLF [10], try to analyze the variance of the infrared and visible images and choose the valuable features for image fusion. The valuable features refer to the target feature for object detection, such as illumination priors, high-frequency features and attribute representation. All these works try to find the interoperability of the fusion models and optimize the fusion mechanism. However, those decompositions are evaluated in a small dataset and more evaluations should be tested in the various images at nighttime to demonstrate the effectiveness of those decomposition theories. Self-attention mechanisms [11] and feature pyramid networks [12] are added into the fusion work to enhance the fusion performance. Performance improvements have been achieved with those modules. However, the numerous modules are placed with numerous convolution layers which add to the burden of the network.

Generative Adversarial Networks (GAN) [13] is a popular generation model used in the image reconstruction. GAN contains a generator and a discriminator in order to generate the similar images with the reference images. The generator consists of several convolution layers with up-pooling operations to enlarge the size of the inputs. The discriminator is a classification with convolution layers and down-pooling operations to discriminate the fake reconstructed images from the true reference images. GAN is introduced into the fusion work of infrared and visible images in recent years. Compared with the traditional fusion models based on features, fusion models with GAN are able to integrate feature extraction, feature fusion and image reconstruction in a single model and produce a promising fused result. However, the reconstructed images by GAN are evaluated by human visual perception which is not suitable for high vision tasks. Moreover, there are few fusion works with GAN for vehicle detection at nighttime, and the variances of the visible and infrared images are not considered in the existing fusion works. Furthermore, the fusion models are not guided by the detection tasks directly and the convergence process is not guided by the detection tasks. Many works succeed to introduce high-vision task driven branches into the Generative Adversarial Networks, such as detection guided [14] and instance aware [15]. However, the supervision manners do not optimize the image generation directly and receive small improvement in the high-vision tasks.

Though many fusion works of the infrared and visible images have been put forward, there is still a challenge for the fusion of infrared and visible images for vehicle detection. First, the great feature complexity of visible and infrared images at nighttime is the bottleneck of the fusion model. The detection performance at nighttime via the fusion of the visible and infrared images is still unsatisfactory. Second, the fusion manners fusing the visible and infrared branches together are limited which cannot distinguish the valuable features from invaluable features for vehicle detection effectively. The detection supervision could not guide the feature extraction in each visible and infrared branch directly. Third, most of the existing fusion works are based on features extracted by the convolutions. The feature pyramid network and self-attention module are introduced into the fusion work to enhance the fusion performance. However, the fusion mechanism in the fusion work are limited and poor improvement is achieved with complex structures. Lastly, numerous detection branches increase the burden of the fusion models, which is not suitable for the training process.

In this paper, in order to improve the vehicle detection performance at nighttime, we propose a fusion model (GF-detection) of visible and infrared image fusion with GAN for the vehicle detection task. A detection task-driven fusion model with GAN is designed. GF-detection contains the visible branch, infrared branch, self-attention fusion model and detection model. Each branch contains a generator, a discriminator and a detection backbone utilized to extract the detection features.The fusion operation is as follows: two tensors from visible images and visible detection features are concatenated with convolution layers in the visible fusion model in the visible branch. Two tensors from infrared images and infrared detection features are concatenated with convolution layers in the infrared fusion model in the infrared branch. Two tensors from sub-branches, such as the visible branch and infrared branch, are concatenated in the self-attention fusion model and are sent to the detection model. Since visible images contain more contours and color information while infrared images contain more targets in the low illumination condition, visible images are suitable to extract the semantic features for vehicle classification, and the infrared images are suitable to extract the salient features for vehicle detection. In this way, different feature extraction strategies are conducted for the visible and infrared images. Semantic features are extracted with deep convolution models in the visible images for the vehicle classification, and salient features are extracted with swallow convolution models in the infrared images for the vehicle detection. The self-attention fusion model contains a channel attention and a spatial attention to fuse the two reconstructed images. There are two fusion stages in our fusion model, the first one is the fusion in the GAN in each branch of the images and detection features, and the second fusion is conducted in the self-attention fusion mechanism.

All innovations are as follows:

- Aiming to fuse the visible and infrared images effectively besides the feature extraction by convolutions, Generative Adversarial Networks (GAN) is introduced into the fusion model for vehicle detection at nighttime. Image fusion is conducted via the image reconstruction.
  Visible and infrared branch are included, and each branch contains a GAN for image fusion. Visible branch transfers the visible images to the infrared images via GAN and infrared branch converts the infrared images to the visible images via GAN. Compared with other fusion models without GAN, fusion models with GAN could exploit more features for image fusion.
- In order to enhance the detection features in reconstructed images, detection features are added to the GAN module. The detection features optimize the image generation directly. Two detection features extracted from the visible and infrared images are added into the generator of the GAN. Compared with the detection branches, the detection features are more effective for vehicle detection without extra burden introduced.
- To extract features of the visible and infrared images with different characters, various structures of the visible branch and infrared branch are designed in the subbranch. Many convolution layers exist in the encoder, detection backbone and decoder in the visible branch, while a few convolution layers exist in the encoder, detection backbone and decoder in the infrared branch. The different feature extraction strategies are suitable for the feature extraction of different source images.
- In order to improve the fusion performance with limited augment of complexity, a self-attention fusion model is employed as the second fusion model. With the channel and spatial attention map learned from the fusion features, more effective fusion features are exploited for vehicle detection tasks. Self-attention fusion model behind the visible and infrared branches could enhance the fusion performance with few parameters increased.
- A vehicle dataset containing paired infrared and visible images at nighttime are collected and labeled as a vehicle dataset for the fusion study of infrared and visible images.

The Section 1 paper describes the background of our research and the highlights of our work. The Section 2 briefly describes the work related to the fusion work of the visible and infrared images for vehicle detection at nighttime. The Section 3 describes the structure of the model in detail, explains the architecture and lists the loss functions. The Section 4

is devoted to experiments and analysis, listing and discussing the results of quantitative comparisons. The Section 5 presents our conclusions.

