Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 1896029, 10 pages
https://doi.org/10.1155/2021/1896029

Research Article

FA-YOLO: An Improved YOLO Model for Infrared Occlusion Object Detection under Confusing Background

Shuangjiang Du,[1] **Baofu Zhang**,[1] **Pin Zhang**,[2] **Peng Xiang**,[1] **and Hong Xue**[1]

[1]*College of Communication Engineering, Army Engineering University of PLA, Nanjing 210072, China*
[2]*College of Field Engineering, Army Engineering University of PLA, Nanjing 210072, China*

Correspondence should be addressed to Shuangjiang Du; shuangjiangdu@163.com

Received 24 July 2021; Accepted 27 October 2021; Published 20 November 2021

Academic Editor: Yin Zhang

Infrared target detection is a popular applied field in object detection as well as a challenge. This paper proposes the focus and attention mechanism-based YOLO (FA-YOLO), which is an improved method to detect the infrared occluded vehicles in the complex background of remote sensing images. Firstly, we use GAN to create infrared images from the visible datasets to make sufficient datasets for training as well as using transfer learning. Then, to mitigate the impact of the useless and complex background information, we propose the negative sample focusing mechanism to focus on the confusing negative sample training to depress the false positives and increase the detection precision. Finally, to enhance the features of the infrared small targets, we add the dilated convolutional block attention module (dilated CBAM) to the CSPdarknet53 in the YOLOv4 backbone. To verify the superiority of our model, we carefully select 318 infrared occluded vehicle images from the VIVID-infrared dataset for testing. The detection accuracy-mAP improves from 79.24% to 92.95%, and the F1 score improves from 77.92% to 88.13%, which demonstrates a significant improvement in infrared small occluded vehicle detection.

1. Introduction

Infrared target detection is a hot topic in object detection due to its specific characteristics and special demands. The infrared images have some inherent defects; for instance, infrared targets captured by the infrared cameras are not distinguished in the shape and boundary, which is easily to be misclassified by the environment information; secondly, compared with the visible images, the infrared images contain much more noise such as the Gaussian noise, which may depress the detection accuracy, if not preprocessed. What is more, as for the infrared remote sensing targets, the pixels are much smaller than the ordinary images [1]. All of these features make the infrared target detection more challenging than the normal detection tasks.

Since the infrared remote sensing targets are small and weak, the current methods are feature fusing [2, 3] and multiscale detection [4] to keep the small-scale features. As for the noise impact, the common method is the use of noise filters to suppress the background, such as the median and Robinson filters [5]. Moreover, the infrared datasets are not as sufficient as the visible datasets, which means that they are insufficient to train the model with infrared images in the same way as that with visible images. Thus, transfer learning [6, 7] is a good way to make up for the deficiency.

Nevertheless, the current papers focus more on the infrared small, dim targets without too much confusing background information, while the infrared object detection under the confusing background is not being sufficiently studied. Usually in this scene, the targets are occluded by the useless information from the wild environment, such as the trees, the shadow, and other ground features. The background information may invalidate the detection performance of the model and cause a false decision; that is to say, the detection falsely regards the negative sample as the targets resulting in a low precision. However, in the current military field, the most application scenarios are in the wild complex environments; thus, it is of vital practical

importance to improve the detection performance of the models so that we can still detect the weak and occluded targets in complex environments precisely. Last but not the least, a good detection model can replace hand labor and increase the efficiency of surveillance and detection, as shown in Figure 1, and our paper tries to solve the detection issues in this field.

In terms of the above issues, our paper proposes the focus and attention mechanism-based YOLO (FA-YOLO) model. First of all, to mitigate the impact of confusing background information, we change the YOLOv4 data flowing structure and introduce the negative sample focusing mechanism during the training process. After several epochs of training, the model selects a number of false-positive samples and maps them into the corresponding locations in the feature map and trains them again. Through focusing on the confusing sample training, the model could learn to be more precise.

Secondly, to enhance the features of small objects, we reconstruct the backbone network of YOLOv4 by adding an attention mechanism to the CSPDarknet53 network. We plug the sequent channel and spatial attention block after each residual block; meanwhile, to increase the reception field, we change the convolutional kernel in spatial attention into a dilated convolutional kernel.

Additionally, we use CycleGAN [8] to create infrared images from the visible images to make up for an insufficient infrared training dataset. Transfer learning is also used to promote the optimization of the model parameters. To further verify the superiority of our model, we also add SSD [9], faster R-CNN [10], and YOLOv3 [11] as the comparison models. Compared with the original YOLOv4 [12] model, the detection accuracy-mAP$_{50}$ of the FA-YOLO models improves from 79.24% to 92.95%, and the F1 score improves from 77.92% to 88.13%, which has a state-of-the-art performance.

The main contributions of our work are as follows:

(1) Use GAN to increase the amount of the infrared images and transfer learning to promote the training process

(2) Add a negative sample focusing mechanism to the YOLOv4 model, let it focus more on the negative sample training to reduce the impact of the confusing background, and thus improve the detection accuracy of the model

(3) Fix the dilated convolutional block attention module (dilated CBAM) into the CSPDarknet53 to enhance the features of small targets

Section 2 surveys the related works. Section 3 explains the FA-YOLO in theory. Section 4 is the experiment, and Section 5 concludes the whole paper.

