IYOLO: Multi-scale infrared target detection method based on bidirectional feature fusion

Xiaofeng Zhao¹, Yebin Xu⁴*, Fei Wu¹, Wei Cai¹ and Zhili Zhang¹
¹Xi'an Research Institute of High Technology, Xi'an, Shanxi, 710025, China
*Corresponding author's e-mail: ghyh15@hxbcd.com

Abstract: Convolutional neural networks are advanced computer vision solution for various tasks, and have made considerable progress. In order to detect infrared targets with details and texture features, we propose a detection method based on YOLOv3. First, we propose a bidirectional feature fusion structure called Improve-FPN(Im-FPN), which allows fast and efficient multi-scale feature fusion. Second, we use the Focal Loss(FL) loss function to balance the problem of sample imbalance. Based on these optimizations, an Infrared-YOLO (IYOLO) target detection model for infrared target is constructed. On the self-constructed infrared dataset: Infrared-VOC, the mAP of IYOLO can achieve 77.1%, which is 4.0% higher than YOLOv3, and the detection speed can reach 55.6FPS. Experimental results show that this method can improve the overall accuracy of detection while ensuring real-time detection.

1. Introduction

Deep learning is driving advances in recognition. Computer vision based on deep learning has achieved unprecedented development. With the support of powerful visible light datasets provided by various computer vision challenges such as ILSVRC, which are not only improved for whole-image classification[1-3], but also achieved great promotion on object detection[4-7], image segmentation[8,9] and other directions. The research of object detection based on deep learning is gradually extended to pictures of other imaging modalities. Infrared images used the thermal radiation of the object as the source of information, and the resulted images mainly highlights the thermal distribution of the object. It’s less affected by other factors such as light and weather, and has the ability to detect all day. Target detection algorithms based on deep learning have become an important research direction in the field of infrared target detection, and have broad application prospects in aerial reconnaissance, autopilot driving, and video surveillance based on UAV platforms.

Target detection algorithms which based on convolutional neural networks (CNN) are mainly divided into two-stage target detection algorithms such as: R-CNN[10], Fast R-CNN[11], Faster R-CNN[4] and one-stage target detection methods based on regression such as: SSD[5], YOLO[6,7]. The one-stage target detection algorithm represented by YOLOv3, associated the prediction frame with the actual frame through traversal, and used the NMS mechanism to suppress the prediction boxes with low correlation to achieve the purpose of detection. YOLOv3 reduced the process of extracting proposal regions, which help it achieved the requirements of real-time detection in many scenarios, and has been widely used. Inspired by the fast and elegant working principle of the one-stage target detection method, this paper applies the target detection algorithm based on CNN to the infrared target detection with detailed and texture features, and proposes a fast and efficient target detection model.
2. Related Works

Recently, target detection based on infrared images is mainly divided into infrared small target detection based on space or sea and infrared target detection with detailed and texture features on the ground. In the research based on point-shaped infrared small targets, Ma et al. [12] used a residual neural network to detect and recognize low-altitude UAV targets. Liu et al. [13] utilized a full convolutional neural network and visual saliency method to detect infrared targets in complex background. Wu et al. [14] used the convolutional neural network and the SENet model to detect the infrared weak and small targets. And Dai et al. [15] utilized a convolutional neural network and a local attention mechanism to detect small infrared targets. Those are mainly used in the detection of infrared small targets in space-based and sea-based backgrounds. At present, there are few researches on infrared targets with detailed and texture features through deep learning. The main challenges are following aspects:

1. Infrared targets with detailed and texture features require high quality for the imaging system, and also need lots of manpower and material resources. Consequently, there are fewer public datasets.

2. The current research is mainly based on visible light images. Although infrared imaging has advantage of less affected by natural factors, the infrared imaging is continuous and easily affected by complex backgrounds, which easily cause missed and false detection.

In order to solve these problems, we used a portable infrared imager to collect data of infrared vehicle targets in the road scene, and generate an Infrared-VOC dataset for training and testing the constructed algorithm. Then, starting from the imbalance of the dataset and multi-scale feature fusion, the YOLOv3 target detection algorithm is improved. This part will be discussed in Section 3.2 and 3.3. Experiments have proved that our optimization algorithm can effectively detect infrared targets with detailed and texture features, while ensuring the real-time detection.

