IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 17, 2021, accepted January 29, 2021, date of publication February 8, 2021, date of current version February 16, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3057723*

# Weak and Occluded Vehicle Detection in Complex Infrared Environment Based on Improved YOLOv4

**SHUANGJIANG DU[1], PIN ZHANG[2], BAOFU ZHANG[1], AND HONGHUI XU[2]**
[1]College of Communications Engineering, Army Engineering University of PLA, Nanjing 210007, China
[2]College of Field Engineering, Army Engineering University of PLA, Nanjing 210007, China

Corresponding author: Pin Zhang (pinzhangthree@sina.com)

This work was supported in part by the National Nature Science Foundation of China under Grant 61371121.

**ABSTRACT** Infrared small target detection is still a challenge in the field of object detection. At present, although there are many related research achievements, it surely needs further improvement. This paper introduced a new application of severely occluded vehicle detection in the complex wild background of weak infrared camera aerial images, in which more than 50% area of the vehicles are occluded. We used YOLOv4 as the detection model. By applying secondary transfer learning from visible dataset to infrared dataset, the model could gain a good average precision (AP). Firstly, we trained the model in the UCAS_AOD visible dataset, then, we transferred it to the VIVID visible dataset, finally we transferred the model to the VIVID infrared dataset for a second training. Meanwhile, added the hard negative example mining block to the YOLOv4 model, which could depress the disturbance of complex background thus further decrease the false detecting rate. Through experiments the average precision improved from90.34% to 91.92%, the F1 score improved from 87.5% to 87.98%, which demonstrated that the proposed algorithm generated satisfactory and competitive vehicle detection results.

**INDEX TERMS** Infrared aerial image, occlusion, vehicle detection, hard negative example mining, YOLOv4.

## I. INTRODUCTION

Object detection technology has been very mature, and it has been widely used in many aspects, among which it has reached the peak in the detection of traditional images. However, when considering some specific application scenarios, such as the dim, weak and severely occluded vehicle detection in the infrared images under a complex background, the object detection technology still needs to be improved. The current researches focus much more on the detection of small infrared objects but ignore the impacts of the complex backgrounds, which is also a challenge for detector, thus this paper studies on detection of the infrared objects occluded and impacted by the complex backgrounds.

Firstly, there are some inherent defects in infrared camera images, for instance, infrared imaging is subject to imaging distance, angle of view and the change of the light, moreover it is easily disturbed by the atmospheric radiation and occlusion of objects in transit of light, thus, the imaging effect is not stable enough [11]. Much noise, low contrast and indistinct boundary between target and background also make the detection of infrared images much harder than that of the normal datasets such as ImageNet and MS COCO.

Secondly, for infrared long-range aerial image, the image size of the target is smaller and the resolution is much smaller than the normal image usually with an average pixels of $30 \times 30$ [12]. Small scale object detection is a hot and challenging task in the field of object detection. For small scale object, the main methods are feature fusing [14] and multi scale fusing [15].

In particular, the object detection in the complex backgrounds studied in this paper, under the interference of small scales and noise, coupled with the occlusion of trees in the backgrounds and other issues, constitutes a great disturbance to the detection of the models. The non-object features in the background may confuse the detector and mislead it to make a false decision, resulting in a high false detection rate (a low precision). However, it is of great practical significance to study how to improve the detection accuracy of the detector so that it can still have a good performance on severely occluded targets in complex environments. Especially in military target

The associate editor coordinating the review of this manuscript and approving it for publication was Bo Pu.

detection, it can liberate human labor force from massive images to be recognized.

Thirdly, it is a common problem that infrared remote sensing datasets with labels are insufficient. There are a huge dataset of visible images while the infrared image datasets are relatively small, which causes a trouble for the training procedure of infrared object detection.

To solve these problems, firstly, this paper proposed the strategy of ''secondary transfer learning''. Considering the visible remote sensing images and infrared camera images have largely consistency in the content, we first get the pre-trained YOLOv4 weights–*yolov4.weights* on MSCOCO dataset and put it on UCAS_AOD dataset [8] for the initial training, thus we gain the weights trained for the first time. Then we transfer the weights to VIVID [9] visible aerial image dataset for the first transfer learning to get the first fine-tuning of weights. Next we put the weights on VIVID infrared aerial image dataset for the second transfer learning to get a second fine-tuning weight after training. After two stages of transfer learning, the weight parameters can be adjusted step by step to achieve the desired requirements, so as to make up for the lack of infrared image dataset.

In view of the high false detection rate caused by the complex background, we proposed to add hard negative example mining block to the YOLOv4 model. Meanwhile, in order to balance the positive samples and negative samples, the positive samples and the negative samples were added at a ratio of approximately 1:3 for secondary training. Finally, the detection accuracy of YOLOv4 model was improved from 88.71% to 91.59%, indicating that the improved network model could meet the expected requirements.

The main contributions of our work are as following:

- The method of ''secondary transfer learning'' is proposed to solve the problem of insufficient dataset
- The more challenging scene of heavily occluded objects detection with complex background is studied
- Hard negative example mining block is added to the YOLOv4 model to improve the detection accuracy of the model

The content of the article is arranged as follows:

- The second section gives a literature review of relevant researches and introduces the general research status of relevant fields.
- The third section discusses in detail how to improve the YOLOv4 model and the operation flow of data processing.
- The fourth section presents the experimental steps and results
- The fifth section summarizes and forecasts the full researches

## II. LITERATURE REVIEW

### A. INFRARED OBJECT DETECTION

For infrared object detection, the common infrared targets are infrared pedestrian detection [1], [20]–[22], infrared vehicle and aircraft detection [3], [4], [5], [23]. The main problem to be overcome is the lack of infrared data, followed by the problem of unclear infrared image features.

