Design Studies and Intelligence Engineering
L.C. Jain et al. (Eds.)

doi:10.3233/FAIA220705

Few-Shot Infrared Ship Detections via Improved TFA with Similarity Contrast and VOVNetv2

Miao Lize[a], Li Ning[a,1], Zhou minglong[a], and Zhou Huiyu[b]
[a] *Nanjing University of Aeronautics and Astronautics, Nanjing, P R China*
[b] *University of Leicester, Leicester, UK, LE1 7RH*

Abstract. The low resolution of infrared images makes it more difficult to detect objects, and the quality of detection results obtained by the CNN based object detection model are worse for few-shot problems. Two-stage Fine-tune Approach (TFA) is effective to improve the precision of detection for few-shot problems. Because of the category imbalance of training samples, TFA has the problem of misclassification. To solve this problem, TFA with similarity contrast (SC-TFA) is proposed. The VOVNetv2 is used as the backbone network to improve the detection accuracy. The similarity contrast detection head is added to the detection module to improving the classify performance. And both cosine similarity and Euclidean distance are used as the similarity measure in the contrast loss function. The effectiveness of the improved TFA for the few-shot problem is verified on the VOC dataset and the infrared ship dataset. The average precision of the novel categories (nAP) of SC-TFA on VOC dataset and the infrared ship dataset reaches 54.92% and 41.1% respectively, which is 4.7% and 3.4% higher than TFA.

Keywords. few-shot learning; object detection; similarity contrast; fine-tune.

1. Introduction

Compared with RGB images, infrared images are less affected by the harsh environment conditions. Infrared image can be used normally under conditions such as strong light irradiation, complete darkness, and haze weather. Infrared images are widely used in military, transportation, security and other fields. In some special scenes, such as military scenes, it is difficult to obtain infrared images for the target, and the commonly used deep learning models cannot obtain qualified detection results because of the few-shot problem.

Existing object detection model have limitations in infrared object detection for few-shot problem. Duan et al. [19] propose to add an auxiliary network to Yolo, which can improve the quality of detection result for infrared object; Shu et al. [20] propose to improve Yolov5 by using DenseNet. These methods can improve the accuracy of infrared object detection, but do not perform well on the problem of few-shot.

The RepMet [1] to learn the multi-modal distribution of different categories of training samples while optimize the model parameters, so that the model can obtain the classification and detection ability in the case of few-shot learning. Xin et al. [4] fine-

[1] Corresponding Author. E-mail: lnee@nuaa.edu.cn

tuning the full connected layer and the detection head to accomplish few-shot object detection. Transfer learning can reduce the data requirements of the model, but it is easy to misclassify due to the unbalanced samples of base classes and novel classes.

To solve the problems mentioned above, TFA is used as the baseline, and improved by adjusting the network structure and loss function. The contributions of our work can be summarized as follows:

In the feature extraction part, dense connection is used to improve the feature extraction ability of the model. ResNet is replaced by VOVNetv2 [7] in the backbone, which accomplish the reuse of the feature map in the channel dimension by densely transferring the feature map to the deep layer. During the fine-tuning stage, a similarity contrast detection head is added to improve the detection accuracy. The loss function uses both cosine similarity and the Euclidean distance of the feature vector, which make feature with the same category label aggregation, and separate the feature with different categories label. It can improve the classification precision of the detection model.

2. Related work

Transfer learning, which is used in few-shot problem, learns knowledge from the source domain with enough training data, and then transfers it to the target domain which has limited training data. TFA pretrained the model on the base dataset, and then in the few-shot fine-tuning stage, only the parameters of the box classification and regression in the model is updated. The dataset used in the few-shot finetuning stage contains training data of the novel classes and the base classes. The training strategy make the model has the detection ability for both the base classes and novel classes, and the model achieves better

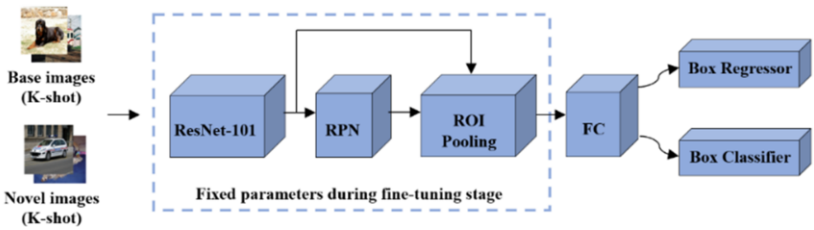

Figure 1. The structure of the model used by TFA

detection results. The structure of TFA as shown in Fig.1.

