

## RESEARCH ARTICLE

# YOLO-Extract: Improved YOLOv5 for Aircraft Object Detection in Remote Sensing Images

ZHIGUO LIU<sup>1</sup>, YUAN GAO<sup>1</sup>, QIANQIAN DU<sup>2</sup>, MENG CHEN<sup>1</sup>, AND WENQIANG LV<sup>1</sup><sup>1</sup>Communication and Network Key Laboratory, Dalian University, Dalian 116622, China<sup>2</sup>College of Engineering Physics and Optoelectronics, Taiyuan University of Technology, Taiyuan 030024, China

Corresponding author: Zhiguo Liu (liuzhiguo\_dldx@163.com)

**ABSTRACT** Compared with natural images, remote sensing targets have small and dense target shapes as well as complex target backgrounds. As a result, insufficient detection accuracy and target location cannot be accurately identified. So, this paper proposes the YOLO-extract algorithm based on the YOLOv5 algorithm. Firstly, The YOLO-extract algorithm optimized the model structure of the YOLOv5 algorithm. The YOLO-extract algorithm not only deleted the feature layer and prediction head with poor feature extraction ability but also a new feature extractor with stronger feature extraction ability was integrated into the network. At the same time, YOLO-extract borrowed the idea of residual network to integrate Coordinate Attention into the network. Secondly, the mixed dilated convolution was combined with the redesigned residual structure to enhance the feature and location information extraction ability of the shallow layer of the model and optimize the feature extraction ability of the model for different scale targets. Finally, drawing on the idea of  $\alpha$ -IoU Loss, Focal- $\alpha$  EIoU Loss was designed to replace CIoU Loss, which makes the model bounding box regression faster and the loss lower. The experimental results on the test data set show that compared with the YOLOv5 algorithm, the YOLO-extract algorithm has a faster convergence speed, reduces the calculation amount by 45.3GFLOPs and the number of parameters by 10.526M, but increases the mAP by 8.1% and the detection speed by 3 times.


**INDEX TERMS** Remote sensing aircraft target, YOLOv5, structure optimization, dilated convolution, focal- $\alpha$  IoU loss.

## I. INTRODUCTION

In recent years, with the development of remote sensing technology, the information content of satellite remote sensing images has increased dramatically, which plays an increasingly obvious role in military applications. So in object detection, remote sensing object detection has become one of the key topics. As an important means of transportation and military equipment, the use of target detection algorithms to locate and identify aircraft is of great significance to airport monitoring and management, military intelligence analysis, and military action decision making. However, the remote sensing targets are collected at high altitudes, so the target size is usually small and easily affected by various factors such as weather, illumination, sea conditions, sensor parameters, etc. In addition, aircraft targets in remote sensing images are

usually densely arranged, which makes it difficult to separate the targets in satellite remote sensing images from the surrounding background, resulting in more difficult feature extraction, low detection accuracy, and failure to meet the requirements of real-time detection.

Several solutions have been proposed for the above target detection problems. Traditional target detection is mainly based on machine learning, but with the development of deep learning, the field of computer vision has brought new changes and developments to target detection and image classification. There have also been great advances in object detection in the above remote sensing images. At present, the mainstream target detection algorithms are mainly divided into two categories: two-stage algorithms and one-stage algorithms. Typical two-stage algorithms are R-CNN (Region Convolutional Neural Networks) [1], Fast R-CNN (Fast Region-Based CNN) [2], and Faster R-CNN (Faster Region-Based CNN) [3]. The two-stage algorithm has

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson .

higher accuracy but a slower speed and loses the spatial information of local objects in the whole image. Typical one-stage algorithms include SDD (Single Shot MultiBox Detector) [4], YOLO (You Only Look Once) series [5], [6] [7], [8] etc. The accuracy of the one-stage algorithm is average, but the detection speed is high.

At present, in the detection of satellite remote sensing images, great progress has been made in both detection accuracy and speed. Reference [9] proposed to use DenseNet [10] to improve YOLOv3 [7] and improve the detection accuracy of remote sensing images by improving the structure in Backbone, but the structure of DenseNet is too complex and the number of parameters is too large, which leads to a decrease in detection speed. Reference [11] made a lightweight improvement on the structure of YOLOv3 [7] and introduced Res2Net [12] to improve the detection accuracy and speed of remote sensing targets. Reference [13] added the PPM(Pyramid Pooling Module) [14] based on YOLOv4 [8] and used the Mish function to replace the original activation function, which improved the detection accuracy and recall rate of aircraft and dockyard in remote sensing images. Reference [15] proposed YOLOv5-aircraft based on YOLOv5, the smooth Kullback-Leibler divergence loss function was used to replace the cross entropy loss function, and the CSandGlass module was designed to replace the residual module, which improved the accuracy and speed of aircraft targets in remote sensing images.