Transfer learning [24] is an effective method to solve the problem of insufficient infrared image datasets. Its possibility lies in the great similarity in shape between targets in infrared images and visible images, so we can firstly put model in a large number of visible image datasets for training, and then transfer it to a small amount of infrared datasets for fine-tuning. The use of transfer learning should meet two conditions: firstly, the target difference between the images of the datasets in the two training stages should not be too large, and there should be similarity; secondly, the number of images of the datasets in the first pre-training stage should be much larger than that of the datasets in the second training stage; otherwise, the purpose of transfer learning cannot meet the expected requirements. In our experiments, the objects in infrared images and the visible images have the similarities in features. They are all the remote sensing images, with the characteristics of remote sensing images, that is to say, small pixels, indistinct boundaries between the objects and environment and complex backgrounds.

Zhang and Zhu [3] compensates for the lack of infrared data set through transfer learning. They first train the model on the VIVID visible light dataset, and then fine-tune their parameters on the VIVID infrared dataset. Hu *et al.* [1] detect the pedestrian in the infrared images, use visible light images in the CVC dataset for pre-training, and then transfer to the CVC infrared dataset for fine-tuning. Moreover data augmentation technique can be used for the expansion of infrared datasets, on the basis of the original data by rotating, flipping, adding noise [3] to double the number of datasets, GAN [1] can also be applied to generate the infrared datasets of different style, or convert visible image to infrared image.

### B. AERIAL IMAGE DETECTION

Aerial images are characterized by small resolution, small scale, indistinct features and background interference. Wang *et al.* [2] proposed a special feature extraction network, MNET, for the detection of aircraft with minimal resolution ($2 \times 2$ pixels) under the background of sea and sky. This network uses feature fusion, which fuses different feature extraction layers to retain small-scale information, and introduces attention mechanism into feature maps to highlight small-scale features. For vehicle detection, Zhang and Zhu [3], [4] improved the YOLOv3 model by reducing the numbers of network layers of feature extraction as well as transfer learning, and calculated the appropriate sizes of anchor boxes through k clustering algorithm, so as to detect low-resolution vehicles in infrared aerial images and then track them.

Zheng *et al.* [5] used YOLOv3 to detect low-resolution infrared vehicles in land battlefield, but its target were not aerial images. Kassim *et al.* [6] used the MASK-RCNN network and added the Data Association and Filtering (DAF) module to the network post-processing module. The principle is to distinguish real targets and background interference by

**TABLE 1. Comparisons of previous works.**

| Method | Pros | Cons | References |
|---|---|---|---|
| Transfer learning | The model structure is simple and clear, easy to training | Only useful for objects of large scales in simple background | [1],[3-5],[24] |
| Feature fusion+ attention mechanism | Extract features of very small objects(2×2 pixels) | Only useful for objects in simple background such as sky and sea | [2],[6] |
| Density map | Detect the small and high density objects | Not useful for single object detection | [16] |
| Extra information | Has the ability of reasoning and predict the occluded parts | Detection on traditional visible datasets | [17-19] |
| Ours (secondary transfer learning + HNEM) | Decrease the impact of complex backgrounds and mitigate the lack of infrared datasets | | |

comparing the differences of targets in several consecutive frames of images before and after, so as to count birds in the continuous images shot by infrared camera. In literature [16], Density Map method was proposed to guide the detector to detect the vehicles in the aerial images. The basic idea is to firstly use the Density Map to calculate the density of the vehicles in the images, then, the image was segmented into several small pieces of different sizes according to the density map of the vehicles, and each piece of the image was detected separately with different anchors and intensity. This method can make the detector to detect different areas in the image specifically, and improve the detection efficiency and accuracy.

### C. OCCLUSION DETECTION

At present, most researches on the recognition of occluded images mainly focus on large scale traditional objects of visible images, such as pedestrian detection [17], faces recognition [18], cars [19] and other occluded objects detection. In [17], a model based on Deformable Part Network (DPN) was proposed, which has the ability of reasoning and can predict the occluded parts. Guo *et al.* [18] proposed the Adaboost Cascade Classifier algorithm based on the Haar-like feature is applied. Firstly, the cascade classifier is applied to identify the eyes and mouth of a human face; then, the physical feature relationship between the eyes and mouth of a person and the face is used to identify and detect the entire masked face. In the context of urban autonomous driving system, vehicles are blocked due to buildings, dynamic changes and limited perspective. In [19], Faster-RCNN network and the Part-Aware Region Proposal Network were used to perceive the partial or global visual information of a vehicle. By extracting and encoding matching the global and local information of a vehicle, the two feature frames were integrated, so that the model could reduce the influence caused by occlusion.

At present, the most solutions to the occlusion detection problem is aimed at the images with high resolutions, large the scales, and a distinct boundary between the backgrounds and the objects. However, there is still room for further research on the detection of weak infrared small targets under severe occlusion and complex background. Meanwhile, the application scenario is more in line with the actual requirements. Therefore, this paper proposes a research point of detecting the severely occluded vehicle target in infrared camera aerial image. The YOLOv4 [7], [13] model was used as the basic detection model, and corresponding improvements were made on the basis of it.

The comparisons of previous works and our proposed methods are shown in TABLE 1.

## III. METHODOLOGY

### A. SECONDARY TRANSFER LEARNING

Considering the small amount of infrared dataset available in this experiment, transfer learning is adopted to make up for it. Different from the method mentioned above, we added the transfer learning for twice. First, the model was trained on the visible remote sensing dataset, and then it was trained on a small amount of infrared dataset. The transfer learning aims to learn to extract the features of the object due to the large amount of visible image datasets.

As for visible image datasets, we used UCAS_AOD visible dataset, which includes two objects, automobile and aircraft. We extracted the dataset of automobile detection, a total of 510 images, rotating, flipping and adding Gaussian noise on the basis, thus expanding the datasets to 3060 pieces.

Secondly, 320 aerial images from VIVID dataset were selected, manually annotated with *labelImg*, and then expanded to 960 images by rotating and flipping for transfer learning. Finally, 310 infrared aerial images from VIVID dataset were selected, manually annotated with *labelImg* for secondary migration and learning. As shown in figure 2, the first row shows the process of augmentation of UCAS_AOD dataset, one original image is expanded to six pieces through horizontal flipping, vertical flipping and adding Gaussian noise. The second row demonstrates the three different datasets. As can be seen from the figures, the vehicle targets in the three images are relatively small, and all are aerial images taken by remote sensing. Therefore, this provides a prerequisite for transfer learning. Figure (d) is the infrared vehicle image with severe occlusion. It is difficult to find the occluded target only through rough observation of human eyes.