FSCE [8] based on TFA proposed a contrastive loss to improve the classification performance of the model. FSCE measures the similarity between the feature with different categories labels, and a contrastive loss function is added to increase the consistency between feature with the same category labels and improve the difference of feature with the different category labels, and finally improves the accuracy of few-shot object detection. The cosine similarity is used in the contrastive loss function to measure the feature similarity. The Euclidean distance of the vectors in the feature space is ignored.

3. The proposed method

TFA is used as the baseline, adopt the two-stage finetune approach and obtain the model that can complete the detection task in the case of few-shot. In the feature extraction

stage, VOVNetv2 [7] is used as the feature extraction network, and the information in the feature map is fully utilized by the dense connection, so as to obtain better detection results. In the stage of fine-tuning, the parameters in the backbone network are fixed, and the similarity contrast detection head is added to the model with reference to FSCE. By improving the contrast loss function, the accuracy of the model for object classification is increased. The improved model structure as shown in Fig.2.

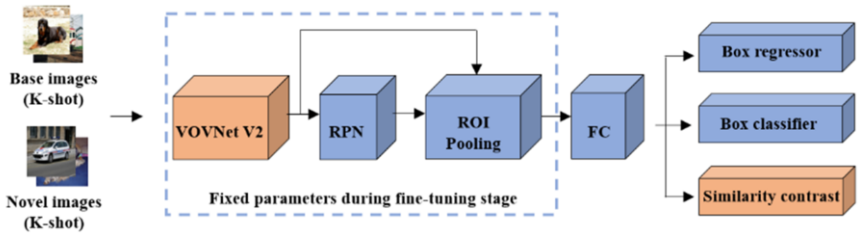

Figure 2. The structure of SC-TFA

3.1. Improve backbone via VOVNetv2

The backbone network used by TFA is ResNet-101, and the residual network alleviates the problems of gradient disappearance and gradient explosion through skip connections, so that the model can be deeper and have stronger feature extraction capabilities. Through the dense connection, the feature extraction ability of the backbone network can be further enhanced, so VOVNetv2 is used to replace ResNet.

VOVNetv2 mainly consists by improved OSA modules. The input image first passes through the stem block composed of 3×3 convolutional layers, and then the OSA module of 4 stages. At the end of each stage, a 3×3 max pool layer with stride of 2 is used for downsampling. The OSA module mainly uses 3×3 convolutional layers. In order to reduce the redundant information of a large number of feature maps generated by aggregation, concat is only performed in the last layer. Then, the weights are allocated to the obtained feature maps through the effective squeeze and excitation modules (eSE) [7], so as to obtain more reliable feature maps. The structure of OSA as shown in Fig.3.

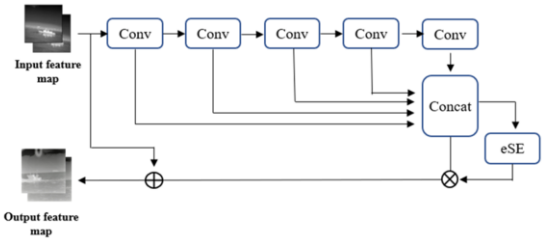

Figure 3. The structure of OSA

3.2. Cosine and Euclidean Distance (CED) loss

In the object detection task, the localization and classification of the object are need to complete. In order to obtain better object classification results, it is necessary to make the eigenvectors of different categories to be separated in the feature space, and the

eigenvectors of the same category to be aggregated. To obtain the feature that are suitable for the task of object detection, in addition to an effective backbone network, an appropriate loss function is also required. Therefore, the similarity contrast detection head is added to the model, and define the cosine and Euclidean distance (CED) loss. The definition of the SC-TFA loss function is as follows:

$$L_{total} = L_{rpn} + L_{cls} + L_{reg} + \alpha L_{CED} \tag{1}$$

L_{total} is the total loss function of the model; L_{rpn} use binary cross-entropy loss to product regional proposal; L_{cls} is a cross-entropy loss which is used to classify the bounding box; L_{reg} use smooth L_1 loss to get the more precise bounding box positioning; L_{CED} used in the similarity measurement branch, and α is a hyper-parameter.