2. Related Works

2.1. Vehicle Datasets in Nighttime

Though a large amount of visible and infrared vehicle datasets are open, most of them are not paired, which is not suitable for image fusion for vehicle detection. Some common scenes with paired infrared and visible videos containing vehicles are released recently, such as OSU Color-Thermal Database [16], VLIRVDIF [17] and VIFB [18]. These datasets are employed for the fusion study of the infrared and visible images. However, most of them contain few vehicles. Drone RGBT Crowding Counting [7] contains infrared and visible images at nighttime containing pedestrians and vehicles for crowed people counting. The vehicles in the Drone RGBT Crowd Counting are unannotated. Many works collecting paired visible and infrared images should be conducted for the fusion study for vehicle detection.

2.2. Vehicle Detection in Nighttime

Most of the recent fusion studies of infrared and visible images for vehicle detection are based on features, such as MFDSSD [19], SKNet [20], multispectral ensemble detection [21], FFECSE [22] and CS-RCNN [23]. The features extracted from the infrared branch and the visible branch by convolution layers are added together with several convolution layers and those fused features are sent to the detection task for vehicle detection. The self-attention mechanism [11] is widely applied in each scale of the convolution layer in order to fuse each scale of infrared and visible features. Feature pyramid networks [12] are employed within several scale fusion layers to enhance the fusion performance. Prior knowledge is also exploited in the fusion model, such as Illumination-aware, IAFR-CNN [24] and IATDNN+IAMSS [25].

Though many works have been conducted on the structure of the fusion models, they are similar, and these works increases the burden of the network without much improvement received in the vehicle detection. The effectiveness of these models should be evaluated in vehicle detection. Furthermore, most codes of those works are not open, and the effectiveness is not validated.

2.3. GAN for Image Fusion of Infrared and Visible Images

The fusion of infrared and visible images with GAN is a study hotpot recently. FusionGAN [26] is the pioneer to introduce GAN into infrared and visible image fusion. DRF [8] divides the infrared or visible images into the scene representation and attribute representation and fuses them together in the image generation. AttentionFGAN [27] introduces multiple classification constraints to simultaneously estimate the two probability distributions of source images. The GAN methods conduct feature extraction, feature fusion and image reconstruction in an implicit manner, however, it is still unstable for the whole image reconstruction. The styles the fused image learns are sensitive to the loss supervision in the image generation. The styles of the fused images are not stable, and the convergence is unpredictable.

Fusion based on GAN is not an end-to-end model for vehicle detection. All features extracted from the infrared and visible images are fused in the reconstructed images and are then extracted from the reconstructed images by the detection model. The image generation of GAN is evaluated by human visual perception and might not be suitable for high-vision tasks, such as vehicle detection. Many detection supervision manners are introduced into the image generation work to improve the generated images detection performance, such as instance aware [28] and detection guide [14], are introduced into the GAN fusion to enhance the detection features. STDFusionNet [28] combines a salient target mask into the fusion of the infrared and visible images. DUNIT [14] tries to introduce detection supervision into the image-to-image translation from night to day.

The supervision manners introduced into the GAN fusion are effective, however, the mechanism of how the extra supervision branch effects image generation is still not clear, and the effects of the added supervision manners should be evaluated both in theory and practice.

3. Method

The overall structure of GF-Detection is illustrated in Figure 3. The whole structure of GF-detection contains the visible branch, infrared branch and self-attention fusion model. Visible branch is aiming to reconstruct the visible images via the fusion of the visible images and detection features. The infrared branch aims to reconstruct the infrared images via the fusion of the infrared images and detection features. The trained detection models extract the detection features from the visible and infrared images into the image reconstruction. The self-attention fusion model fuses two fused features and sends them to the detection model.

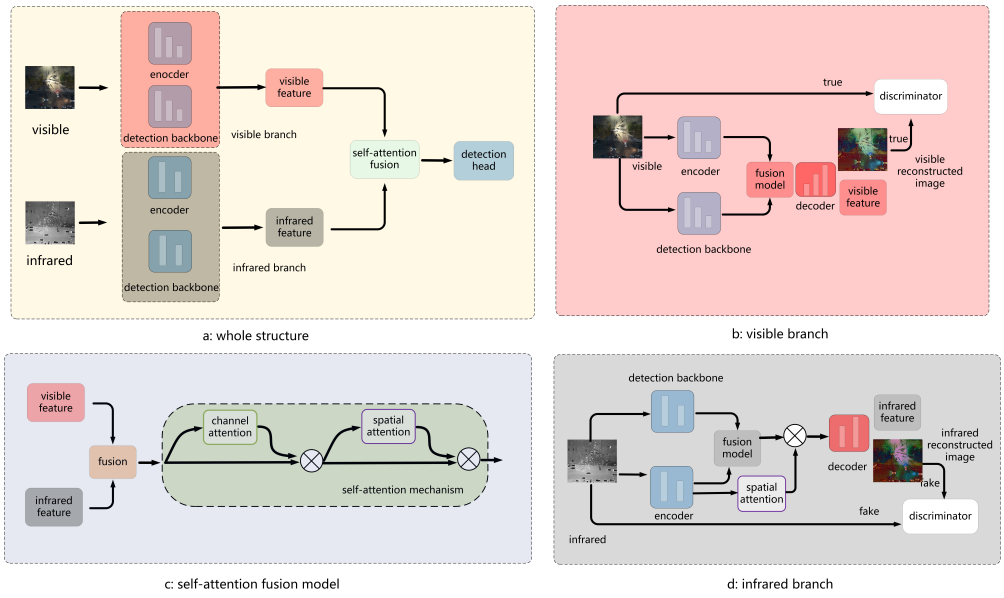

Figure 3. The overall structure of GF-Detection is illustrated in figure. The whole structure of GF-detection contains the visible branch, infrared branch and self-attention fusion model. Visible branch is aiming to reconstruct the visible images via the fusion of the visible images and detection features. The infrared branch aims to reconstruct the infrared images via the fusion of the infrared images and detection features. The trained detection models extract the detection features from the visible and infrared images into the image reconstruction. The self-attention fusion model fuses two fused features and sends them to the detection model.

GAN is employed in order to reconstruct the images with different source images and detection features. For example, the visible images are reconstructed with visible images and the detection features from the visible images and infrared images, then these images are discriminated with the true visible images and the infrared images. The reconstructed visible images combine the style features learned from the generative adversarial networks and the detection features extracted by the trained detection models.

