Airborne infrared aircraft target detection algorithm based on YOLOv4-tiny

Xiangguan Hou a, Jianwei Ma b, Shaofei Zang c
Henan University of Science and Technology, Luoyang 471000, China

a houxiangguan@haust.edu.cn, b lymjw@163.com, c 510883896@qq.com

Abstract. Aiming at the problem that the current aerial infrared aircraft targets are blurred, low contrast, and susceptible to noise interference, which can lead to inaccurate recognition, an improved YOLOv4-tiny infrared aircraft target detection method based on cavity convolution is proposed. First, in order to make full use of the shallow features, add a parallel branch after the feature layer whose output size of YOLOv4-tiny is; then, after the first network output layer, add three parallel holes of 1, 3 respectively. 5. The depth of 5 can separate the hollow convolution layer to expand the feature map receptive field; finally, the feature fusion network is improved, the final output feature layer of the network, the improved anti-residual block extraction feature, and the size is adjusted after convolution. The prediction result is processed and output by yolo head. Experiments on the aerial infrared aircraft data set show that compared with the original YOLOv4-tiny, the detection accuracy is increased by 4.29% and the detection effect is significantly improved under the premise of less loss of detection frame rate and small weight.

Keywords: YOLOv4-tiny; hole convolution; anti-residual; infrared aircraft target.

1. Introduction
Preface Infrared [1] target detection in the air is one of the key technologies in the field of precision guidance. It is widely used in early warning systems for military installations such as aircraft infrared search and tracking systems, infrared imaging and guidance systems. However, due to the small size, fuzzy outline and low contrast of aerial infrared aircraft targets, they are susceptible to noise interference, making aerial infrared targets present many challenges in the actual detection process. Therefore, the research of aerial infrared target detection algorithm still has strong challenges. Target detection algorithms are usually divided into two categories, namely traditional target detection algorithms and target detection algorithms based on deep learning. Because traditional target detection algorithms represented by hand-designed features have low detection accuracy and poor robustness, this paper chooses deep learning target detection algorithms to detect aerial infrared targets.

The current target detection algorithms based on deep learning can be divided into two types: two-stage detection network and single-stage detection network. The two-stage detection network has the characteristics of high detection accuracy and accurate positioning, such as R-CNN [2], Fast R-CNN [3], Faster R-CNN [4,5], DCN [6], etc. Among them, R-CNN extracts about 2000 candidate frames for each image and uses the convolutional neural network to extract features, but accuracy is still low and
the candidate frame selection is time-consuming; for the low-precision problem, Fast R-CNN uses the final convolutional layer and fully connected ROI pooling layer is added between layers to adapt to candidate frames of any scale and improve the detection effect; for the time-consuming problem of candidate frame selection, Faster R-CNN uses a region selection network (RPN) to generate regional frames to improve the detection speed; 2017 Dai et al. The deformable convolution network (DCN) is proposed, and the deformable convolution and deformable ROI pooling layer are designed to improve the shape of the receptive field and enhance the positioning ability. In summary, the two-stage detection network has high detection accuracy, but the candidate area selection is time-consuming and cannot meet the real-time detection requirements.

The current classic single-stage detection algorithms include SSD [7], YOLO [8], etc. SSD uses VGG-16 to extract features, but when fusing features, the scale information is independent of each other, and other scale information cannot be fully utilized, and the detection effect is poor. In order to strengthen the connection between different feature layers of SSD, Cheng et al. [9] proposed the DSSD algorithm in 2017, Design a prediction module similar to the residual block to fully mine the feature information of each layer; YOLOv3 [10] uses the residual block to extract features and uses FPN for information fusion, which can simultaneously meet the accuracy and speed requirements of real-time detection, but there are still parameters Excessive detection speed and accuracy need to be improved. Recently, Blochkovskiy et al. [11] proposed the YOLOv4 algorithm, which introduced the SPP structure to improve the size of the receptive field, used PANet instead of FPN for multi-channel feature fusion, and selected CSP-Darknet53 as the backbone network Further improve the detection accuracy and speed, but the model parameters are too large, the weight file is too large, and deployment is difficult.