The content of similarity measure includes angle and Euclidean distance，and the L_{CED} is defined as follows:

$$L_{CED} = L_{COS} + L_{ED} \tag{2}$$

where L_{COS} is used to calculate the loss of angle similarity, and L_{ED} is used to calculate the loss of Euclidean distance. For the N features of the similarity contrast detection head input, the calculation process of L_{COS} is as follows:

$$L_{COS} = \frac{1}{N}\sum_{i=1}^{N} f(u_i \geq \phi) \cdot L_{x_i} \tag{3}$$

$$L_{x_i} = \frac{-1}{N_{y_i}-1}\sum_{j=1,j\neq i}^{N} f(y_i = y_j) \cdot log\frac{exp(\tilde{x}_i \cdot \tilde{x}_j/\tau)}{\sum_{k=1,k\neq i}^{N} exp(\tilde{x}_i \cdot \tilde{x}_k/\tau)} \tag{4}$$

$$\tilde{x}_i \cdot \tilde{x}_j = \frac{x_i \cdot x_j}{\|x_i\| \cdot \|x_j\|} \tag{5}$$

where $f(\cdot)$ is a conditional judgment function, output 1 when the given condition is true, otherwise output 0; x_i is the feature of *i-th* region proposal; u_i is the value of Intersection over Union (IOU); y_i is the ground truth; N_{yi} is the number of region proposal which label is y_i; τ is a hyper-parameter.

The loss function is constructed by the cosine similarity of the features. Cosine similarity has the disadvantage of ignoring distance. As shown in Fig.4, feature A is more similar to feature B than feature C when only cosine similarity is considered.

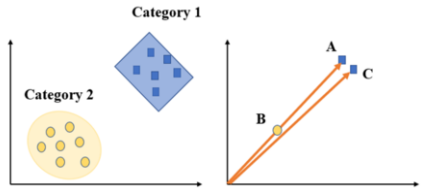

Figure 4. Two-dimensional schematic diagram of the distribution of different categories of eigenvectors in the feature space

Therefore, it is necessary to add the loss function of the distance metric, and the calculation process of L_{ED} is as follows:

$$L_{ED} = \frac{1}{N}\sum_{i=1}^{N} f(u_i \geq \phi) \cdot L_{dx_i} \tag{6}$$

$$L_{dx_i} = \frac{-1}{N_{y_i}-1}\sum_{j=1,j\neq i}^{N} f(y_i = y_j) \cdot log\frac{d(x_i,x_j)}{\sum_{k=1,k\neq i}^{N} d(x_i,x_k)} \tag{7}$$

$$d(x_i,x_j) = 1 - \frac{1}{1+\sqrt{x_i \cdot x_i + 2 \cdot x_i \cdot x_j + x_j \cdot x_j}} \tag{8}$$

L_{ED} uses Euclidean distance as the similarity measure of feature, which aggregate features of the same category and increase the distance between features of different categories.

4. Experiment

In this section, the split of the dataset, the setting of model hyperparameters, and visualize the detection results will be detailed.

4.1. Dataset

During the experiments, the PASCAL VOC dataset and infrared ship dataset are used to train and test the model. The VOC dataset with a total of 20 classes of objects is divided into 15 base classes and 5 novel classes. For the infrared ship dataset, there are a total of 7 classes of ships, and be randomly divided to 5 base classes and 2 novel classes. The infrared ship dataset has a total of 8400 infrared images, of which 1200 are used as the test set, and the remaining images are used as the training set.

For the few-shot problem of K-shot (K=5, 10), using random sampling to select K instances for each category. In order to prevent data leakage, the data used in the training process is sampled from the train set of datasets, and the test set of datasets is used for testing.