The trained detection models are employed to extract the detection features into the image generation process in order to enhance the detection features. The self-attention fusion model designs a fusion model of the visible and infrared branches with self-attention mechanism in order to enhance the fusion features for vehicle detection.

Two stages of fusions are employed in the whole structure. The first stage of fusions is carried out in the visible branch and infrared branch. The detection feature and style feature are fused by the image reconstruction by GAN. The second stage fusion is conducted in the self-attention fusion model, which fuses the visible and infrared features.

The detailed parameters of GF-detection are listed in the Table 1.

Table 1. Parameters of the main module of GF-detection.

Module	Submodule	Parameters
generator	encoder	Conv(1,3,7,2,0)+maxpooling, Conv(3,64,4,2,1)+BatchNorm2d, Conv(64,128,4,2,1)+BatchNorm2d, Conv(128,256,4,2,1)+BatchNorm2d, Conv(256,512,4,2,1)+BatchNorm2d
	decoder	Convtranspose(512,256,4,2,1)+BatchNorm2d, Convtranspose(256,128,4,2,1)+BatchNorm2d, Convtranspose(128,64,4,2,1)+BatchNorm2d, Convtranspose(64,3,4,2,1)+BatchNorm2d, uppooling
discriminator	encoder	Conv(3,64,7,2,0)+maxpooling, Conv(64,128,4,2,1)+BatchNorm2d+LeakyReLU, Conv(128,256,4,2,1)+BatchNorm2d+LeakyReLU, Conv(256,512,4,2,1)+BatchNorm2d+LeakyReLU, FC(204800,4096)
fusion	encoder	Conv(3,64,7,2,0)+maxpooling, Conv(64,256, 7,2,0)+BatchNorm2d+LeakyReLU, Conv(256,512, 7,2,0)+BatchNorm2d+LeakyReLU, Conv(512,1024, 7,2,0)+BatchNorm2d+LeakyReLU, Conv(1024,2048,7,2,0)+BatchNorm2d+LeakyReLU
	fusion	Conv(512,512, 3,1,1)+BatchNorm2d+LeakyReLU, Conv(1024,1024,3,1,1)+BatchNorm2d+LeakyReLU, Conv(2048,2048,3,1,1)+BatchNorm2d+LeakyReLU
	decoder	Convtranspose(2048,1024,3,2,1)+BatchNorm2d+LeakyReLU, Convtranspose(1024,512,4,2,1)+BatchNorm2d+LeakyReLU, Convtranspose(512,256,4,2,1)+BatchNorm2d+LeakyReLU, Convtranspose(256,64,4,2,1)+BatchNorm2d+LeakyReLU, Convtranspose(64,3,4,2,1)+BatchNorm2d+LeakyReLU, uppooling
Self-attention fusion model	channel attention model	AdaptiveAvgPool2d+FC(512,32)+LeakyReLU+ Conv(32,512,1,1) AdaptiveMaxPool2d+FC(512,32)+LeakyReLU+ Conv(32,512,1,1)+Sigmoid()
	spatial attention model	torch.mean+torch.max+Conv(2,1,7,1,1)+Sigmoid()
RetinaNet	RetinaNet with ResNet(50)	

3.1. Visible Branch

The visible branch contains a fusion model, a generator and a discriminator. Visible branch is utilized to fuse the visible images and detection features. Detection features extracted from the visible images and style features extracted from the encoder of the generator are fed into the visible fusion model. The structure of the visible branch is in Figure 3b. Five convolution layers are assigned to the encoder, decoder and detection backbone to extract the high-level features for vehicle detection. The fusion model is as follows:

$$Fusion_V = Fusion(V + Detection_V) \tag{1}$$

where $Fusion_V$ is the visible branch, $Fusion$ is the fusion model, V is the visible images, $Detection_v$ is the detection feature extracted from the visible images.

The generator fuses those inputs and generates the synthesis images with fused features, those generated images are fed into the discriminator. The discriminator distinguishes the synthesis images with the visible images and infrared images. The adversarial loss function in the visible branch is as follows:

$$\begin{array}{c}\mathcal{L}_{V\_GAN}(G_V, D_V, Fusion_V, I) = \mathbb{E}_{I\sim p_{data}(I)}[\log D_I(I)] \\ +\mathbb{E}_{Fusion_V\sim p_{data}(Fusion_V)}[\log(1 - D_I(G_V(Fusion_V)))]\end{array} \tag{2}$$

where $\mathcal{L}_{V\_GAN}(G_V, D_V, Fusion_V, I)$ is the adversarial loss in the visible branch. $Fusion_V \sim p_{data}(Fusion_V)$ and $I \sim p_{data}(I)$ are distribution characters of the fusion model and infrared images. Generator G_V generates images with the styles similar to the one of the infrared images. Discriminator D_I distinguishes $G_V(Fusion_V)$ and the infrared images I.

The total loss function of the visible images:

$$L_V(V, G_V) = L_{V\_GAN}(G_V, D_V, Fusion_V, V, I) \tag{3}$$

3.2. Infrared Branch

The structure of the infrared branch is shown in Figure 3d. Three convolution layers are adopted in the encoder, decoder and detection backbone to extract the salient details for vehicle detection. The spatial attention model is added in the fusion model.

Spatial attention formula is as follows:

$$M_S(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) * F \tag{4}$$

where $M_S(F)$ is the channel attention model, F is the input feature, MLP is the convolution layers, $AvgPool$ is the mean operation and $MaxPool$ is the max operation.

Similar to the visible branch, the fusion model, adversarial loss and the infrared branch are calculated as follows:

$$Fusion_I = M_S(Fusion(I + Detection_I)) \tag{5}$$

where $Fusion_I$ is the infrared branch, $Fusion$ is the fusion model, I is the infrared images, $Detection_I$ is the detection feature extracted from the infrared images.