The whole procedure is shown is figure 3.

First, get the official weights of YOLOv4 model – *yolov4.weights* which was trained in MS COCO dataset. MS COCO contains more than 200,000 images and

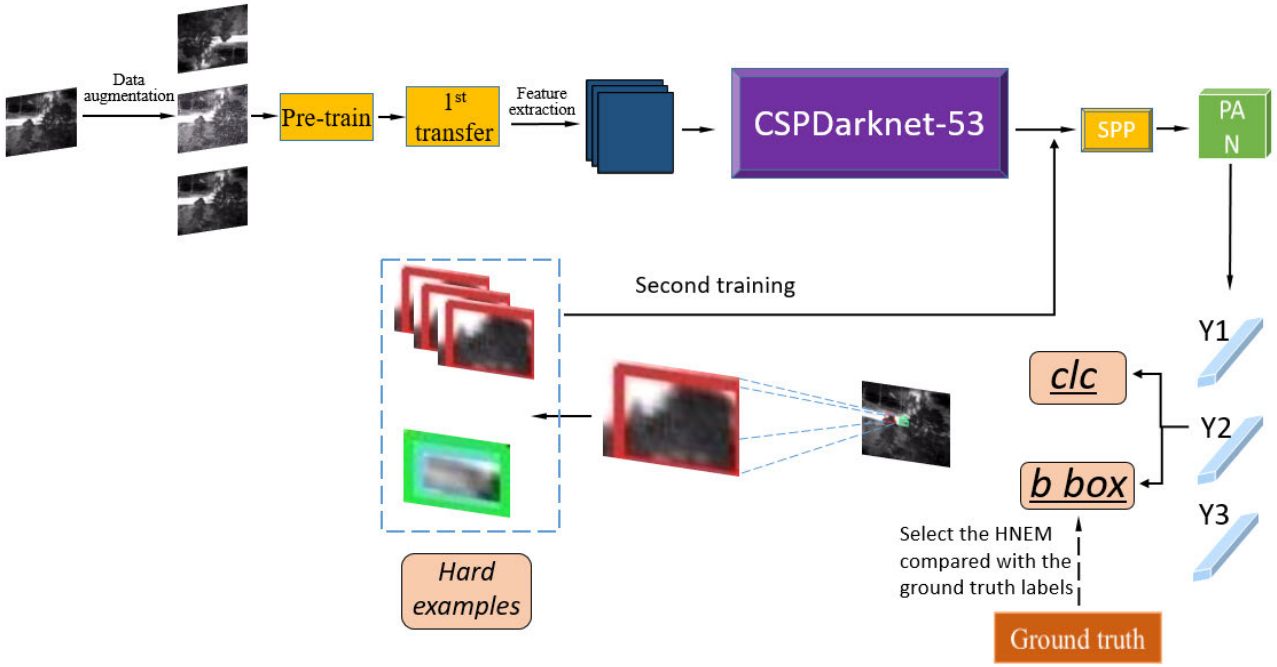



**FIGURE 1.** The whole procedure of the YOLOv4 model with HNEM optimizer.

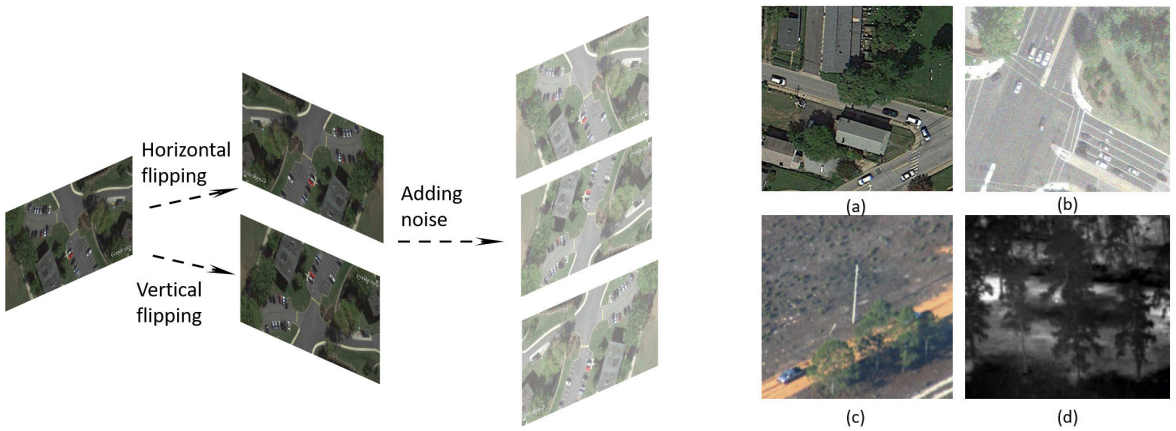



**FIGURE 2.** The left row shows augmentation process of UCAS_AOD visible remote sensing dataset through flipping and adding noise. The right row shows images of three datasets, which are (a) UCAS_AOD dataset, (b) UCAS_AOD dataset with noise added, (c) VIVID visible remote sensing dataset, and (d) VIVID infrared remote sensing dataset.

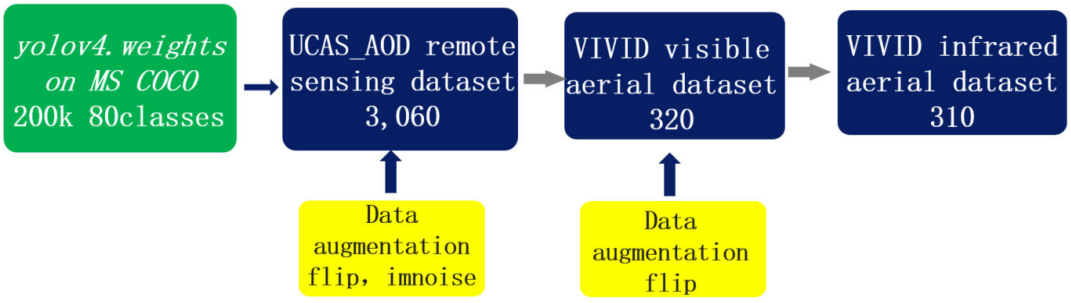



**FIGURE 3.** The secondary transfer learning takes place firstly from UCVS_AOD to VIVID visible dataset and again from VIVID visible dataset to infrared dataset.