2. Related Works

This section briefly surveys the related works in infrared small target detection and attention mechanism.

2.1. Infrared Small Target Detection. Infrared object detection mainly contains infrared person detection [6, 13–15], infrared vehicle detection [7, 16, 17], infrared aircraft detection [5], and infrared creature recognition and counting [18]. Usually, the lack of an infrared dataset for training and the unclear infrared image features are the problems that need to be overcome.

Transfer learning [6, 7, 19] is usually used for the insufficient training datasets; thus, it is also effective in the infrared dataset training. The possibility mainly relies on the similar image features of the two datasets. The similarity and the huge pretraining dataset are the two conditions needed for transfer learning. The generative adversarial network (GAN) [6] is another method applied to make up for the insufficient infrared datasets through generating infrared images in different styles from visible images.

Wang et al. [2] propose the MNET network, using only three downsampling operations to preserve the features of small infrared targets and using dense connection of the feature map to keep the size all the same; Xu and Wu [3] also use DenseNet and expand it to four scales of anchor boxes in YOLOv3; Zhang et al. [20] uses a double multiscale feature pyramid network to combine different semantic and resolution feature levels.

2.2. Attention Mechanism. CBAM [21] is a simple yet effective attention module for feedforward convolutional neural networks, generating both channel and spatial attention maps separately. It is a lightweight and general module, and it can be integrated into any CNN architectures seamlessly with negligible overheads and is end-to-end trainable along with base CNNs. BAM [22] is also a two-dimensional attention module, which is placed at each bottleneck of models where the downsampling of feature maps occurs. AS-YOLO [23] adds the CBAM after the fusion of different scale feature maps in the PANet so as to enhance the fused features. Gao et al. [24] add a channel attention module (ECANet [25]) after all residual modules of CSPDarknet53 in YOLOv4, and its module mainly consists of two parts, namely, dimensionless local cross-channel interaction and one-dimensional convolution operation with the size of an adaptive convolution kernel. Chen et al. [26] construct a multilevel feature pyramid, use the attention model to obtain the salient features of different levels, and fuse the salient features of different levels for SAR ship detection in multiscale and complex scenarios.

3. The Proposed Method

3.1. Work Flow. The whole procedure of the FA-YOLO is shown in Figure 2. After pretraining, use CycleGAN to generate enough infrared images and put them to the detection model for the final training. The FA-YOLO consists of dilated CBAM and hard example mining module and could detect the small targets and delete the confusing negative sample.

During the transfer learning process, we use UCAS-AOD as the pretraining dataset; it contains 510 visible

Figure 1: Some examples of VIVID-infrared images carefully selected, in which 40% area of the vehicles is occluded by the complex background.

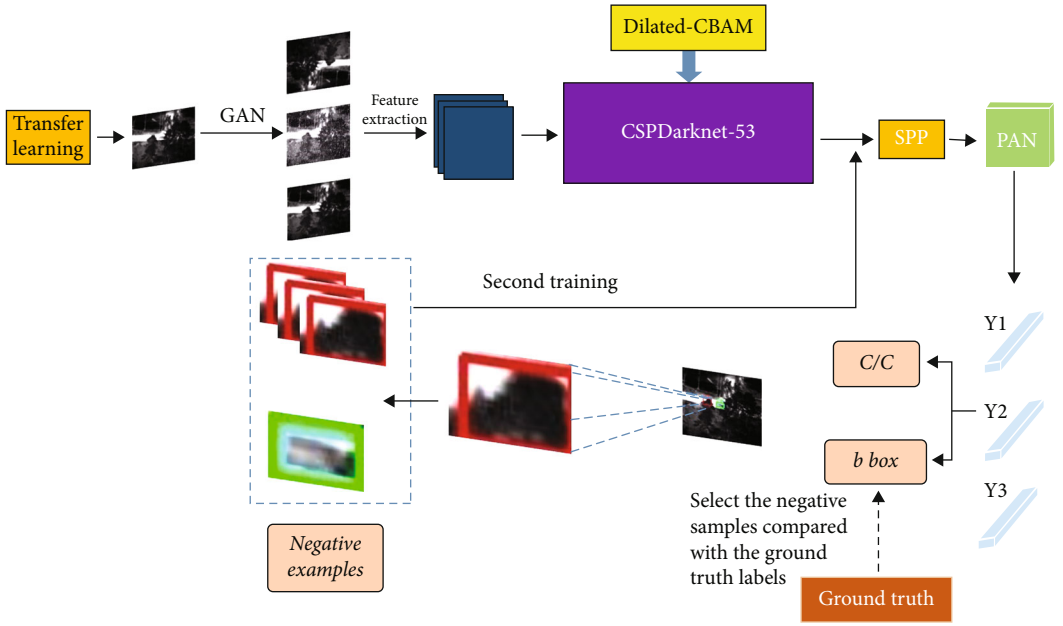

Figure 2: The whole procedure of the FA-YOLO model. The data augmentation adopts transfer learning and GAN, the backbone uses an attention mechanism, and the training procedure adds a negative example focusing mechanism.

vehicle images; and through flipping and adding noise, we augment them to 3060 images.