3. IYOLO Target Detection Model

The constructed IYOLO target detection model is shown in Figure 1.

![Network structure of IYOLO](image)

The input image is downsampled and encoded in the feature extraction network to reduce the image resolution. Then multi-scale feature fusion is performed on the extracted features at different levels to achieve feature fusion and reuse. Finally, the fused feature map is used as the input of the target detection network to perform decoding, target classification, coordinate prediction, bounding box regression and complete the detection of the target in the image.

3.1 Backbone Network

The infrared image accepts the thermal radiation information of the target for imaging. It has features such as low imaging resolution, continuous imaging area and contour blur. In a complex background, the feature extraction of the target is more susceptible to interference. In order to perform sufficient feature extraction, we used DarkNet-53 as the backbone of the feature extraction network, and extracted the input image features of 320×320 pixels. Then fed them to the feature fusion network.

3.2 Bi-directional Feature Fusion Network

The FPN[16] structure perform top-down feature fusion on the multi-scale features, which obtained by
the feature extraction network, so that multi-dimensional information can complement each other. It can make full use of information of multi-scale and make up for the phenomenon of insufficient utilization of information at each layer effectively. However, the size is changed during the upsample process and feature information is easily lost. In this article, the Im-FPN structure is used to replace the FPN. Several feature fusion structures are shown in Figure 2.

![Figure 2. Comparison chart of feature fusion network. (a) FPN. (b) Im-FPN.](image)

Inspired by residual connection and PANet[17], the bi-directional feature pyramid network proposed was used V-shaped fusion to complement the features in this paper. Compared with PANet, the calculation of the start and end of the middle layer is omitted, while the number of model parameters is reduced. Compared with FPN, it can make full use of the different levels of features obtained by the feature extraction network when a small parameter is added.

### 3.3 Focal Loss

The Focal Loss[18] (FL) can suppress the weight of simple samples, promote the model to have a better balance ability, and help the detection of small sample categories. FL is defined as:

\[
FL(q, p) = -\alpha_t(q(k)) \sum_{k=1}^{K} \log(p(k))q(k)
\]  

(1)

Here:

\[
\alpha_t = \begin{cases} 
\alpha & \text{if } p = 1 \\
1 - \alpha & \text{otherwise}
\end{cases}
\]  

(2)

\(\lambda\) is the focus parameter, the weight for simple samples can be smoothly adjusted. \(q(k)\) express the prediction confidence of the corresponding category, \(\varepsilon\) is the attenuation factor, \(K\) is the number of categories. This paper obtains through experiments: \(\lambda = 2.0\), \(\alpha_t = 0.25\), \(K = 2\), \(\varepsilon = 0.1\).

The total loss is the sum of the prediction box regression, classification and local loss, and the total loss is defined as:

\[
L_{all} = L_{GIOU} + L_{S-FL} + L_{loc}
\]  

(3)

### 4. Experiments

In this section, we present the target detection results achieved by infrared-VOC dataset. The dataset was taken by the portable infrared thermal imaging camera: Tix660, and consist of cars and buses in the road scene.

The IYOLO was fine-tuned by using SGD with initial learning rate \(10^{-3}\), 0.9 momentum, batch size 48. The experimental platform used in this paper is Linux 18.04, the CPU is Inter Core i9-9900K CPU @3.60Ghz, the GPU is Nvidia P6000, 24GB. CUDA 11.0 and cudnn 8.0.4.30 were used for accelerated training under the pytorch 1.7.1 framework.

#### 4.1 IYOLO for Object Detection

In this experimental study, SSD, Faster RCNN, YOLOv3, and IYOLO are compared. The following Table 1 demonstrates mean average precision (mAP), average precision(AP) and frames per second(FPS) on the Infrared-VOC dataset.

The IYOLO achieved the highest mAP score at 0.5 IOU level, which reached 77.1%. In the one-
stage algorithm, it is 4% higher than YOLOv3, and 7.4% higher than SSD, while the detection speed reaches 55.6FPS, which meets the needs of real-time monitoring. Meanwhile, it is 1.5% higher than Faster R-CNN, and the detection speed is 7.94 times. Experiments have proved that IYOLO can significantly improve the average accuracy of target detection while maintaining a high real-time monitoring speed.