The generator fuses those inputs and generates the synthesis images with fused features, those generated images are fed into the discriminator. The discriminator distinguishes the synthesis images with the infrared images. The adversarial loss function in the infrared branch is as follows:

$$\begin{array}{c}\mathcal{L}_{I\_GAN}(G_I, D_I, Fusion_I, V) = \mathbb{E}_{V\sim p_{data}(V)}[\log D_V(V)] \\ +\mathbb{E}_{Fusion_I\sim p_{data}(Fusion_V)}[\log(1 - D_V(G_I(Fusion_I)))]\end{array} \tag{6}$$

where $\mathcal{L}_{I\_GAN}(G_I, D_I, Fusion_I, V)$ is the adversarial loss in the infrared branch. $Fusion_I \sim p_{data}(Fusion_I)$ and $V \sim p_{data}(V)$ are distribution characters of the fusion model and visible images. Generator G_I generates images with the styles similar to the one of the visible

images. Discriminator D_V distinguishes $G_I(Fusion_I)$ and the true visible images V. The total loss function of the infrared fusion images:

$$L_I(I, G_I) = L_{I\_GAN}(G_I, D_I, Fusion_I, I, V) \tag{7}$$

3.3. Self-Attention Fusion Model

The self-attention fusion model contains a fusion module and a self-attention module. The structure is visualized in Figure 3c. The fusion module fuses the visible features from the visible branch and the infrared branch. The self-attention module contains a channel attention mechanism and a spatial attention mechanism in order to enhance the fusion features for vehicle detection. The fusion model in the self-attention fusion model is calculated as follows:

$$Fusion_S = Fusion(G_V(Fusion_V) + G_I(Fusion_I)) \tag{8}$$

where $Fusion_S$ is the fusion model in the self-attention fusion model. $Fusion$ is the fusion module, $G_V(Fusion_V)$ is the reconstructed visible images, $G_I(Fusion_I)$ is the reconstructed infrared images.

The channel attention model calculates the weight of each channel through the sigmoid activation function by the mean or max value of the input channels, and then multiplied with the input channels to enhance the channel features. The spatial attention model calculates the weight of each point in the feature map though the sigmoid activation function by the mean or max value of the input features, and then multiplied with the input features to enhance the spatial features.

The channel attention formula is as follows:

$$M_c(F) = \sigma(FC(AvgPool(F)) + FC(MaxPool(F))) * F \tag{9}$$

where $M_c(F)$ is the channel attention model, F is the input feature, FC is the fully connected layers, $AvgPool$ is the mean operation and $MaxPool$ is the max operation.

The whole loss function of the self-attention fusion model is as follows:

$$\begin{aligned} Fusion_S = \sigma(&MLP(AvgPool(F)) \\ &+MLP(MaxPool(F))) * F \end{aligned} \tag{10}$$

3.4. Detection Model

RetinaNet is adopted as the detection model for vehicle detection. Focal loss is employed to tackle the imbalance of the positive and negative sampling. The detection model weight is refreshed in the training process. The whole loss in the detection model is as follows:

$$p = Detection(Fusion_S) \tag{11}$$

$$L\_detection(g, p) = L\_regression(g, p) + FL(p_t) \tag{12}$$

where $L\_detection(g, p)$ is the whole loss in the detection model, $L(g, p)$ is the L1 loss function, $detection$ is the detection model, g is the ground-truth annotations, p is the predicted boxes. $FL(p_t)$ is the focal loss, p_t is the modified classification results.

3.5. Total Loss Function

The total loss function in the GF-detection model is as follows:

$$\begin{aligned} L_{GF\_detection}(V, I, G) = \quad & L_{V\_GAN}(G_V, D_V, Fusion_V, I) \\ &+L_{I\_GAN}(G_I, D_I, Fusion_I, V) \\ &+L_{detection}(g, p) \end{aligned} \tag{13}$$

4. Experiment and Discussion

Though variety datasets of paired visible and infrared images are released, there is still a lack of the vehicle datasets with paired visible and infrared images at nighttime. Drone RGBT Crowd counting is an open paired visible and infrared dataset at nighttime for crowd counting. It is collected with Drone. In this paper, we propose a vehicle datasets named RGBT-Vehicle containing paired visible and infrared images by choosing and labeling the vehicle images from Drone RGBT Crowd counting datasets. A total of 825 paired images with visible and infrared images are utilized for vehicle detection. The visible and infrared images are registered and share the same image size. Pixel to pixel mapping has been conducted in the visible and infrared images before the fusion.

Lighting conditions vary in different vehicle scenes. The contour and color information of vehicles are affected by low illumination conditions, especially the black vehicles hidden in the dark. The size of vehicles varies from 20×30 to 100×50 pixels. Much interference, such as woods, buildings and stones, share the same characters with the vehicles which hinders vehicle detection. Vehicles contain cars, trucks and buses, etc. The number of different vehicle categories varies greatly, among which cars are the largest. All vehicles are treated as the same category and no more classification works are performed due to the imbalance of various vehicle targets.

A server with 2080Ti GPU is used in our experiments. The learning rate of the detection model is 0.001, which is reduced to 0.1 times per 10 epochs. The learning rate of the GAN is 0.001, which is reduced to 0.1 times per 10 epochs. The batch sizes of detection model and GAN are 4. The images fed into the GAN and detection model are 512×512 pixels. As a single vehicle category, the detection performance is evaluated by the Precision and Recall curve (PR curve). The PR Curve evaluates the detection results in a range of IoU thresholds. Furthermore, we set the IoU threshold as 0.5 to obtain the precision and recall in order to achieve a quantitative result.

4.1. Experiments on RGBT-Vehicle

Six detection models, RetinaNet trained on the visible images (visible), RetinaNet trained on the infrared images (infrared), the fusion model based on features (Fusion), the fusion model with BGR channels (BGR), distangle with detection task oriented (Distangle-oriented) and GF-detection models are tested. Figure 4 illustrates the structure of the comparisons.

RetinaNet trained on the visible images (Figure 4a) is tested on the visible images. RetinaNet trained on the infrared images (Figure 4b) is tested on the infrared images. The fusion model based on features (Figure 4c) refers to the models which fuse features from the visible branch and infrared branch and fed into the detection model. The fusion model with BGR channels (Figure 4e) refers to the fusion model dividing the visible branch into blue, green and red branch and fusing four branches. The distangle with detection task-oriented (Figure 4d) refers to the fusion model with GAN dividing the reconstructed images with mutual features and identical features. Only complementary features are fed into the fusion model. The detection model weight is refreshed during the training to adapt to the reconstructed images.