Then, we use the CycleGAN network to transfer the VIVID-visible images to infrared images, as shown in Figure 3. Overall, the final infrared dataset contains 500 images for training from the VIVID-infrared dataset generated by CycleGAN.

3.2. Negative Sample Focusing. As shown in Figure 1, the vehicles in VIVID-infrared images selected by us are heavily impacted by environment; the features of vehicles are mixed with the confusing background information, which is even difficult for human eyes to recognize. The complex background information may interfere with the detection model by causing too much false-positive examples. To mitigate the impact of the background information and depress the damage of the false positives, herein, we revise the YOLOv4 model with a negative sample focusing mechanism which could focus on training the confusing negative samples and distinguish the targets from the complex background.

After the NMS of the YOLOv4 model, the YOLO-head layer outputs several predicted boxes with location parameters (b_x, b_y, b_h, and b_w) and class possibilities c. Through calculation, we could gain the IoU of each predicted box towards the corresponding target box. In general, as for each predicted box, when the IoU > 0.7, prediction is corrected; otherwise, it should have been recognized as the background but was falsely predicted as the targets, that is to say, the negative samples. When doing a detection task, there would be so much negative samples that impact the performance of the model.

$$D_n = \left\{ c_i \mid \mathrm{oU}_i < 0.7, c_i > c_j, 1 \le i, j \le N \right\}. \tag{1}$$

In consequence, we need to revise the model, and let it focus more on such negative samples. As shown in equation (1), select the predicted boxes, of which the IoU < 0.7 into D; these are the negative samples. Figure 4 shows the negative sample focusing mechanism in the FA-YOLO model. In the training procedure, every time when doing backpropagation, the model gets the four location parameters (b_x, b_y, b_h, and b_w) of the false positives (FP) and uses the location parameters to map the FPs to the corresponding area in

Figure 3: Infrared images generated by CycleGAN. The first row is the visible images, and the second row is the generated infrared images.

the layers before the multiheads (as shown in the red areas). In theory, the locations in different layers have a congruent relationship through the convolution operation, and we could use reverse convolution operation to find the location relationship between the shallow layer and the deep layer. Then, we transfer them to the corresponding locations in the feature map output by the CSPDarknet53 and optimize the model with these samples again.

Every time after normally training for m epochs, select the first n samples in dataset D_n, find the negative samples and their corresponding feature maps output by the backbone, put them into the forward-propagation operation, and optimize the loss values of the NS. When doing the negative sample training optimization, to make the parameter optimize faster, we freeze the backbone parameters and just upgrade the subsequent parameters.

3.3. Dilated CBAM. Given the problem that infrared vehicle targets are small and the features are not obvious from the background, it is not easy for the model to extract and conserve the features. In this way, the attention mechanism, channel attention and spatial attention, is added to the YOLOv4 network to enhance the small targets, making the key features distinguishable.

Our attention contains both channel attention and spatial attention, given the input feature map $F \in \mathbb{R}^{C\times H\times W}$ from the upper layer, and the dilated CBAM sequentially generates a 1D channel attention map $M_c \in \mathbb{R}^{C\times 1\times 1}$ and a 2D spatial attention map $M_s \in \mathbb{R}^{1\times H\times W}$ as illustrated in Figure 5. The overall attention process can be summarized as

$$\begin{aligned} F' &= M_c(F) \otimes F, \\ F'' &= M_s\left(F'\right) \otimes F'. \end{aligned} \tag{2}$$

3.3.1. Channel Attention. In channel attention, we use the module from CBAM [21], which aggregates the spatial information of a feature by using both average pooling and max pooling, generating two different spatial context descriptors: F^c_{avg} and F^c_{max}. Both of the two descriptors are forwarded to a multilayer perception (MLP) to generate a different channel attention map and then added and activated by the sigmoid function to the final channel attention map. The channel attention is computed as

$$\begin{aligned} M_c(F) &= \sigma(\text{MLP}(\text{Avgpool}(F)) + \text{MLP}(\text{Maxpool}(F))) \\ &= \sigma\left(W_1\left(ReLU\left(W_0\left(F^c_{\text{avg}}\right)\right)\right) + W_1(ReLU(W_0(F^c_{\text{max}})))\right). \end{aligned} \tag{3}$$

3.3.2. Spatial Attention. In spatial attention, we change the convolutional layer in CBAM into a dilated convolution kernel to increase the receptive field so as to link the information of the targets and the background. However, Yu et al. [27] point out that dilated convolutions can cause gridding artifacts, which often occur when a feature map has higher frequency content than the sampling rate of the dilated convolution. To remove the gridding artifacts, we add two more dilated convolutional kernels with smaller dilated rates after the first dilated one with a dilated rate of 4, as shown in the first row in Figure 5.