80 classes. Put it on the UCAS_AOD dataset as the initial weights for pre-training. Then evaluate the models, chose the model weight with the high AP as the initial weight for the next transfer learning. The weights pre-trained are transferred

to the VIVID visible aerial dataset for the first transfer training, and in the same way, place the selected model weight on the infrared aerial dataset for the second transfer training. After two stages of transfer training, the model can get better optimization and approach the optimal solution step by step.

### B. HARD NEGATIVE EXAMPLE MINING

In the infrared image, as shown in figure 2, the vehicle in the image is heavily occluded by trees, and the boundaries between the background and the target is not obvious, resulting in many confusing backgrounds. Considering that the complex backgrounds will interfere with the performance of the detector and increase the rate of false detection, a hard negative example mining module was added at the end of the YOLOv4 model.

The basic idea is to first calculate the class confidence of each bounding boxes $C_1$ predicted by the classifier, and the IOU values $C_2$ between the bounding boxes and the ground truth labels. Generally speaking, the bounding boxes of which the IOU value $C_2 < 0.4$ means a false positive sample (FP), i.e. predicted as an object whereas it is the background information. For these bounding boxes, the higher the class confidence $C_1$ is, the harder the classifier could identify them correctly. And these are what regarded as hard negative examples in this experiment. We sorted these samples in descending order by the confidence value $C_1$, thus got the sample dataset D:

$$\mathrm{D} = \left\{ C_1^i \mid C_2^i < 0.4, C_1^i > C_1^j, 1 \le i, j \le N \right\} \tag{1}$$

$$\mathrm{D}' = \left\{ C_1^i \mid C_2^i > 0.4, C_1^i < C_1^j, 1 \le i, j \le n \right\} \tag{2}$$

As shown in figure 4. The YOLOv4 model have three YOLO Heads of different numbers. Each of the heads are fed with a layer of different scales. Suppose the input scale is $608 \times 608 \times 3$, the three layer before the heads are $19 \times 19 \times 1024$, $38 \times 38 \times 512$ and $76 \times 76 \times 256$ respectively.

When we got the false positive examples (FP), they could be mapped into the layers (as shown in the red areas), then we transferred them into the feature map generated by the backbone. Then we did the forward propagation again, just calculate the loss values of the hard negative examples compared with the ground truths and do the optimization.

At the beginning of the training, the training set was put into the model for training. After trained for M epochs, select the top N samples in dataset D, then put them into feature extraction layers of the model for another epoch of training, which could make the model for further identification of the misleading background and find the differences between the vehicles and the complex environment so as to reduce the false detection rate.

Meanwhile, the balance of the amounts of negative examples and positive examples should be taken into consideration. Otherwise, if the amount of negative examples is much larger than that of the positive examples, the parameters of the model may not converge to a better solution [10]. In this way, while adding the hard negative examples, we got the hard

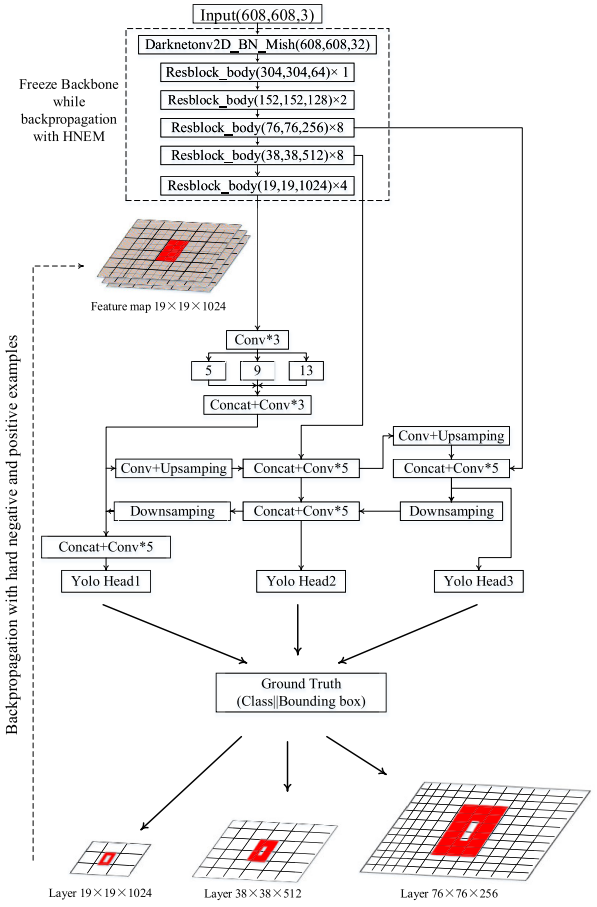



**FIGURE 4.** Add hard negative example mining module to YOLOv4.

**TABLE 2.** Comparison of different versions of YOLO.

| Version | Backbone | If multi-scale | Loss function | mAP | FPS |
|---|---|---|---|---|---|
| YOLO | 24conv+2full | | MSE | 63.4 | 45 (VOC 2007) |
| YOLOv2 | Darknet-19 | | MSE | 78.6 | 40 (VOC 2007) |
| YOLOv3 | Darknet-53 | ✓ | MSE+ Cross Entropy Loss | 31.0 | 35 (MS COCO) |
| YOLOv4 | CSPDarknet-5 | ✓ | CIoU+ Cross Entropy Loss | 41.2 | 38 (MS COCO) |

positive examples dataset $\mathrm{D}'$ in equitation (2), of which the IOU value $C_2 > 0.4$, were put into the model at the same time with a ratio of positive and negative examples $\mathrm{n/N} = 1:3$.

The hard negative and positive examples were put into the training set for the second training. For a better result of the second training with the hard examples, we froze the CSPDarknet-53 layers, the backbone, just fine-tune the parameters of the neck network and head network during the second training.