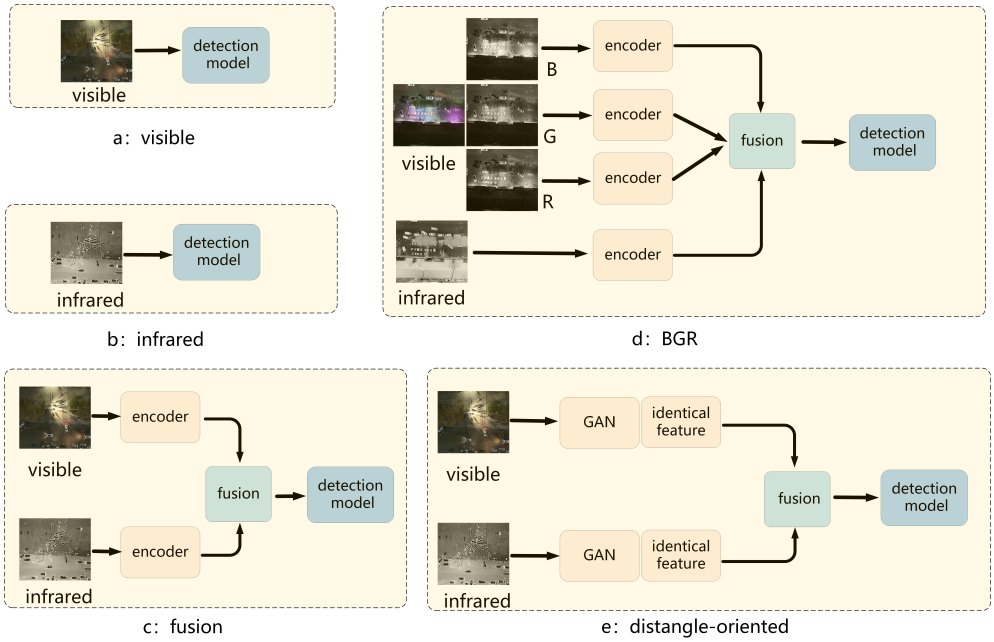

Figure 4. The structures of six comparisons.

Figure 5 shows the PR-Curve of six fusion models. Figure 5 visualizes that our GF-detection model achieves the best detection performance. Compared with single images with visible or infrared models, fusion models with two source images achieve an improvement in the vehicle detection.

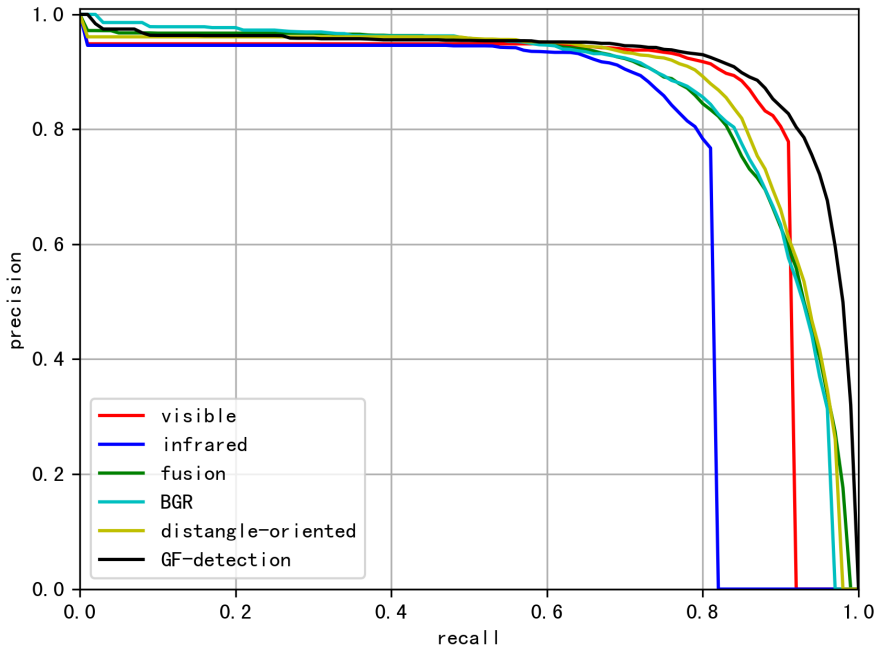

Figure 5. PR-Curve of six fusion models.

Table 2 lists the precision and recall scores with IoU set as 0.5 of six fusion models. It demonstrates that our fusion model achieves the best detection performance in the IoU set as 0.5. Fusion models achieve an improvement compared with the models of single branch. Distangle-oriented and GF-detection achieve an competitive result with GAN, and it proves that GAN could be introduced into the fusion work. Fusion with detection results prove that the detection features are valuable for fusion work.

Table 2. Comparison results of six fusion models with IoU set as 0.5.

Fusion Model	Precision	Recall	F1
visible	71.0	81.0	75.6
infrared	60.2	72.4	65.8
fusion	81.3	92.6	86.5
BGR	81.0	90.4	85.4
distangle-oriented	82.9	92.3	87.3
GF-detection(ours)	**86.7**	**94.4**	**90.3**

Figures 6 and 7 visualize the detection results of six models on the visible and infrared images. These images illustrate that the GF-detection model preserves the most detection boxes with the least missing alarms and false alarms are also the least at nighttime. Detection models based on the single branch suffer a great number of missing alarms and false alarms. The fusion models of visible and infrared images can distinguish the vehicle targets and the interference of the background in weak or strong light conditions and receive an improvement in vehicle detection.