Firstly, apply the average pooling and max pooling operation along the channel axis and concatenate them to generate an efficient feature descriptor, $F^c_{\text{avg}} + F^c_{\text{max}} \in \mathbb{R}^{2\times H\times W}$. Then, put the descriptor forward to the $M_s(F)$ to generate the spatial attention. The $M_s(F)$ is composed of three dilated convolution layers, i.e., 3×3 kernels with dilated rates of 4, 2 and 1, respectively. In short, the spatial attention is computed as

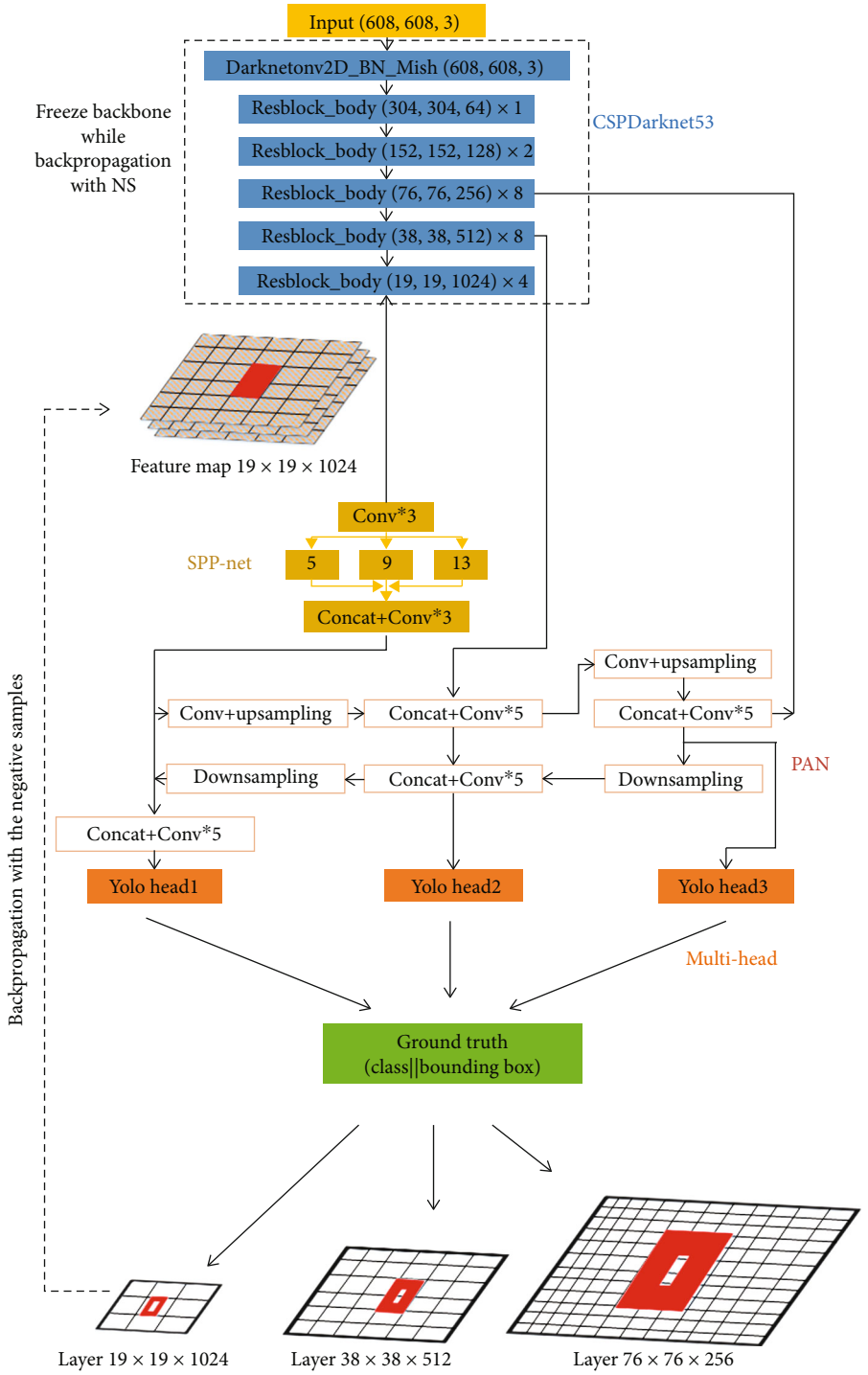

Figure 4: The negative sample focusing mechanism in the FA-YOLO model. The red parts are the negative samples mapped into the feature map.

$$M_s(F) = \sigma\left(f_4^{3\times3}\left(f_2^{3\times3}\left(f_1^{3\times3}(\text{Avgpool}(F); \text{Maxpool}(F))\right)\right)\right) = \sigma\left(f_4^{3\times3}\left(f_2^{3\times3}\left(f_1^{3\times3}\left(F_{\text{avg}}^s; F_{\text{max}}^s\right)\right)\right)\right). \tag{4}$$

The CSPDarknet53 has $1 + 2 + 8 + 8 + 4 = 23$ residual blocks, we plug the dilated CBAM after each block, thus getting an attention-based CSPDarknet53 feature extraction network, and each residual block with the dilated CBAM is a new basic unit of the attention-based CSPDarknet53.