In the divided dataset, the number of training samples corresponding to base categories is not limited. It is used in the base-training to provide pre-trained model parameters for subsequent training. Both base categories and novel categories in the fine-tuning process can only use K instances to train the model. In the process of testing the model, the AP50 of the base categories(bAP) and the AP50 of the novel categories(nAP) are calculated to evaluate the quality of the model. Since the model is mainly aimed at the few-shot object detection, nAP is used as the main evaluation of the model.

4.2. Implementation Details

During training, the batch size is set as 16. To alleviate overfitting, multiple data augmentation strategies are employed, including mosaic, random scale and adjust chroma. Mosaic [13] cuts and splices 4 images to enrich the background of the image; random scale randomly enlarges and reduces the size of the original image, enriching the object data of different scales; adjust chroma transforms the RGB color components of the image, and enhances the model to color robustness of transformations.

The SGD with momentum 0.9 and weight decay 0.0001 is used as optimizer. The learning rate is set as 0.01 during the base training and set as 0.001 during the fine-tuning. In order to make the model converge better the learning rate decay strategy is adopted. For alleviating the large oscillation of the model in the early stage of training, warm-up strategy is used, using a small learning rate for the first 200 steps, and then gradually increasing to the base learning rate.

4.3. Experiment Results

In the detection process, TFA is used as the baseline. The backbone network and loss function of detection head in the model are adjusted. MobileNetV2, ResNet, ResNeXt and VOVNetv2 is compared by the bAP of the base training result. The results of the comparison are shown in Table 1.

It can be seen from the detection results that using VOVNetv2 as the backbone network can improve the detection results of the model. Using MobileNetV2 can obtain a higher detection speed, but the detection accuracy is obviously inferior to other backbone networks. Since the base training stage needs to provide the parameters of the backbone network for the finetune stage, the accuracy loss will be amplified in the few-shot training. Therefore, VOVNetv2 is used as the backbone,

Table 1. Comparison of detection results of different backbone networks in the base training stage

Backbone	**bAP**	**FPS**
MobileNetV2	67.88%	20.60
ResNet-101	79.80%	13.16
ResNeXt-101	78.99%	7.09
VOVNetv2	80.31%	10.11

In the finetune stage, the parameters of the backbone are fixed, and fine-tun the parameters of the detection head by the samples of novel categories. Through the similarity contrastive detection head, the angle and Euclidean distance difference of the feature vector in the projection space are used to optimize the model parameters to obtain the better detection results.

TFA is used as the baseline to conduct ablation experiments, and compared the effects of VOVNetv2 and CED loss on the few-shot object detection. nAP is used as the evaluation standard, and the detection results of the model at 5-shot and 10-shot are obtained, and the detection results are shown in Table 2.

Table 2. Ablation experiment results of improved models on VOC dataset

Models	Backbone	Contrastive Loss	nAP	
			5-shot	10-shot
Model-1	ResNet-101	L_{COS}	45.26%	53.17%
Model-2	VOVNetv2	L_{COS}	47.34%	53.74%
Model-3	ResNet-101	L_{CED}	47.24%	54.58%
Model-4	VOVNetv2	L_{CED}	49.07%	54.92%

Model-1 in the table use ResNet-101 as the backbone and use only cosine similarity in the similarity contrast loss function. The 5-shot and 10-shot nAP of this model on the test set dataset respectively reach 45.26% and 53.17%. Model-2 uses VOVNetv2 as the backbone, which increases the 5-shot and 10-shot nAP by 2.08% and 0.57% compared to the baseline. Model-3 uses L_{CED} on the similarity contrastive detection head, the 5-shot and 10-shot nAP of the model are increased respectively by 1.98%. and 1.41%. Model-4 uses both VOVNetv2 and L_{CED} to obtain the best detection results, the 5-shot and 10-shot nAP have increased to 49.07% and 54.92%.

Table 3. Compare SC-TFA with common few-shot models on VOC dataset

Models	Backbone	nAP	
		5-shot	10-shot
RepMet	InceptionV3	34.4%	37.2%
Meta R-CNN	ResNet-101	41.2%	48.1%
TFA	ResNet-101	48.7%	50.2%
FSCE	ResNet-101	45.2%	53.2%
SC-TFA (ours)	VOVNetv2	49.1%	54.9%

SC-TFA can obtain qualified detection results with a small number of samples. Compared with common models for few-shot problems, our model is able to achieve higher nAP in both 5-shot and 10-shot cases. The test results are shown in Table3.