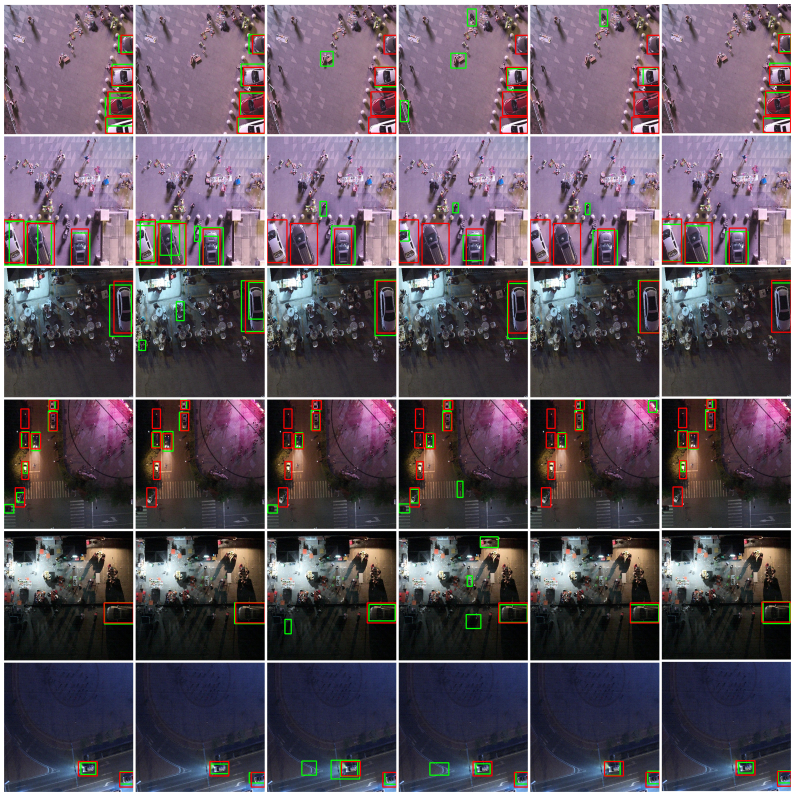

Figure 6. Samples of six fusion models in the visible images. From left to right are visible, infrared, fusion, BGR, Distangle-oriented and GF-detection. Red box denotes the ground-truth annotations, green is the predicted box.

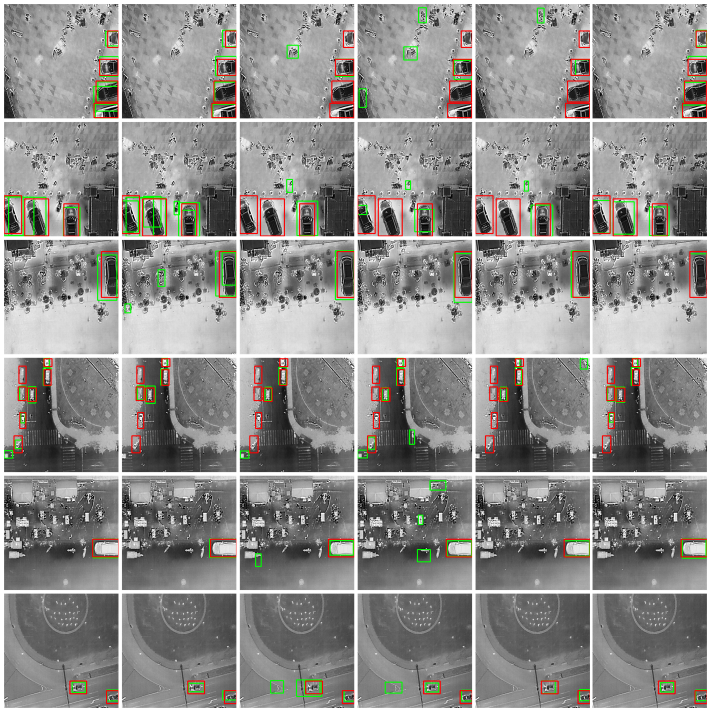

Figure 7. Samples of six fusion models in the infrared images. From left to right are visible, infrared, fusion, BGR, Distangle-oriented and GF-detection. Red box denotes the ground-truth annotation, green is the predicted box.

4.2. Ablation Study

GF-detection introduces GAN, detection features and self-attention fusion models to improve the fusion performance of the visible and infrared images at nighttime. Four models designed as fusions with detection features (visible and infrared branches are fused in the fusion model directly), fusion with GAN (visible GAN branch and infrared GAN branch are fused in the fusion model, the detection model weight is refreshed during the training to adapt to the reconstructed images), fusion with GAN and detection features (visible GAN branch with detection features and infrared GAN branch with detection features in the fusion model) and fusion with GAN, detection features and self-attention (visible GAN branch with detection features and infrared GAN branch with detection features in the self-attention fusion model) namely GF-detection. The PR-Curve is visualized in Figure 8 and the precision-recall scores at IoU 0.5 are listed in Table 2.

Table 2 illustrates that three models we design to fuse visible and infrared images for vehicle detection at nighttime are effective and necessary. Furthermore, three models work together to achieve the best fusion performance. Those loss functions, such as the adversarial loss and detection feature enhanced detection loss, play a positive effect on the fusion performance.

In particular, compared with fusion based on features, fusion models with GAN achieve a better detection performance, which demonstrates the value of the image reconstruction in the fusion work. Two-layer fusion mechanism enables GF-detection fusing the visible and infrared branch effectively.

Compared with fusion GAN, the fusion model with GAN and detection features receives a high score in the precision, recall and F1. It demonstrates the value of the detection features to the adversarial networks. The self-attention fusion models share similar values.

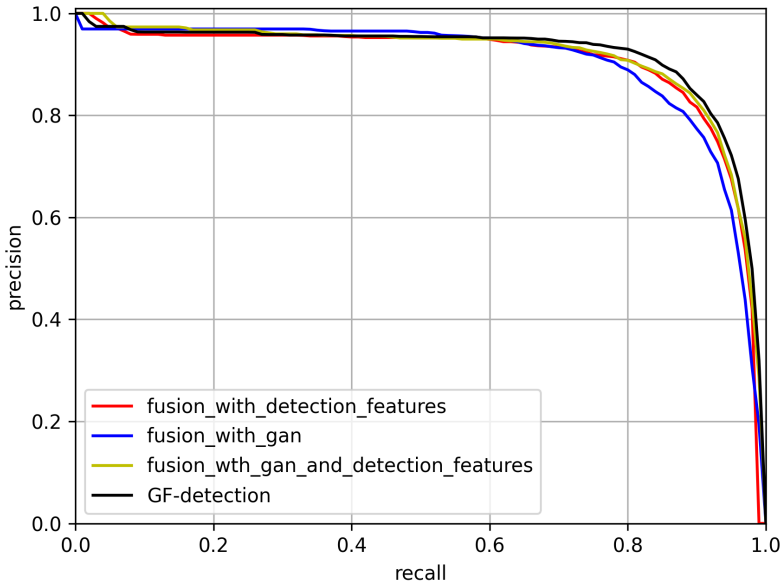

Figure 8. PR-Curve of four fusion models.