4. Experiment

4.1. Dataset and Environment. The pretraining dataset is UCAS-AOD visible dataset, with a total of 3060 images.

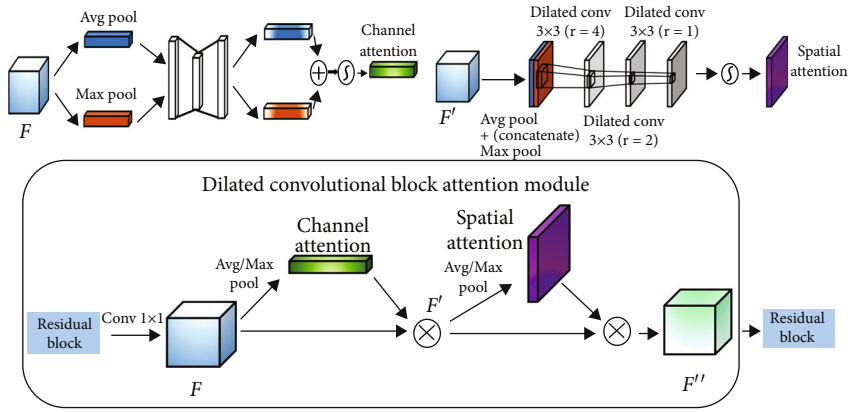

Figure 5: Dilated convolutional block attention module in CSPDarknet53. The first row is the channel attention and spatial attention, respectively, and the second row is the whole structure of the dilated CBAM plugged into the CSPDarknet53.

Table 1: Result comparisons between different models.

Model	Precision (%)	Recall (%)	AP_{50} (%)	F1 (%)	FPS
SSD-VGG16	78.11	74.06	76.95	76.03	59.24
YOLOv3	84.24	80.66	84.53	82.41	40.02
Faster R-CNN-Resnet	70.76	90.74	88.30	79.51	9.35
YOLOv4 (no transfer)	82.20	74.06	79.24	77.92	40.72
YOLOv4 (transfer learning)	89.22	85.85	90.34	87.50	41.86
YOLOv4+NSF	96.21	81.06	91.92	87.98	40.74
YOLOv4+NSF+dilated CBAM	98.11	80.00	92.95	88.13	35.61

The final infrared datasets contain 500 images from the VIVID generated from CycleGAN and were manually annotated by labelImg. The testing dataset contains 100 infrared images from the VIVID-infrared dataset, and the vehicle in each image is heavily occluded and impacted by the confusing background information. During the experiments, the GPU is RTX 2080Ti.

4.2. Comparison Experiments. To verify the superiority of the FA-YOLO, we launch extensive comparison experiments. The SSD, YOLOv3, faster R-CNN, and original YOLOv4 model are put on the dataset for training and testing. Furthermore, we also launch an experiment to verify the efficiency of transfer learning. Based on the YOLOv4 model, we used no transfer learning as the comparisons and just train the model on the infrared dataset.

4.3. FA-YOLO Experiments. Finally, we apply the negative sample focusing mechanism and dilated CBAM to the YOLOv4 model sequently. For the negative sample focusing mechanism, each time when normal training for 9 times, select the first 120 negative samples for one time focusing training. As for the dilated CBAM, we add the module to the CSPDarknet53, since the structure has changed and we first train the model in VOC-2007 for 1,000 epochs and get the weight file. Then, we keep all the procedures and parameters consistent with those in the original experiments.

4.4. Experiment Results. The experiment results are shown in Table 1. The mean average precision (mAP_{50}) and F1 score are adopted as the metrics of the detection accuracy, as shown in the following equations:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{6}$$

$$\text{F1 score} = 2\frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}. \tag{7}$$

It could be concluded from the table and the *P-R* curve in Figure 6 that our FA-YOLOv4 has the highest mAP_{50} and F1 score among all the other models. When using a transfer learning strategy, the mAP_{50} improves by 11.1% and the F1 score improves by 9.58%; when using the negative sample focusing mechanism, the mAP_{50} improves by 12.68% and the F1 score increases by 10.06%. When adding the dilated CBAM to the YOLOv4, the mAP_{50} improves by 13.71% to 92.95% and the F1 score improves by 10.21% to 88.13%.

Figure 7 shows the part of the detected images on the testing set of FA-YOLO, from which we draw the conclusion that the attention module could detect the small, weak, and occluded targets well. Figure 8 is the heat map—the

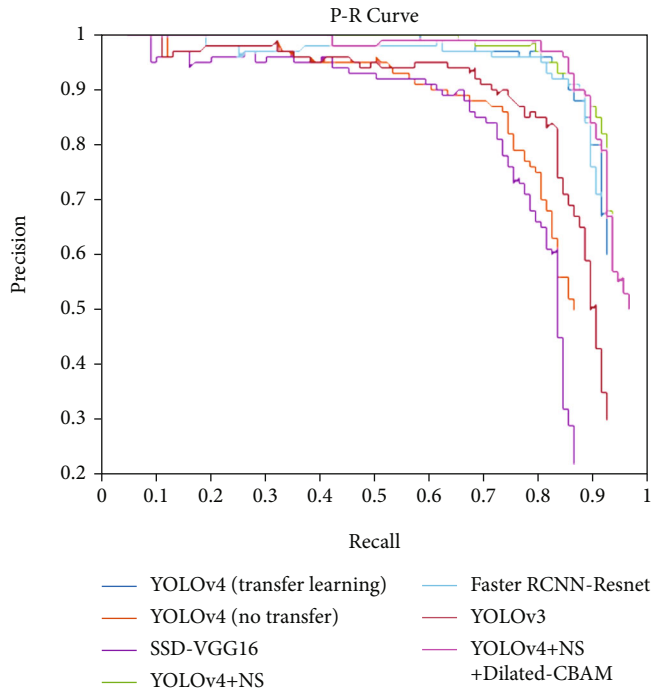

Figure 6: The precision-recall curves of different detection models in this experiment.