The detection results are visualized as shown in Figure 5. When detecting some objects with special angles, it is missed by TFA, FSCE, but SC-TFA can detect them, as shown in Figure 5 (b), (e), (h). In the case of complex occlusion, it is difficult for the model to accurately locate and classify the object. TFA did not detect heavily occluded aircraft, and the detection results of FSCE and SC-TFA were also unqualified， as shown in the Figure 5 (c) , (f) and (i).

SC-TFA also got qualified detection results on the infrared ship dataset, the nAP of 2-way 5-shot and 2-way 10-shot is 34.6% and 41.1%. The experimental results have shown that the improvement of the model is also effective for the infrared ship dataset. The detection results on infrared ship dataset are visualized as shown in Figure 6. When the object is of regular size, because of the improved backbone, SC-TFA can obtain the qualified detection results under the condition of both regular and low resolution, as shown in Figure 6 (a) and (b). For small objects and truncated objects, the improved model can also obtain good detection results, as shown in Figure 6 (c) and (d). For slightly occluded objects, the improved model still performs well, as shown in Figure 6 (e). For ships in the port scene, due to the density and serious occlusion of objects, and some objects are unavoidable to be lost, SC-TFA can detect part of all objects, as shown in Figure 6 (f).

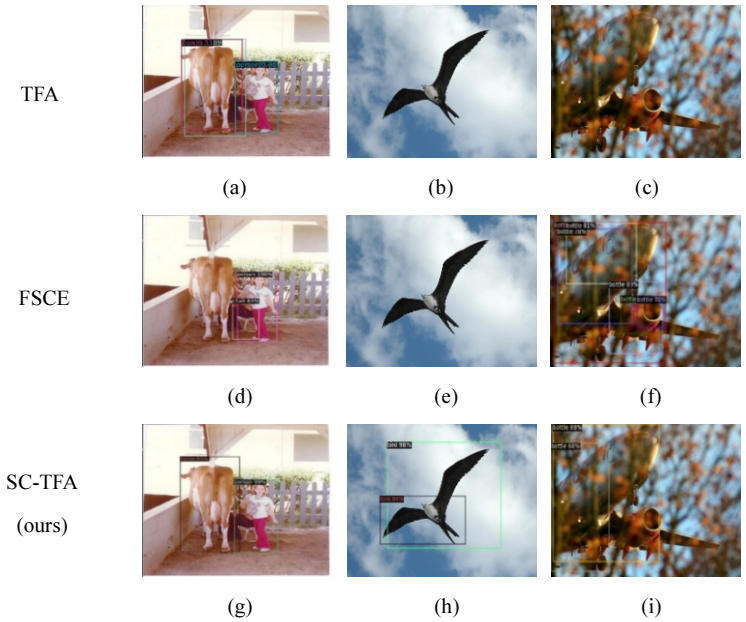

Figure 5. Comparison of test results of TFA, FSCE and SC-TFA on VOC dataset for 10-shot object detection.

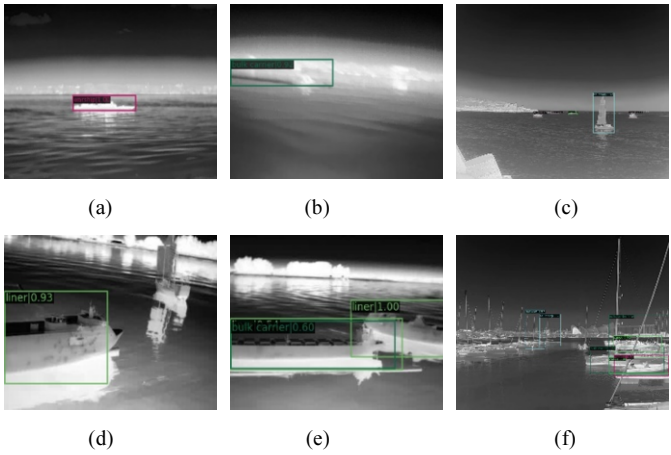

Figure 6. The detection result of SC-TFA on infrared ship dataset for 2-way 10-shot.