4.3. Images Reconstructed by GAN

GAN is employed in the fusion of the visible and infrared images. Different supervisions lead to different styles of the reconstructed images learned. In this section, three GAN models with various loss supervision are evaluated for the vehicle detection, namely GAN, detection task oriented GAN and GAN with detection features enhanced. GAN refers to the fusion model with GAN. Detection task-oriented GAN refers to the fusion model with GAN and a detection model based on the reconstructed images. The detection model weight is refreshed to adapted to the reconstructed images. Detection task oriented GAN. GAN with detection features enhanced refers to the GAN with detection features extracted from the trained detection model. Figures 9 and 10 visualize some samples of the four GAN models and Table 3 lists the detection results of those models.

Table 3. Comparison results of four fusion models with IoU set as 0.5.

Fusion Model	Precision	Recall	F1
fusion with detection features	84.6	92.3	86.5
fusion with GAN	85.1	94.0	89.3
fusion with GAN and detection features	86.0	**94.4**	90.0
GF-detection (ours)	**86.7**	**94.4**	**90.3**

Figures 9 and 10 illustrate that the reconstructed images by GAN are blurred and lack details. More noises are introduced. The color of those images is wired and unrealistic. Images reconstructed by GAN with detection task-oriented own more salient targets, but still suffer unrealistic colors. Images reconstructed by GAN with detection features enhanced is colorful and has many details of the vehicle. The style of those images is the closest to the real scenes and least noise is mixed with the reconstructed images.

GF-detection introduces the two layer fusion with two layer detection task supervisions and preserves the most details in the reconstructed images. Detection task-oriented GAN introduces the single detection branch after the fusion model and preserves fewer details. GAN generates images without extra detection task supervision and keeps the least details.

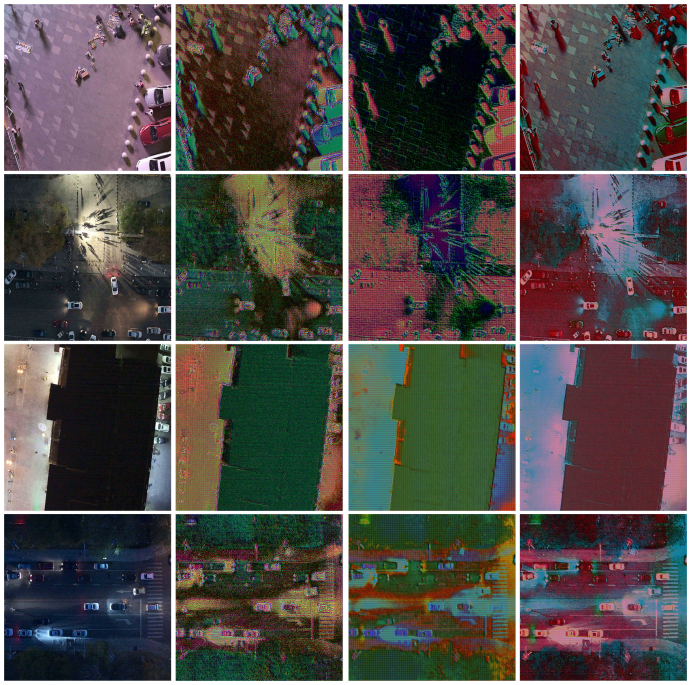

Figure 9. Samples of the four GAN models. From left to right are the visible images, images generated by GAN, images generated by detection task oriented GAN and GAN with detection features enhanced in our models.

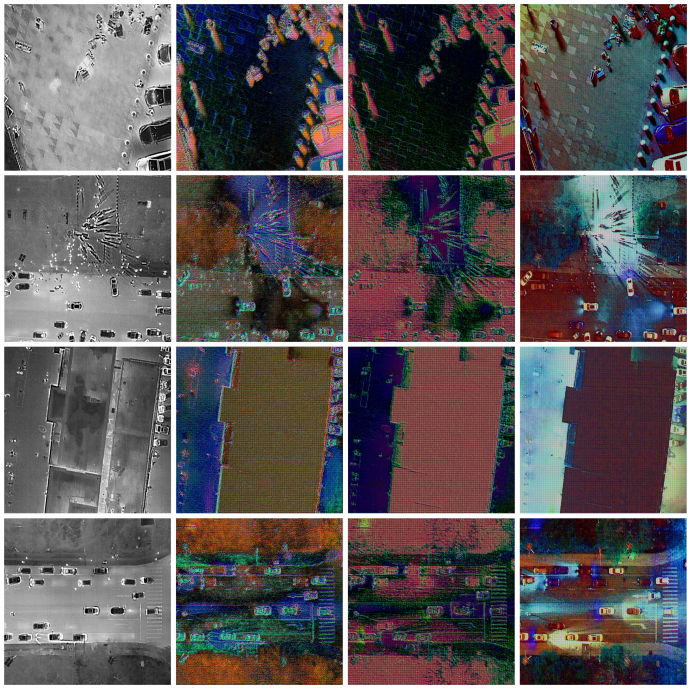

Figure 10. Samples of the four GAN models. From left to right are the infrared images, images generated by GAN, images generated by detection task oriented GAN and images generated by GAN with detection features enhanced.

5. Conclusions

In this study, we design a fusion network of infrared and visible images based on Generative Adversarial Networks for nighttime vehicle detection in the remote sensing scenes. We introduce the Generative Adversarial Networks (GAN) into the fusion task. With the reconstructed images by GAN, the infrared and visible images are fused. Furthermore,

we introduce detection features into the image reconstruction to enhance the detection features for vehicle detection. A self-attention fusion model is designed to fulfill another fusion. Extensive experiments based on paired infrared and visible images demonstrate the effectiveness of our model in the fusion work of visible and infrared images for nighttime vehicle detection.