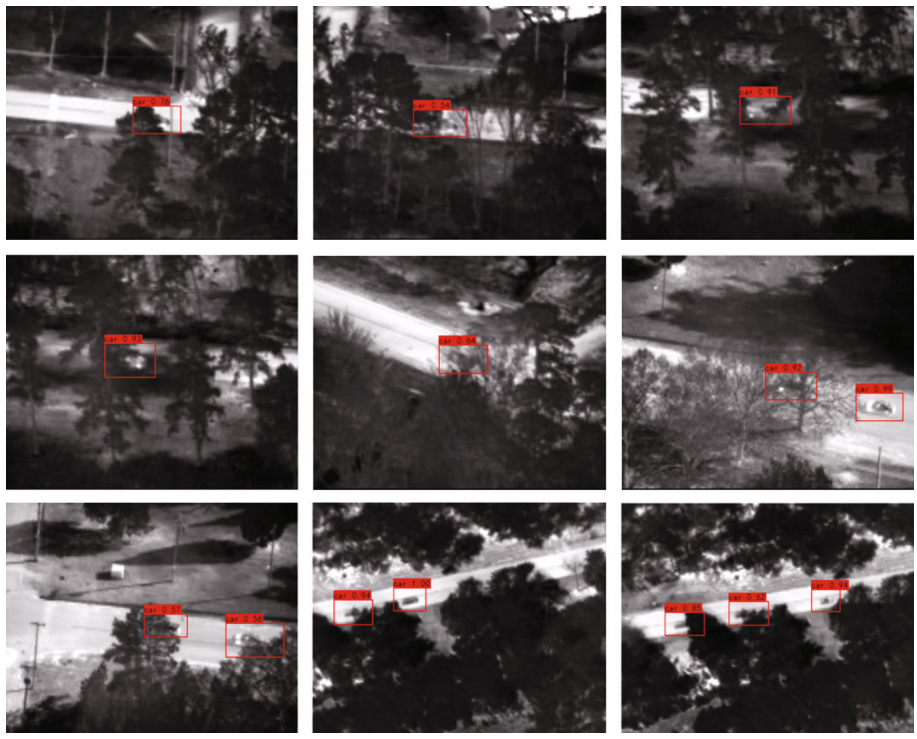

Figure 7: The detected images on the testing dataset of FA-YOLO.

explanation of the CSPDarknet53 with the dilated CBAM; we use Grad-CAM [28] to visualize the output of the backbone when inputting an image. The attention module could focus on the target information and filter the background information well in most targets, but there is still some confusing background information which may mislead the detection model.

Figure 9 shows the comparison detection results of SSD, faster R-CNN, and FA-YOLO on testing images. The blue boxes are the ground truth boxes (GT), the green boxes are