In summary, ensuring real-time detection and high detection accuracy at the same time is an important basis for the performance indicators of target detection algorithms. Especially for aerial infrared aircraft detection tasks, the target moves fast and requires high detection accuracy, so this paper uses the lightweight version of YOLOv4 YOLOv4-tiny algorithm to detect aerial infrared aircraft targets. In order to improve the detection accuracy of the algorithm, the following improvements are made:

1) Add a branch after the feature layer whose output size of YOLOv4-tiny is 26*26*128. After DBL transformation and then 1*1 convolution to adjust the size, it is fused and added with the original feature map, and the fusion result is the same as the original The 13*13*256 output feature maps of the feature map are added together as the first feature output layer;

2) After the first network output layer, add three parallel depth separable convolutional layers with a hole rate of 1, 3, and 5 respectively to expand the feature map receptive field. After the addition and fusion, 1*1 convolution Adjust the size, and then output the second feature layer after the DBL layer;

3) Improve the feature fusion network. By adding 1*1 convolution before and after the depth separable convolution, the dimension first increases and then decreases, constructing an improved anti-residual block, the first output feature layer, and the improved anti-residual After the feature is extracted from the difference block, a part is adjusted by convolution and processed by yolo_head to output the first prediction result, and the other part is upsampled and combined with the second feature map, and the second prediction result is output after yolo_head processing.

2. YOLOv4-tiny network
The YOLOv4-tiny model is a lightweight version of the YOLOv4 model. It is a lightweight real-time detection algorithm suitable for embedded platforms. Compared with YOLOv4, the detection accuracy is reduced, but it achieves model compression while greatly improving Detection speed.

2.1. YOLOv4-tiny algorithm structure
Its backbone network CSPDarknet53-Tiny is a feature extraction network composed of three DBL modules and three Resblock_body blocks stacked. It uses two scales to predict the target. After the prediction is completed, non-maximum values are introduced to suppress redundant detection frames. So that each target has a unique prediction frame, and its structure is shown in Fig. 1. The DBL layer is composed of a convolutional layer, a normalization processing layer, and an activation layer group. The
Resblock body block is composed of many small residual blocks and a large residual edge. The original residual block stack is split and divided into left and right parts, the main part continues to stack the residual blocks, and the other part is used as a large residual edge and directly connected to the end. The structure comparison diagram of this structure and the conventional residual block is shown in Fig. 2.

Fig. 1 YOLOv4-tiny principle diagram

Fig. 2 Comparison of anti-residual block and conventional residual block

2.2. Network loss function
The loss function of YOLOv4-tiny is the same as that of YOLOv4. It is based on the loss function of YOLOv3 and introduces CIOU loss instead of MSE mean square error loss as the positioning loss of the prediction frame, DIOU [12] loss considers both the overlapping area of the bounding box and the distance between the center points of the two boxes, making it more accurate in positioning the target, while the CIOU loss takes into account the consistency of the aspect ratio of the two boxes based on the DIOU loss. The effect of positioning. The loss function of YOLOv4-tiny is shown in formula 1:

\[ \text{loss(object)} = L_1 + L_2 + L_3 \] (1)
It is divided into positioning loss, confidence loss and classification loss, where and are calculated as shown in the following formula:

\[ L_{\text{pos}} = -\sum_{i=0}^{5} \sum_{j=0}^{5} I_{i,j}^{\text{obj}} \left[ C_i \log(C_i') + (1 - C_i') \log(1 - C_i') \right] - \lambda_{\text{noobj}} \sum_{i=0}^{5} \sum_{j=0}^{5} I_{i,j}^{\text{obj}} \left[ C_i \log(C_i') + (1 - C_i') \log(1 - C_i') \right] \] (2)

The confidence loss in formula (2) is calculated in the form of binary cross entropy, which contains two parts: the box confidence loss with the target and the confidence loss without the target. To judge whether the j-th box in the i-th grid is responsible for this goal.