Given the problems of the YOLOv5 algorithm in identifying aircraft targets in remote sensing images, such as low accuracy, slow detection speed, and difficulty in target feature extraction, this paper proposes the YOLO-extract algorithm based on the analysis of the model structure of the YOLOv5 algorithm. The YOLO-extract algorithm optimizes the YOLOv5 model structure and introduces a coordinate attention mechanism to improve the detection accuracy of aircraft targets. Secondly, the dilated convolution and residual structure are introduced to improve the feature extraction capabilities of the model, and finally, the improved loss function is used to accelerate the convergence speed of the model. In the second part of this paper, based on the analysis of the model structure of YOLOv5, the design scheme of the YOLO-extract algorithm is described; in the third part, the composition of the experimental data set, the experimental environment, and the comparative analysis of the simulation results are described; the final discussion research conclusions and future research directions are presented.

## II. MODEL INTRODUCTION AND IMPROVEMENT

### A. YOLOv5 MODEL INTRODUCTION

YOLOv5 is a one-stage target detection algorithm with five versions of n, s, m, l, and x, and each version has a different network depth and width. The YOLOv5 algorithm model includes four parts: input, backbone extraction network, neck, and detection head. Its network structure diagram is shown in Fig.1. Due to its deep network structure, the YOLOv5 model has a large loss of semantic information and position

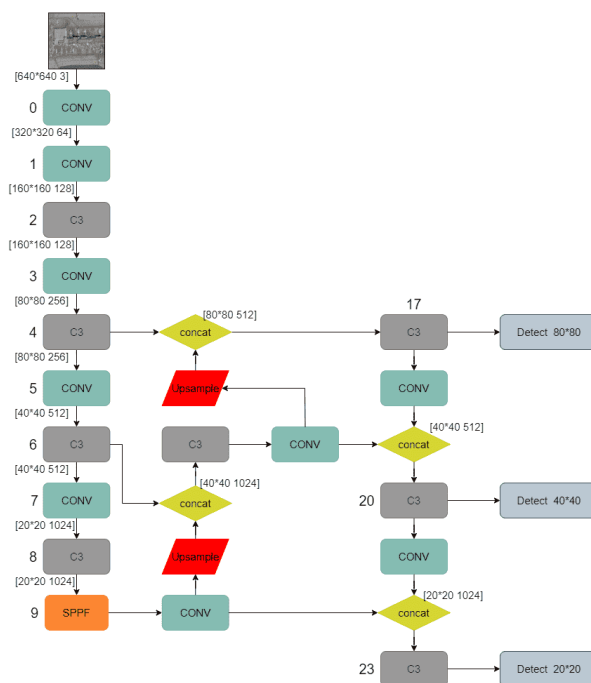


FIGURE 1. YOLOv5 model.

information of the target, resulting in a large loss in the bounding box regression and a slow detection speed, which is not suitable for remote sensing aircraft target detection. Therefore, this paper designs a YOLO-extract algorithm based on the analysis of the YOLOv5 6.0 algorithm.

### B. MODEL STRUCTURE IMPROVEMENT

#### 1) SIMPLIFIED MODEL STRUCTURE

In CNN, low-level feature maps have high resolution, rich location information, and less semantic information. High-level feature maps have a small resolution, less location information, and rich semantic information. In the YOLOv5 model, Backbone reduces the original image from  $640 \times 640$  to  $20 \times 20$  after three down-sampling for feature fusion. Although continuously shrinking the feature map can make the neuron have a larger receptive field and rich semantic information, for the aircraft target in the remote sensing image in this paper, because it occupies fewer pixels in the image, the features after continuous down-sampling of the location information in the figure is greatly lost, causing the model to fail to notice the target. Moreover, the semantic loss caused by continuous down-sampling will have the opposite effect on the feature learning and detection tasks of aircraft targets. as shown in Fig.2.

As shown in Fig.2, the first line is the attention of the  $80 \times 80$  resolution prediction head in the YOLOv5 model to the aircraft target, the second line is the attention of the  $40 \times 40$  resolution prediction head, and the third line is the  $20 \times 20$  resolution prediction head. It can be seen from Fig.2 that only the prediction head of  $80 \times 80$  resolution in the YOLOv5 model can pay attention to the aircraft target.