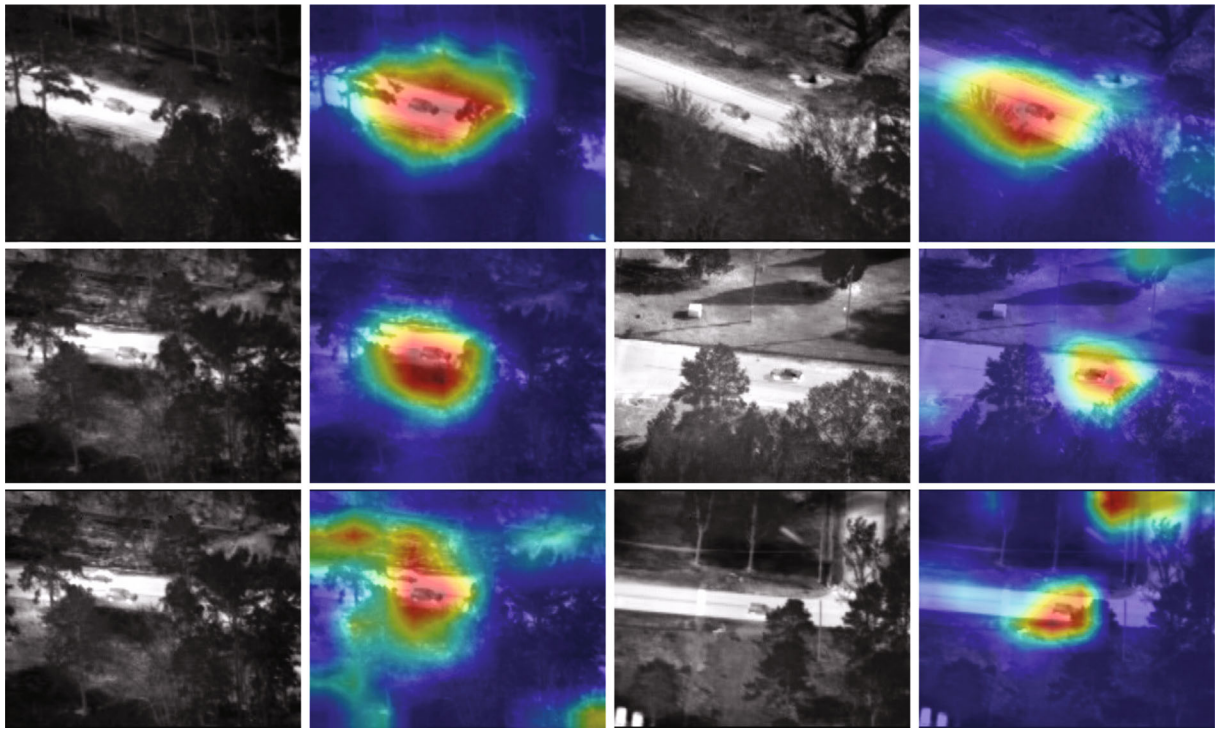

Figure 8: The dilated CBAM explanation, of which the red parts mean the key importance to the detection task.

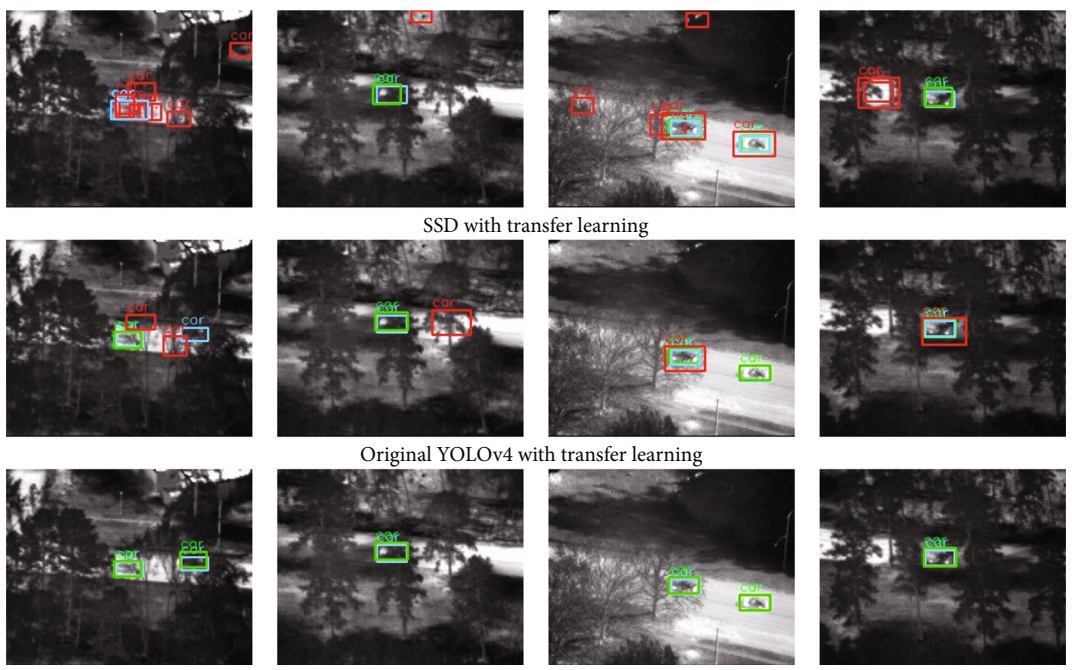

Figure 9: The result comparison of the three models. The blue box means the GT, the green box means the TP, and the red box means the FP—negative samples. The FA-YOLO could efficiently depress the negative samples.

the true positive samples detected by the model (TP), and the red boxes are the background information falsely recognized as positive samples by the models (FP), in other words, the negative samples. From formulas (5), (6), and (7), the FPs will decrease the detection accuracy and the TPs are what we really need. What is more, the comparison of the three row images indicates that the baseline models could not distinguish the confusing background information correctly, while the FA-YOLO could solve this problem well.

5. Conclusion

In our paper, the FA-YOLO model is proposed to the application of infrared occlusion vehicle detection in wild complex background, where the confusing background information causes great impact on the target detection. By using GAN and transfer learning, our model has a sufficient dataset for training and optimization. By using the negative sample focusing mechanism during the training procedure, it could mitigate the complex background information and occlusion influences, thus making the model more accurate for distinguishing the targets and the background. Finally, by plugging the attention mechanism module into CSPDarknet53, the YOLOv4 could enhance the features of small targets so as to improve the detection accuracy. Through extensive experimental verification and comparison, the detection accuracy-mAP_{50} on the VIVID-infrared occluded vehicle improves by 13.71% and the F1 score increases by 10.21%, which shows a significant improvement of our method and superiority of the proposed model.

Data Availability

The [experiment results and algorithm codes] data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] Q. Chen and A. Hamdulla, "Summary about detection and tracking of infrared small targets," in *2019 12th International Conference on Intelligent Computation Technology and Automation (ICICTA)*, pp. 250–253, Xiangtan, China, 2019.

[2] K. Wang, S. Li, S. Niu, and K. Zhang, "Detection of infrared small targets using feature fusion convolutional network," *IEEE Access*, vol. 7, pp. 146081–146092, 2019.

[3] D. Xu and Y. Wu, "Improved YOLO-V3 with DenseNet for multiscale remote sensing target detection," *Sensors*, vol. 20, no. 15, p. 4276, 2020.