After several repeated iterations until reaching the pre-set training epochs, the experiment verified the validation of the hard negative example mining module, which the average

**TABLE 3.** Comparisons of 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 RCNN-Resnet** | 70.76 | 90.74 | 88.30 | 79.51 | 9.35 |
| **YOLOv4** (no transfer learning) | 82.20 | 74.06 | 79.24 | 77.92 | 40.72 |
| **YOLOv4** (one transfer learning) | 91.28 | 83.96 | 88.95 | 87.46 | 40.77 |
| **YOLOv4** | 89.22 | 85.85 | 90.34 | 87.50 | 41.86 |
| **YOLOv4+HNEM** | 96.21 | 81.06 | **91.92** | **87.98** | 40.74 |

precision of the model was improved by 1.58% from 90.34% to 91.92%.

### C. MODEL STRUCTURE

The whole optimizing of the improved model is shown in figure 1. We adopted the YOLOv4 as our basic model, because it has many better properties compared with the previous versions of YOLO [25]–[27]. For backbone part, it has a deeper structure than before with Darknet-53 which is much more powerful than Darknet-19 in YOLOv2 but still more efficient than ResNet-101 or ResNet-152 [27]. What's more, in YOLOv4 structure, it adds the cross-stage-partial to Darknet-53 compared with YOLOv3, which can gain a higher accuracy as well as a high speed [13]. The cross-stage connection and the residual part could maintain the small scale features, thus we did not need to change the connection relationship of the original YOLOv4 structure.

Here we listed a table comparing the versions from YOLO to YOLOv4, as shown in TABLE 2. We compared the backbone structures, the loss functions, the mAP and FPS on Pascal VOC 2007 or on MS COCO. YOLOv4 is better in detection accuracy and detection speed. In the publish literatures, YOLOv4 models all gain a much better performance than YOLO and YOLOv2, this is why we chose YOLOv4 model, and we didn't do any comparison experiments on YOLO and YOLOv2.

Before training, we used K-means cluster method to define the sizes of the anchor boxes. We set k = 9, after experiment, the result showed 9 different size of anchor boxes, they were (10, 25), (12, 44), (12, 38), (14, 23), (16, 32), (18, 55), (19, 22), (24, 26), (44, 35), while the pixel size of the image was fixed to 416 × 416.

## IV. EXPERIMENT

### A. EXPERIMENT ENVIRONMENT

The infrared training dataset of this experiment contains 310 images, which were selected from three subsets, pktest01, pkest02 and pkest03 of the VIVID dataset, and were manually marked into YOLO format. The testing dataset contains 101 infrared images from the three VIVID subsets selected out manually that have severe occlusion, and the occlusion rate of the vehicle is more than 40%. There are 256 data tags. The GPU used in the experiment was RTX 2080Ti, the Python version was 3.8, and it was carried out in Pytorch 1.7 environment.

### B. EXPERIMENT STEPS

The whole experiment was carried out in three stages, and the training adjustment was carried out on three datasets respectively.

#### 1) COMPARISON EXPERIMENTS

In the first stage, the comparison models were first used for training and testing. To proof our model's advanced performance, we selected SSD-VGG16, YOLOv3, and Faster RCNN-Resnet model as the baseline model. These models are the typical models of one-stage and two-stage detection models. The training epoch of the model was 120, the initial learning rate was 10e-3, and the learning rate decayed to 0.95 time per epoch.

#### 2) ORIGINAL MODEL EXPERIMENTS

In the second stage, the YOLOv4 original model was used for training and to testing.

To validate the effect of our secondary transfer learning. We also set comparison experiments. Based on the YOLOv4 model, we used no transfer learning and just one transfer learning as comparisons. As for no transfer learning, we put the official weights on the infrared image dataset directly for training. As for the one transfer learning, we firstly put official weights on the UCAS_AOD dataset for training, then transfer the pre-trained weights on the infrared image dataset for training. The initial parameters of the training stage were set as follows: the training epoch was 120. The batch size was 8, using Cosine Annealing Algorithm; the initial learning rate of the first 60 individual epochs was 10e-3, and the initial learning rate of the last 60 epochs was 10e-4; the weight decay was 10e-4. Due to the multi-scale fusion network of the YOLOv4 model itself, its average precision was greatly improved compared with the baseline model.

#### 3) IMPROVED MODEL EXPERIMENTS

In the third stage, hard negative examples mining module was added on the basis of YOLOv4 model. All training parameters remained unchanged, and M = 9 and N = 120 were set in Section 3. The average precision of the model reached