5. Conclusion

In this paper, the model named SC-TFA is proposed for few-shot infrared ship detection. SC-TFA obtain a reliable few-shot object detection result by fine-tuning training with improved backbone and similarity contrast loss function. In the process of base training, using VOVNetv2 as backbone, which can extract and aggregate image features of different receptive fields. In the process of fine-tuning stage, using cosine similarity and Euclidean distance of features in the similarity contrast loss function, so as to obtain more effective features for infrared object classification. The improved model has

increased detection accuracy compared with TFA on the VOC dataset. SC-TFA can obtains the qualified results for few-shot infrared object, and favorable detection results can still be obtained when detecting complex and occluded objects.

Funding

This work received support from Science and Technology on Electro-optic Control Laboratory and Aviation Science Foundation Project (ASFC-20175152036), National joint project of foreign experts (1004/011951G20130) and Key Project on Artificial intelligence (1004-56XZA19008). The authors are also grateful for the support of their colleagues at the Key Laboratory of Radar Imaging and Microwave Photonics, Ministry of Education.

References

[1] L. Karlinsky, J. Shtok, S. Harary, E. Schwartz, A. Aides, R. Feris, R. Giryes, and A. M. Bronstein. Repmet: Representative-based metric learning for classification and few-shot object detection. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 5192–5201, 2019.

[2] Fu K, Zhang T F, Zhang Y, et al. Meta-SSD: towards fast adaptation for few-shot object detection with Meta-learning [J] . IEEE Access, 2019, 7: 77597-77606.

[3] Y Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. 2019. Meta R-CNN: Towards General Solver for Instance- Level Low-Shot Learning. In ICCV2019. IEEE, 9576–9585.

[4] Xin Wang, Thomas E. Huang, Joseph Gonzalez, Trevor Darrell, and Fisher Yu. 2020. Frustratingly Simple Few-shot Object Detection. In ICML 2020 (Proceedings of Machine Learning Research, Vol. 119). 9919–9928.

[5] T. Wang, X. Zhang, L. Yuan and J. Feng, Few-shot Adaptive Faster R-CNN, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 7166-7175.

[6] REN S, HE K, GIRSHICK R, et al. Faster R-CNN：towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39 (6): 1137-1149.

[7] Y. Lee and J. Park, CenterMask: Real-Time Anchor-Free Instance Segmentation, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 13903-13912.

[8] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. 2021. FSCE: Few-shot Object Detection via Contrastive Proposal Encoding. In CVPR 2021. Computer Vision Foundation / IEEE, 7352–7362.

[9] LEE Y, HWANG J, LEE S, et al. An energy and gpu-computation efficient backbone network for real-time object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.2019:0-0.

[10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In CVPR, 2017.

[11] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In ECCV, 2018.

[12] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In CVPR, 2018.

[13] Chien-Yao Wang, Alexey Bochkovskiy, and HongYuan Mark Liao. Scaled-yolov4: Scaling cross stage partial network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13029–13038, 2021.

[14] Oriol Vinyals Aaron van den Oord, Yazhe Li. Representation learning with contrastive predictive coding. Advances in Neural Information Processing Systems, 31, 2018.

[15] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In International Conference on Learning Representations, 2019.

[16] T. Lin, P. Doll´ar, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 936–944, 2017

[17] Lu Zhang, Shuigeng Zhou, Jihong Guan, and Ji Zhang. 2021. Accurate Few-shot Object Detection With Support-Query Mutual Guidance and Hybrid Loss. In CVPR 2021. Computer Vision Foundation / IEEE, 14424–14432
[18] Wang X L, Shrivastava A, Gupta A, et al. A-Fast-RCNN: hard positive generation via adversary for object detection [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 3039-3048.
[19] Duan Huijun, Wang Zhigang, Wang Yan. Two-channel saliency object recognition algorithm based on improved YOLO network[J]. Laser & Infrared, 2020,50(11):1370-1378.
[20] Shu Lang, Zhang Zhijie, Lei Bo. Research on Dense-Yolov5 Algorithm for Infrared Target Detection[J]. Optics & Optoelectronic Technology,2021,19(01):69-75.