Author Contributions: Methodology, P.G. and T.T.; software, P.G.; validation, P.G. and N.Z.; formal analysis, P.G.; investigation, P.G., T.T., T.Z., L.L. and J.T.; writing—original draft preparation, P.G.; writing—review and editing, P.G.; supervision, P.G.; project administration, P.G. and J.T.; funding acquisition, T.T. and J.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under 408 Grant 42071339.

Data Availability Statement: Not applicable.

Acknowledgments: We sincerely thank the editor and reviewers for their constructive comments.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GAN	Generative Adversarial Networks
GF-detection	fusion model with GAN for vehicle detection task
IoU	intersection over Union
PR-Curve	precision and recall curve

References

1. Lin, T.; Rivano, H.; Le Mouël, F. A survey of smart parking solutions. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 3229–3253. [CrossRef]
2. Peng, H.; Zhang, Y.; Yang, S.; Song, B. Battlefield image situational awareness application based on deep learning. *IEEE Intell. Syst.* **2019**, *35*, 36–43. [CrossRef]
3. Mandal, M.; Shah, M.; Meena, P.; Devi, S.; Vipparthi, S.K. AVDNet: A Small-Sized Vehicle Detection Network for Aerial Visual Data. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 494–498. [CrossRef]
4. Razakarivony, S.; Jurie, F. Vehicle detection in aerial imagery: A small target detection benchmark. *J. Vis. Commun. Image Represent.* **2016**, *34*, 187–203. [CrossRef]
5. Bozcan, I.; Kayacan, E. Au-air: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; pp. 8504–8510.
6. Liu, X.; Yang, T.; Li, J. Real-time ground vehicle detection in aerial infrared imagery based on convolutional neural network. *Electronics* **2018**, *7*, 78. [CrossRef]
7. Shaniya, P.; Jati, G.; Alhamidi, M.R.; Caesarendra, W.; Jatmiko, W. YOLOv4 RGBT Human Detection on Unmanned Aerial Vehicle Perspective. In Proceedings of the 2021 6th International Workshop on Big Data and Information Security (IWBIS), Depok, Indonesia, 23–25 October 2021; pp. 41–46.
8. Xu, H.; Wang, X.; Ma, J. DRF: Disentangled representation for visible and infrared image fusion. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–13. [CrossRef]
9. Chen, J.; Li, X.; Luo, L.; Mei, X.; Ma, J. Infrared and visible image fusion based on target-enhanced multiscale transform decomposition. *Inf. Sci.* **2020**, *508*, 64–78. [CrossRef]
10. Deng, C.; Liu, X.; Chanussot, J.; Xu, Y.; Zhao, B. Towards perceptual image fusion: A novel two-layer framework. *Inf. Fusion* **2020**, *57*, 102–114. [CrossRef]
11. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 October 2018; pp. 3–19.
12. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
13. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*; MIT Press Direct: Cambridge, MA, USA, 2014; Volume 27.
14. Bhattacharjee, D.; Kim, S.; Vizier, G.; Salzmann, M. Dunit: Detection-based unsupervised image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 4787–4796.

15. Ma, S.; Fu, J.; Chen, C.W.; Mei, T. Da-gan: Instance-level image translation by deep attention generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 5657–5666.
16. Teutsch, M.; Muller, T.; Huber, M.; Beyerer, J. Low resolution person detection with a moving thermal infrared camera by hot spot classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 209–216.
17. Ellmauthaler, A.; Pagliari, C.L.; da Silva, E.A.; Gois, J.N.; Neves, S.R. A visible-light and infrared video database for performance evaluation of video/image fusion methods. *Multidimens. Syst. Signal Process.* **2019**, *30*, 119–143. [CrossRef]
18. Zhang, X.; Ye, P.; Xiao, G. VIFB: A visible and infrared image fusion benchmark. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 104–105.
19. Chen, Y.; Shin, H. Multispectral image fusion based pedestrian detection using a multilayer fused deconvolutional single-shot detector. *J. Opt. Soc. Am. A* **2020**, *37*, 768–779. [CrossRef] [PubMed]
20. Ding, L.; Wang, Y.; Laganière, R.; Huang, D.; Luo, X.; Zhang, H. A robust and fast multispectral pedestrian detection deep network. *Knowl.-Based Syst.* **2021**, *227*, 106990. [CrossRef]
21. Takumi, K.; Watanabe, K.; Ha, Q.; Tejero-De-Pablos, A.; Ushiku, Y.; Harada, T. Multispectral object detection for autonomous vehicles. In Proceedings of the on Thematic Workshops of ACM Multimedia 2017, Mountain View, CA, USA, 23–27 October 2017; pp. 35–43.
22. Xiao, X.; Wang, B.; Miao, L.; Li, L.; Zhou, Z.; Ma, J.; Dong, D. Infrared and visible image object detection via focused feature enhancement and cascaded semantic extension. *Remote Sens.* **2021**, *13*, 2538. [CrossRef]
23. Zhang, Y.; Yin, Z.; Nie, L.; Huang, S. Attention based multi-layer fusion of multispectral images for pedestrian detection. *IEEE Access* **2020**, *8*, 165071–165084. [CrossRef]
24. Li, C.; Song, D.; Tong, R.; Tang, M. Illumination-aware faster R-CNN for robust multispectral pedestrian detection. *Pattern Recognit.* **2019**, *85*, 161–171. [CrossRef]
25. Guan, D.; Cao, Y.; Yang, J.; Cao, Y.; Yang, M.Y. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Inf. Fusion* **2019**, *50*, 148–157. [CrossRef]
26. Ma, J.; Yu, W.; Liang, P.; Li, C.; Jiang, J. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Inf. Fusion* **2019**, *48*, 11–26. [CrossRef]
27. Li, J.; Huo, H.; Li, C.; Wang, R.; Feng, Q. AttentionFGAN: Infrared and visible image fusion using attention-based generative adversarial networks. *IEEE Trans. Multimed.* **2020**, *23*, 1383–1396. [CrossRef]
28. Ma, J.; Tang, L.; Xu, M.; Zhang, H.; Xiao, G. STDFusionNet: An infrared and visible image fusion network based on salient target detection. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–13. [CrossRef]