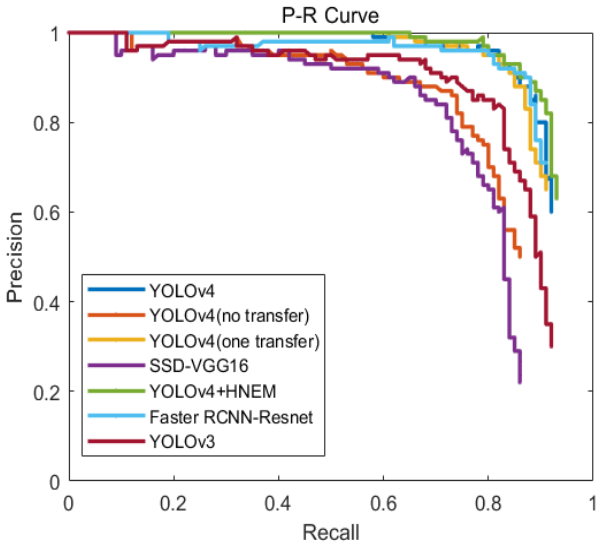



**FIGURE 5.** P-R curves of different models.

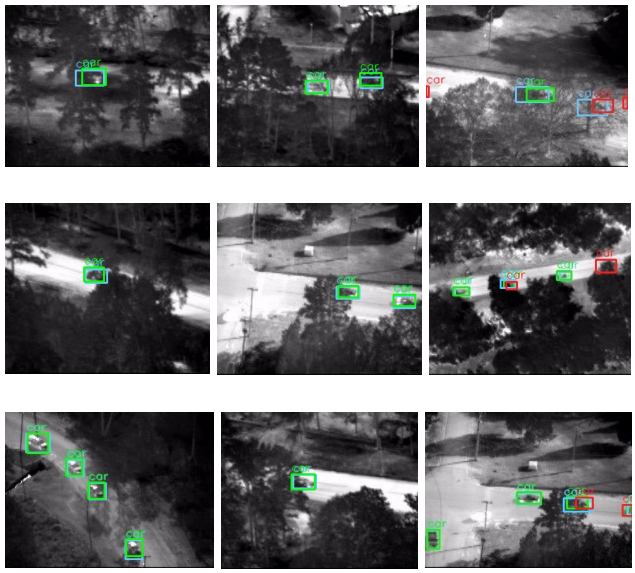


**FIGURE 7.** The part of the detection results of the testing set on YOLOv4 + HNEM.

to 91.92% after the inclusion of hard examples, which improved by 1.58%, and the F1 score improves by 0.48%, indicating that the inclusion of hard negative examples mining module did promote the detection performance of the model.

### C. RESULTS AND ANALYSIS

The precision rate, recall rate, average precision (AP), F1 score, and FPS of different models are shown in TABLE 3. Average Precision is a normal evaluation metric for object detection, but in binary classification, to mitigate the influence of imbalance of positive and negative examples, we also combined the F1 score as a comprehensive evaluation metric, as shown in equitation (3−5). YOLOv4 original model for the testing set has a higher average precision and F1 score compared with the other models, this is due to the superiority of YOLOv4 model structure itself, and the cosine annealing algorithm was used in the process of training enabling the parameter in the process of training to achieve a more optimal solution. When added the hard negative example mining module, the average precision of the model improved by about 1.58%, and the F1 score improved by 0.48%, explaining that the HNEM module does have certain effect to the improvement of model, but in the case of IOU = 0.5, the precision rate significantly improves while the recall rate slightly goes down instead of rising, which illustrates that the model's ability to distinguish the hard negative examples increased but detection of hard positive examples does not change apparently. This reason is caused by comprehensive factors, and is likely to be caused by the imbalance of positive and negative samples in the dataset. Next, it is planned to change the loss function in the YOLOv4 model [11] to balance positive and negative samples and achieve the effect of HNEM.

Figure 5 is the precision-recall curves of all the models, from which the difference in detection accuracy of the models can be intuitively distinguished.

Figure 7 shows the part of the detection results of the testing set on YOLOv4 + HNEM. The blue box represents

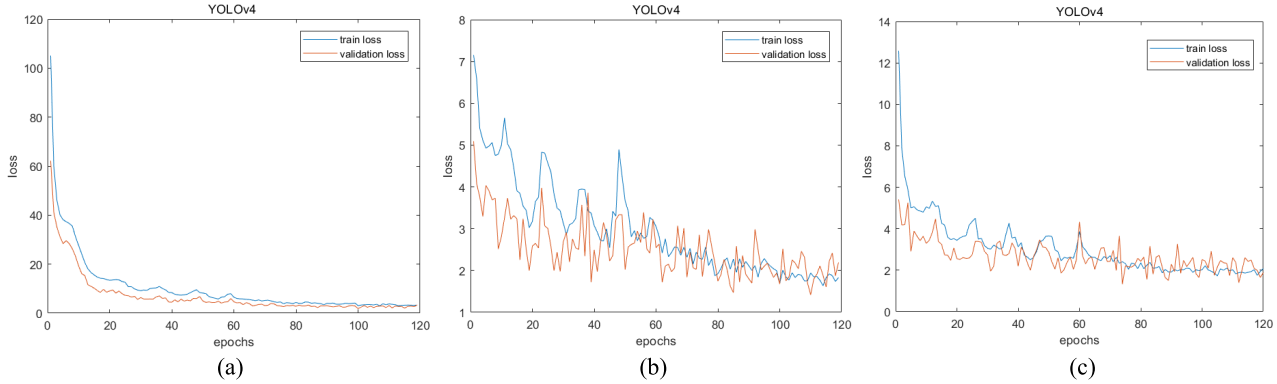



**FIGURE 6.** The train loss and validation loss of YOLOv4 models without transfer learning (a), with one transfer learning (b) and two transfer learnings (c).