\[ L_{\text{conf}} = -\sum_{i=0}^{5} I_{i,j}^{\text{obj}} \sum_{c=\text{classes}} \left[ p_j(c) \log(p_j^*(c)) + (1 - p_j(c)) \times \log(1 - p_j^*(c)) \right] \] (3)

Equation (3) is the category loss, used to determine whether the center of the object falls in the grid i, and the function value is also calculated in the form of binary cross entropy. The CIOU calculation diagram is shown in Fig. 3:

![CIOU calculation diagram](image)

The loss of CIOU positioning is as follows:

\[ L_1 = 1 - IOU + \frac{\rho^2(b, b')}{c^2} + \alpha \nu \] (4)

Among them, C is the diagonal distance surrounding the smallest outsourcing area of the two frames, is the Euclidean distance between the center points of the two frames, is a positive number, is a parameter used to measure the consistency of the aspect ratio, and is defined as follows:

\[ \alpha = \frac{\nu}{(1 - IOU) + \nu} \quad \nu = \frac{4}{\pi^2} \left(\arctan \frac{w_{\text{gt}}}{h_{\text{gt}}} - \arctan \frac{w'}{h'}\right)^2 \] (5)

In formula 5, and are the width and height of the real frame, and are the width and height of the prediction frame. If the width and height of the real frame and the prediction frame are similar, this time is 0, and the penalty term does not work. When it is not 0, the larger the value, the larger the value of, so the function of the penalty term is to control the width and height of the predicted frame to be as close as possible to the aspect ratio of the real frame.

It can be seen from the model structure that the YOLOv4-tiny algorithm reduces the size of the model through a smaller number of network layers, making the model detection more real-time, suitable for the deployment of embedded platforms. However, the model extracts the last two features for detection, resulting in low utilization of shallow features, insufficient scale span, and missed detection of small targets; in addition, the FPN network used to fuse features is too simple, making the model to extract the fusion of the two layers of features is insufficient. Aiming at the above two problems, this paper improves YOLOv4-tiny from two aspects: feature extraction network and feature fusion network.

### 3. Improve YOLOv4-tiny's target detection algorithm

#### 3.1. Improved feature extraction network

This section addresses the problem that model's feature extraction network CSPDarknet53-tiny cannot make full use of the information of each scale. The low-level features (128) are transformed by DBL, and the size is adjusted by 1*1 convolution and merged with the previous features, and the fusion result is combined with the middle-level features (256) as the first feature network output layer to enhance the model’s feature extraction ability for small targets; at the same time, after the first feature output layer,
three parallel connections with different void rates are introduced. The depth of the separable hollow convolutional layer is to expand the feature map receptive field. The fusion result is adjusted by 1*1 convolution and used as the second feature network output layer. The improved feature extraction network structure diagram is shown in Fig. 4, where SepConvBN is a convolution block containing depth separable hole convolution [13]:

![Improved Feature Extraction Network Structure Diagram](image)

**Fig. 4** The improved feature extraction network structure diagram

Hole convolution is a kind of abnormal convolution. Hole convolution increases the receptive field of the model without increasing the amount of calculation. The idea is to expand the distance between each convolution kernel pixel and change the distance from the original 1 becomes greater than 1, and this distance is called the rate of voids (rate). As shown in Fig. 5, when the convolution kernel is, the difference between normal convolution and hole convolution:

![Comparison of Normal Convolution and Hole Convolution](image)

(a) Normal convolution  (b) Hole convolution (rate=2)

**Fig. 5** Comparison of normal convolution and hole convolution

Fig. 5 (a) is a normal convolution, and Fig. 5 (b) is a hole convolution with a void ratio of 2. Compared with the former, the receptive field of the latter is significantly increased. However, while the hole convolution expands the receptive field, some pixel values in the image do not participate in the convolution operation, so that the convolution has the phenomenon of information loss; to solve this problem, when designing three parallel expansion convolutions in this section, The void rate is set to 1, 3, and 5 respectively, that is, the small void rate focuses on short-distance information, and the large
void rate focuses on long-distance information. At the same time, the hole convolution is replaced with a depth separable hole convolution to reduce the amount of parameter calculation.