Table 1. Test results of different target detection algorithms based on Infrared-VOC dataset

| Model      | Backbone  | FPS / (frame-s⁻¹) | mAP@0.5 / % |
|------------|-----------|--------------------|-------------|
| Faster R-CNN | ResNet-101 | 7                  | 75.6        |
| SSD        | VGG16     | 46                 | 69.7        |
| YOLOv3     | DarkNet-53| 58.8               | 73.1        |
| IYOLO      | DarkNet-53| 55.6               | 77.1        |

Figure 3 shows the detection results of the target on the Infrared-VOC dataset, the input image size is 320×320 pixel, the color of bounding box is matched randomly.

4.2 Ablation Study
In this experiment, we investigate the ablative effect of the IYOLO, an ablation experiment was performed on the Infrared-VOC dataset. We adopted the YOLOv3 as a baseline, and ablation learning was performed by changing different modules. The results are shown in Table 2.

Through experimental comparison, we found that YOLOv3 reuses the feature extraction network with FPN, which mAP is 73.1%, but the gap between AP values. Meanwhile, We found that after using Focal loss, the inter-class AP values are reduced, which has an obvious inter-class balance effect. More surprisingly, the mAP of IYOLO reached 77.1% by combining Im-FPN and FL, which is 4.0% higher than YOLOv3. As shown in the Table 2, the Im-FPN designed in this paper can reduce the parameters and improve the detection accuracy.

Table 2. Ablation experiment based on Infrared-VOC dataset

| YOLO v3 baseline | Focal loss | Im-FPN | AP@0.5 / % | mAP@0.5 / % |
|------------------|-----------|--------|------------|-------------|
|                  | √         |        | 75.2       | 71.0        | 73.1        |
|                  | √         | √      | 75.0       | 74.6        | 74.8        |
|                  | √         | √      | 76.7       | 77.6        | 77.1        |
5. Conclusions
The IYOLO proposed in this paper effectively solves the problem of low accuracy for infrared target detection by improving the feature fusion network and combining the Focal loss to balance each category in the dataset. On the Infrared-VOC dataset, mAP of IYOLO achieved an excellent detection performance of 77.1%. It fully explains the important role of the feature fusion network in the target detection model, and provides a reference for subsequent research on feature fusion.

Acknowledgments
This work was supported by the National Natural Science Foundation of China (No.41404022) and Natural Science Foundation of Shanxi Province (No.2015JM4128).

References
[1] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 1097-1105.
[2] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
[3] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
[4] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*.
[5] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). Ssd: Single shot multibox detector. In *European conference on computer vision* (pp. 21-37). Springer, Cham.
[6] Redmon, J., Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.
[7] Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.
[8] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431-3440).
[9] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Springer, Cham.
[10] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580-587).
[11] Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440-1448).
[12] Ma, Q., Sun X.J., Zhang, Y., & Jiang, Y.C. (2020). Low-altitude UAV detection and recognition method based on infrared images. Journal of Projectiles, Rockets ,Missies and Guidance (03), 150-154.
[13] Liu, J.M., & Meng, W.H. (2020). Infrared small target detection combining full convolutional neural network and visual saliency. Acta Photonica Sinica (07), 46-56.
[14] Wu, S.C., & Zuo, Z.R. (2019). Infrared small target detection based on deep convolutional neural network. J. Infrared Millimeter Waves (03), 371-380.
[15] Dai, Y., Wu, Y., Zhou, F., & Barnard, K. (2021). Attentional local contrast networks for infrared small target detection. IEEE Transactions on Geoscience and Remote Sensing, PP(99), 1-12.
[16] Lin, T. Y., Dollár, Piotr, Girshick, R., He, K., Hariharan, B., & Belongie, S. (2016). Feature pyramid networks for object detection.
[17] Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path aggregation network for instance segmentation.

[18] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, Piotr. (2017). Focal loss for dense object detection. IEEE Transactions on Pattern Analysis & Machine Intelligence, PP(99), 2999-3007.