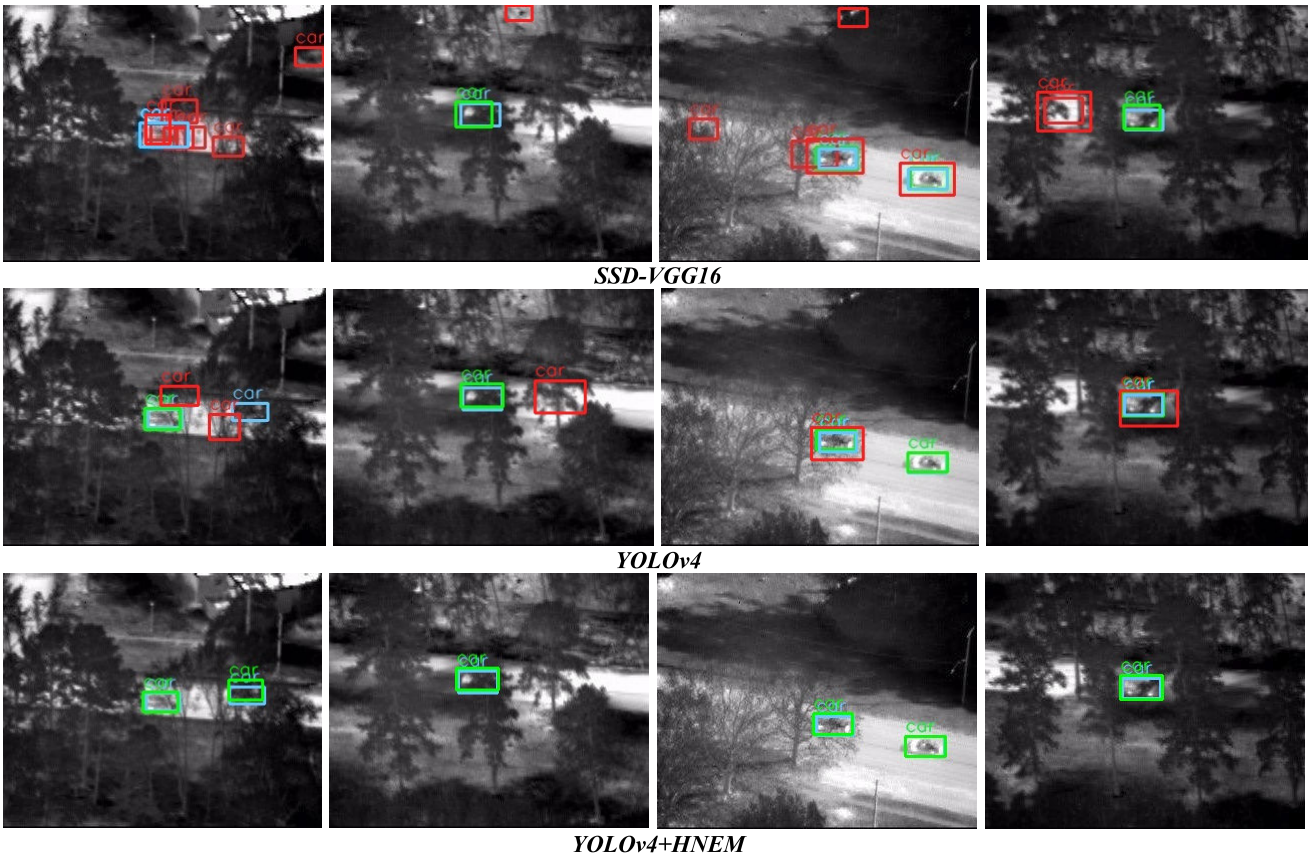



**FIGURE 8.** The detection result diagram of three representative models is selected. It can be seen from the diagram that the improved model has a good identification accuracy for complex background.

the ground truth label (GT), the green box represents the positive sample true detected by the model (TP), and the red box represents the positive sample falsely detected by the model (FP). The relationships among them are listed as the below quotations:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{TP}{GT} \tag{4}$$

$$\text{F1score2} = \frac{Precision * Recall}{Precision + Recall} \tag{5}$$

As can be seen from the above equations, the more the red boxes are, the lower the model's precision is, and the fewer the green boxes are, the lower the model's recall rate is. When the target is heavily occluded, it is difficult for the model to detect it, or some of the deceptive features from the backgrounds are still identified as the target with high confidence, as shown in the third column in Figure 7.

Figure 6 is maps of the training and validation losses of YOLOv4 models with different training strategies, including secondary transfer learning, one transfer learning and without transfer learning. We aim to verify the promoting effect of transfer learning in our model. From the three figures, we can find that without transfer learning, the parameters of the model could not converge to a very small value, and the converging speed is much slower due to the lack of sufficient dataset, and the final testing results also proved this conclusion. As for secondary transfer learning and one transfer learning, the training losses did not vary much, but the converging speed of secondary transfer learning is faster, and the validation loss seems to match with the training loss better than that of one transfer learning.

Figure 8 shows part of the detection results of the three models on some occluded vehicle images. Some misclassified objects from the non-improved models can be well identified by the improved model. It means that the HNEM module can depress the impact of the deceptive backgrounds information such as the trees and the change of light in the complex real environment.

## V. CONCLUSION

In this paper, the YOLOv4 model was introduced into the scenario of weak severely occluded vehicle detection in the infrared aerial images under complicated background, and on this basis, used the secondary transfer learning to overcome the problem of insufficient datasets, the hard negative example mining method at the same time to reduce the high false detection rate of the original model because of the complex background and occlusion influences. Through experimental verification, this method has certain feasibility and improvements. At the same time, it can be seen from the experiment

that there is still possibility for further improvement of the object detection in this scene, and the next step is to further improve the detection accuracy of the detection model in this scene by changing the structure of the model and the loss function.

## REFERENCES

[1] J. Hu, Y. Zhao, and X. Zhang, "Application of transfer learning in infrared pedestrian detection," in *Proc. IEEE 5th Int. Conf. Image, Vis. Comput. (ICIVC)*, Beijing, China, Jul. 2020, pp. 1–4, doi: 10.1109/ICIVC50857.2020.9177438.

[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, doi: 10.1109/ACCESS.2019.2944661.

[3] X. Zhang and X. Zhu, "Vehicle detection in the aerial infrared images via an improved YOLOV3 network," in *Proc. IEEE 4th Int. Conf. Signal Image Process. (ICSIP)*, Wuxi, China, 2019, pp. 372–376, doi: 10.1109/SIPROCESS.2019.8868430.

[4] X. X. Zhang and X. Zhu, "Moving vehicle detection in aerial infrared image sequences via fast image registration and improved YOLOv3 network," *Int. J. Remote Sens.*, vol. 41, no. 11, pp. 4312–4335, 2020, doi: 10.1080/01431161.2020.1717666.

[5] G. Zheng, X. Wu, Y. Hu, and X. Liu, "Object detection for low-resolution infrared image in land battlefield based on deep learning," in *Proc. Chin. Control Conf. (CCC)*, Guangzhou, China, Jul. 2019, pp. 8649–8652, doi: 10.23919/ChiCC.2019.8866344.

[6] Y. M. Kassim, M. E. Byrne, C. Burch, K. Mote, J. Hardin, D. R. Larsen, and K. Palaniappan, "Small object bird detection in infrared drone videos using mask R-CNN deep learning," *Electron. Imag.*, vol. 2020, no. 8, pp. 85-1–85-8(8), 2020, doi: 10.2352/ISSN.2470-1173.2020.8.IMAWM-085.

[7] Y. Wang, L. Wang, Y. Jiang, and T. Li, "Detection of self-build data set based on YOLOv4 network," in *Proc. IEEE 3rd Int. Conf. Inf. Syst. Comput. Aided Educ. (ICISCAE)*, Sep. 2020, pp. 640–642, doi: 10.1109/ICISCAE51034.2020.9236808.

[8] 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 *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Quebec City, QC, Canada, Sep. 2015, pp. 3735–3739, doi: 10.1109/ICIP.2015.7351502.