3.2. Improved feature fusion network

In YOLOv4-tiny, the FPN structure is used to perform simple feature fusion on the two effective feature layers of the output. The process is as follows: the last effective feature layer is convolved and then up-sampled, on the one hand, the first prediction is output by yolo_head processing as a result, on the other hand, it is stacked with the effective feature layer of the previous output, and then processed by yolo_head to output the second prediction result. In this section, based on the FPN of the YOLOv4-tiny network, an improved anti-residual block is introduced to enhance the feature fusion capability of FPN. Fig. 6 shows the structure comparison diagram of the standard residual block, the de-residual block and the improved de-residual block, where is the number of input channels, is the number of output channels, and is the multiple of the compression or expansion channel.

![Fig. 6 Comparison of structure of standard residual block, anti-residual block and improved anti-residual block](image)

The residual network uses jump connections to solve the degradation problem in the neural network. It can be seen from the comparison of Fig. 6 that the standard residual block (Fig.6 a) first compresses the number of channels through 1*1 convolution, and then uses 3*3 volume in the low channel the feature is extracted by product, and finally the channel number of the feature map is restored through 1*1 convolution. The feature channel is first compressed and expanded. During the compression process, there is a problem of damaging the feature expression and causing information loss; while the anti-residual block (Fig.6 b) on the contrary, first expand the channel through 1*1 convolution, then use 3*3 convolution to extract features, and finally use 1*1 convolution to map to the original number of channels; in order to deepen the network depth and reduce the computational complexity, this section uses improvements the latter anti-residual block (Fig.6 c) is improved, and the 3*3 conventional convolution in Fig.6 b is replaced with a 3*3 deep convolution.

The structure diagram of the feature fusion network designed in this chapter is shown in Fig. 7:
It can be seen from the figure that the last layer of the feature extraction network outputs a feature map with 512 channels. After two improved inverse residual blocks, one branch is convolved and then decoded by yolo_head to output the first prediction. As a result, after convolution and upsampling, another branch is added to the feature map of the penultimate layer with 256 channels. As a result, the feature is extracted by two improved anti-residual modules, and then sent to yolo_head performs decoding processing and outputs the second prediction result. The improved overall algorithm structure diagram is shown in Fig. 8:

**Fig. 7 Improved feature fusion network**

4. Experimental analysis
The air infrared aircraft data set used in the experiment has 6 categories, including 5024 training images, 56 verification images, and 502 test images. The category names are as follows: BAF (back to fuselage), BAT (back to tail flame), LAF (lateral fuselage), LAT (lateral tail flame), BWF (backward fuselage), BWT (backward tail flame), this data set category is all small targets, and the target itself is susceptible to background interference, detection The difficulty is greater. A single image contains 3.8 targets on average, and the type distribution is shown in Table 1:
Table 1. Type distribution of infrared data set

| Classes | BAF | BAT | LAF | LAT | BWF | BWT |
|---------|-----|-----|-----|-----|-----|-----|
| number  | 3385| 2730| 6438| 3155| 352 | 4904|

The experimental parameter settings are shown in Table 2:

| Learning rate | Learning rate decay | Early termination | Stride |
|---------------|---------------------|-------------------|--------|
| stage 1 (0-50) | stage 2 (50-100)    | factor patience   |        |
| 10^-3          | 10^-4               | 0.5               | 6      | 10   | 4      |

The following table 3 shows the comparison of the detection indicators between the YOLOv4-tiny algorithm and the improved YOLOv4-tiny algorithm designed in this chapter on the air infrared aircraft data set and the visible light roller data set.

Table 3. Test results of air infrared aircraft data set

| Network model     | Average accuracy (mAP) | Frame rate (frame/s) | Weight (MB) |
|-------------------|------------------------|----------------------|-------------|
| YOLOv3            | 80.44%                 | 6.67                 | 235         |
| YOLOv4-tiny       | 76.14%                 | 30.37                | 22.6        |
| Improved YOLOv4-tiny | 80.43%              | 28.18                | 23.3        |

It can be seen from Table 3 that on the infrared aircraft data set, the detection accuracy of the YOLOv3 model is higher, but the detection speed is slow, while the detection speed of YOLOv4-tiny is faster, 4.55 times that of YOLOv3, and the detection accuracy is only 4.3 points different from YOLOv3 Percentage points, and the weight file size is only about one-tenth of the former. The model is very easy to deploy, and it has achieved a good balance between detection speed and accuracy.