[4] X. Wang, Y. Ban, H. Guo, and L. Hong, "Deep learning model for target detection in remote sensing images fusing multilevel features," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 250–253, Yokohama, Japan, Jul. 2019.

[5] P. Wang, W. Wang, and H. Wang, "Infrared unmanned aerial vehicle targets detection based on multi-scale filtering and feature fusion," in *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, pp. 1746–1750, Chengdu, China, Dec. 2017.

[6] J. Hu, Y. Zhao, and X. Zhang, "Application of transfer learning in infrared pedestrian detection," in *2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC)*, Beijing, China, Jul. 2020.

[7] X. Zhang and X. Zhu, "Vehicle detection in the aerial infrared images via an improved YOLOV3 network," in *2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP)*, Wuxi, China, 2019.

[8] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, Venice, Italy, 2017.

[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. E. Reed, "SSD: single shot multibox detector," in *European Conference on Computer Vision*, Springer, Cham, 2015.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[11] J. Redmon and A. Farhadi, "YOLOv3: an incremental improvement," 2018, http://arxiv.org/abs/1804.02767.

[12] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: optimal speed and accuracy of object detection," 2020, http://arxiv.org/abs/2004.10934.

[13] Y. Song, M. Li, X. Qiu, W. Du, and J. Feng, "Full-time infrared feature pedestrian detection based on CSP network," in *2020 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*, pp. 516–518, Vientiane, Laos, Jan. 2020.

[14] Y. Wang and X. Bai, "Intensity inhomogeneity suppressed fuzzy C means for infrared pedestrian segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 9, pp. 3361–3374, 2019.

[15] Y.-Y. Chen, S.-Y. Jhong, G.-Y. Li, and P.-H. Chen, "Thermal-based pedestrian detection using faster R-CNN and region decomposition branch," in *2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, Taipei, Taiwan, Dec. 2019.

[16] X. X. Zhang and X. Zhu, "Moving vehicle detection in aerial infrared image sequences via fast image registration and improved YOLOv3 network," *International Journal of Remote Sensing*, vol. 41, no. 11, pp. 4312–4335, 2020.

[17] G. Zheng, X. Wu, Y. Hu, and X. Liu, "Object detection for low-resolution infrared image in land battlefield based on deep learning," in *2019 Chinese Control Conference (CCC)*, pp. 8649–8652, Guangzhou, China, Jul. 2019.

[18] Y. M. Kassim, M. E. Byrne, C. Burch et al., "Small object bird detection in infrared drone videos using mask R-CNN deep learning," *Electronic Imaging*, vol. 2020, no. 8, 2020.

[19] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: a survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 5, pp. 1019–1034, 2015.

[20] X. Zhang, K. Zhu, G. Chen et al., "Geospatial object detection on high resolution remote sensing imagery based on double multi-scale feature pyramid network," *Remote Sensing*, vol. 11, no. 7, p. 755, 2019.

[21] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," in *Computer Vision – ECCV 2018*, pp. 3–19, Springer, 2018.

[22] J. Park, S. Woo, J. Y. Lee, and I. S. Kweon, "Bam: bottleneck attention module," 2018, http://arxiv.org/abs/1807.06514.

[23] J. Sun, H. Ge, and Z. Zhang, "AS-YOLO: an improved YOLOv4 based on attention mechanism and SqueezeNet for person detection," in *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pp. 1451–1456, Chongqing, China, 2021.

[24] C. Gao, Q. Cai, and S. Ming, "YOLOv4 object detection algorithm with efficient channel attention mechanism," in *2020 5th International Conference on Mechanical, Control and*

Computer Engineering (ICMCCE), pp. 1764–1770, Harbin, China, 2020.

[25] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: efficient channel attention for deep convolutional neural networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11531–11539, Seattle, WA, USA, 2020.

[26] C. Chen, C. He, C. Hu, H. Pei, and L. Jiao, "A deep neural network based on an attention mechanism for SAR ship detection in multiscale and complex scenarios," *IEEE Access*, vol. 7, pp. 104848–104863, 2019.

[27] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 636–644, Honolulu, HI, USA, 2017.

[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, Venice, Italy, 2017.

[29] Y. Wang, L. Wang, Y. Jiang, and T. Li, "Detection of self-build data set based on YOLOv4 network," in *2020 IEEE 3rd International Conference on Information Systems and Computer Aided Education (ICISCAE)*, pp. 640–642, Dalian, China, Sep. 2020.

[30] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation robust object detection in aerial images using deep convolutional neural network," in *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 3735–3739, Quebec City, QC, Canada, Sep. 2015.

[31] R. Collins, X. H. Zhou, and S. Keat, "An open source tracking testbed and evaluation web site," in *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, pp. 35–42, Piscataway, NJ, USA: IEEE Press, Jun. 2005.

[32] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.

[33] D. Zhao, L. Gu, K. Qian, H. Zhou, T. Yang, and K. Cheng, "Target tracking from infrared imagery via an improved appearance model," *Infrared Physics and Technology*, vol. 104, 2020.

[34] L. Kang, Y. Zhang, B. Zou, and C. Wang, "High-resolution PolSAR image interpretation based on human images cognition mechanism," in *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 1849–1852, Milan, Italy, 2015.

[35] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, Las Vegas, NV, USA, Jun. 2016.