[9] R. Collins, X. H. Zhou, and S. Keat, "An open source tracking testbed and evaluation Web site," in *Proc. IEEE Int. Workshop Perform. Eval. Tracking Surveill.* Piscataway, NJ, USA: IEEE Press, Jun. 2005, pp. 35–42.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[11] D. Zhao, L. Gu, K. Qian, H. Zhou, T. Yang, and K. Cheng, "Target tracking from infrared imagery via an improved appearance model," *Infr. Phys. Technol.*, vol. 104, Jan. 2020, Art. no. 103116, doi: 10.1016/j.infrared.2019.103116.

[12] L. Kang, Y. Zhang, B. Zou, and C. Wang, "High-resolution PolSAR image interpretation based on human images cognition mechanism," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Milan, Italy, Jul. 2015, pp. 1849–1852, doi: 10.1109/IGARSS.2015.7326152.

[13] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*. [Online]. Available: http://arxiv.org/abs/2004.10934

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

[15] X. Wang, Y. Ban, H. Guo, and L. Hong, "Deep learning model for target detection in remote sensing images fusing multilevel features," in *Proc. IGARSS-IEEE Int. Geosci. Remote Sens. Symp.*, Yokohama, Japan, Jul. 2019, pp. 250–253, doi: 10.1109/IGARSS.2019.8898759.

[16] C. Li, T. Yang, S. Zhu, C. Chen, and S. Guan, "Density map guided object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Seattle, WA, USA, Jun. 2020, pp. 737–746, doi: 10.1109/CVPRW50498.2020.00103.

[17] C. Zhou and J. Yuan, "Occlusion pattern discovery for object detection and occlusion reasoning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 7, pp. 2067–2080, Jul. 2020, doi: 10.1109/TCSVT.2019.2909982.

[18] Z. Guo, W. Zhou, L. Xiao, X. Hu, Z. Zhang, and Z. Hong, "Occlusion face detection technology based on facial physiology," in *Proc. 14th Int. Conf. Comput. Intell. Secur. (CIS)*, Hangzhou, China, Nov. 2018, pp. 106–109, doi: 10.1109/CIS2018.2018.00031.

[19] W. Zhang, Y. Zheng, Q. Gao, and Z. Mi, "Part-aware region proposal for vehicle detection in high occlusion environment," *IEEE Access*, vol. 7, pp. 100383–100393, 2019, doi: 10.1109/ACCESS.2019.2929432.

[20] Y. Song, M. Li, X. Qiu, W. Du, and J. Feng, "Full-time infrared feature pedestrian detection based on CSP network," in *Proc. Int. Conf. Intell. Transp., Big Data Smart City (ICITBS)*, Vientiane, Laos, Jan. 2020, pp. 516–518, doi: 10.1109/ICITBS49701.2020.00112.

[21] Y. Wang and X. Bai, "Intensity inhomogeneity suppressed fuzzy C-means for infrared pedestrian segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 9, pp. 3361–3374, Sep. 2019, doi: 10.1109/TITS.2018.2875159.

[22] 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 *Proc. Int. Symp. Intell. Signal Process. Commun. Syst. (ISPACS)*, Taipei, Taiwan, Dec. 2019, pp. 1–2, doi: 10.1109/ISPACS48206.2019.8986298.

[23] P. Wang, W. Wang, and H. Wang, "Infrared unmanned aerial vehicle targets detection based on multi-scale filtering and feature fusion," in *Proc. 3rd IEEE Int. Conf. Comput. Commun. (ICCC)*, Chengdu, China, Dec. 2017, pp. 1746–1750, doi: 10.1109/CompComm.2017.8322839.

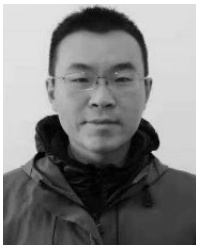
[24] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 5, pp. 1019–1034, May 2015, doi: 10.1109/TNNLS.2014.2330900.

[25] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[26] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6517–6525, doi: 10.1109/CVPR.2017.690.

[27] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

**SHUANGJIANG DU** was born in Xiangyang, Hubei, China, in 1996. He received the B.S. degree in aircraft engineering from the National University of Defense Technology, Changsha, Hunan, China, in 2018. He is currently pursuing the master's degree in optical information with the College of Communication Engineering, Army Engineering University of PLA. His research interests include machine learning, remote sensing, and object detection and camouflage.



**PIN ZHANG** received the M.S. degree in optical engineering and the Ph.D. degree in science and technology of weapon from the PLA University of Science and Technology, in 2013 and 2016, respectively. His research interests include electromagnetic camouflage, microwave photonics, and optical remote sensing. He received the Special Support of China Postdoctoral Fund, the General First-Class Support of China Postdoctoral Fund, the Jiangsu Provincial Natural Science Fund, and the Technical Fund of the Foundation Strengthening Plan of the Military Commission of Science and Technology.

**BAOFU ZHANG** received the B.S. and M.S. degrees from Xidian University, in 1987 and 1990, respectively, both in technical physics. He is currently a Professor with the College of Communication Engineering, Army Engineering University of PLA, and an Expert of Nanjing Workstation of China PLA General Political Department. He has published more than 30 articles in core journals, four monographs, one textbook for the 11th Five-Year Plan for higher education. His research interests include microwave photonics, photoelectric measurement, intelligent signal processing, and photoelectric reconnaissance and countermeasures. He is also a member of the Expert Group of Doctoral Supervisors, and a Senior Member of the China Electronic Association and the China Institute of Communications. He received three First prizes for Teaching Achievements from the PLA University of Science and Technology and two Third prizes for Military Science and Technology Progress. He was a reviewer of *Acta Electronica Sinica* and *Journal on Communications and Acta Optica Sinica*.

**HONGHUI XU** was born in Shantou, Guangdong, China, in 1995. He received the B.S. degree in aircraft engineering from the National University of Defense Technology, Changsha, Hunan, China, in 2018. He is currently pursuing the master's degree in mechanical engineering with the College of Field Engineering, Army Engineering University of PLA. His research interests include machine learning and computer vision, especially object detection.