Compared with YOLOv4-tiny, the improved YOLOv4-tiny method in this paper increases the model volume by 0.7 megabytes and the detection frame rate drops by about 2 frames. At the same time, the detection accuracy is increased by 4.29%. The main reason is the increased fusion in the feature extraction network. The branch and three parallel depths can separate the convolutional layer with holes, which can better extract features. The addition of improved anti-residual blocks enhances the model's utilization of features and improves accuracy; while the detection frame rate decreases, because as the model complexity increases, the model reasoning time has increased, but the detection speed of 28.18 frames still meets the needs of real-time detection.

The following figure shows the comparison between the algorithm proposed in this article and the original YOLOv4-tiny algorithm on the air infrared aircraft data set:
5. Summary

In order to improve the detection accuracy of infrared aircraft in the air, an improved YOLOv4-tiny infrared aircraft detection algorithm is proposed based on the YOLOv4-tiny algorithm. First, add a branch after the feature layer whose output size of YOLOv4-tiny is 26*26*128, and merge it with the 13*13*256 feature map of the original feature map after convolution to strengthen the feature extraction capability of the model; Secondly, three parallel expansion ratios of 1,3,5 are used to expand the receptive field of the model with separable expansion convolutional layers to extract more effective features; then, the feature fusion network is improved, and the first output After the feature layer is added, the constructed improved anti-residual block is added, so that the feature map is operated in high dimensions and then the size is adjusted to increase the model's ability to fuse small target features; finally, on the air infrared aircraft small target data set The experimental verification shows that the improved YOLOv4-tiny network has a significant improvement in the detection ability of small targets compared with the original network.

References

[1] Q. Liyong, X. Lixin, G. Min. The Influence of Infrared Image Complexity on Target Detection Performance [J]. Infrared and Laser Engineering,2013, 42(1): 253-261.
[2] R. Girshick,J. Donahue,T. Darrell, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. IEEE Conference on Computer Vision and Pattern Recognition,2014,580-587.
[3] R. Girshick. Fast r-cnn[C]. IEEE International Conference on Computer Vision,2015,1440-1448.
[4] S.Q. Ren,K.M. He,R. Girshick, 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.
[5] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, Yichen Wei; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 764-773.
[6] J. Redmon,S. Divvala,R. Girshick, et al. You only look once: Unified, real-time object detection[C]. IEEE Conference on Computer Vision and Pattern Recognition,2016,779-788.
[7] W. Liu,D. Anguelov,D. Erhan, et al. SSD: Single shot multibox detector[C]. European Conference on Computer Vision, Cham,2016,21-37.
[8] J. Redmon,S. Divvala,R. Girshick, et al. You only look once: Unified, real-time object
[9] DSSD: Deconvolutional Single Shot Detector. FU C Y, LIU W, RANGA A.et al. https://www.researchgate.net/publication/312759848_DSSD_Deconvolutional_Single_Shot_Detector. 2017.

[10] REDMONJ, FARHADI A. YOLOv3: an incremental improvement [EB/OL].(2018-4-8)[2019-6-29].

[11] Bochkovskiy A, Wang CY, Liao H Y M. YOLOv4: Optimal Speed and Accuracy of Object Detection[C]/IEEE conference on Computer Vision and Pattern Recognition.2020. arXiv:2004.10934v1[ cs. CV].

[12] DistanceIoU Loss: Faster and Better Learning for Bounding Box Regression. Zheng Zhaohui, Wang Ping, Liu Wei,et al. https://arxiv.org/abs/ 1911 08287 . 2019.

[13] Yu F, Koltun V. Multi-scale Context Aggregation by Dilated Convolutions. [C]. Proceedings of the International Conference on Learning Representations,2016:23-24.