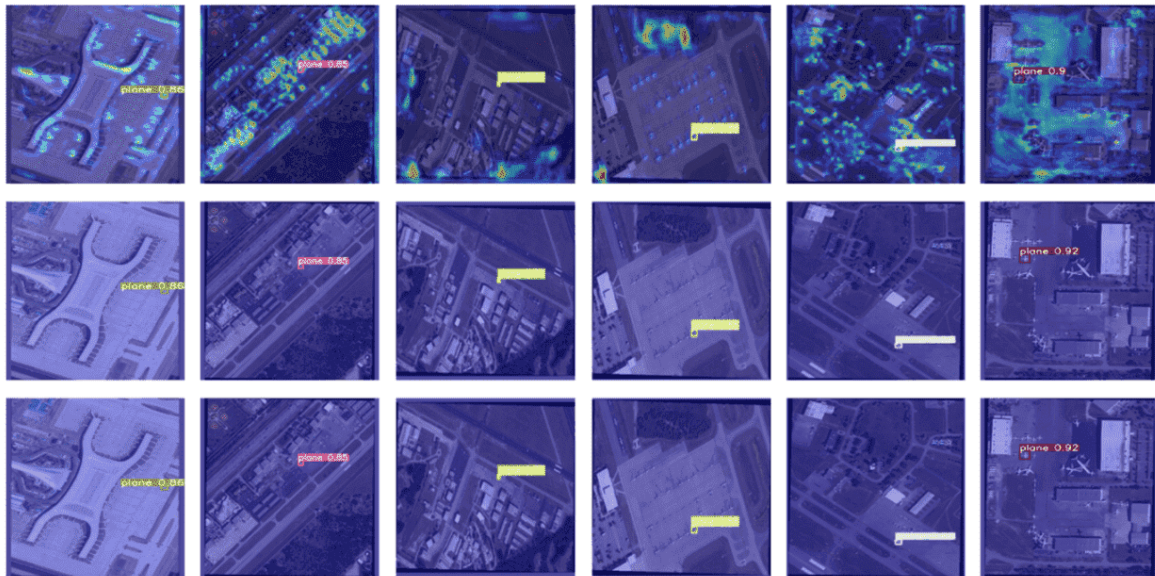


FIGURE 2. The attention of different detection heads to the targets.

Due to the large multiple of down-sampling, the aircraft target in the feature map is seriously lost, and the other two prediction heads cannot identify the aircraft target and detect the position of the aircraft, resulting in a poor detection effect of the YOLOv5 model for small targets. In addition, in the feature fusion based on multi-scale, due to the irreversibility of down-sampling, the lost information cannot be recovered by up-sampling, so the model is disturbed by the feature layer with lower resolution, and the prediction accuracy is low.

Therefore, the prediction head with a resolution of  $40 \times 40$  and a resolution of  $20 \times 20$  is pruned in the model, and the number of down-sampling in Backbone is reduced at the same time, the feature extractor with 32 times down-sampling and 16 times down-sampling is deleted. To prevent the low-resolution feature maps in Neck from interfering with the model during feature fusion and reduce semantic loss, the low-resolution convolution module and up-sampling module in Neck are deleted. It not only reduces the loss of down-sampling in target feature extraction but also greatly reduces the number of model parameters.

### 2) NEW FEATURE EXTRACTOR AND PREDICTION HEAD

After the above mentioned simplification of the model structure, only the  $80 \times 80$  resolution feature extractor in the model extracts the target features, and more comprehensive location information and semantic information cannot be extracted. Therefore, a 4 times down-sampling feature extractor is added to the model Backbone to enhance the feature extraction of aircraft targets. An up-sampling module is added to the feature fusion device to perform feature fusion with the new feature extractor, Conveying richer semantic information and location information. And a new prediction head is added to the fused feature map to enhance the detection of aircraft targets. As shown in Fig.3.

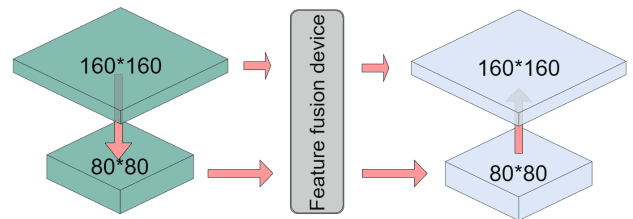


FIGURE 3. The new feature extractor.

The YOLOv5 model uses a combination of FPN (Feature Pyramid Network) [16] and PAN (Path Aggregation Network) [17]. FPN transfers high-level semantic features and PAN transfers low-level location information to the deep. Therefore, adding a new feature extractor can not only better extract the features of aircraft targets but also transfer more semantic information and location information. The optimized YOLOv5-extract model structure is shown in Fig.4. The optimized model uses a more sensitive feature extractor to extract the features of the aircraft target during the training process, which can more accurately predict the position of the target and reduce the difficulty of detection. Parameters dropped from 46.117M to 35.518M, the detection accuracy increases from 0.878 to 0.922.

### 3) COORDINATE ATTENTION MECHANISM

After the above modification of the model structure, the accuracy of the detection of small targets and dense targets in the model has been improved significantly. To further improve the detection accuracy of the model in complex backgrounds, Coordinate Attention [18] (referred to as CoordAttention, hereinafter referred to as CA) is integrated into the network model. Allowing the network to focus on a larger area reduces computational overhead

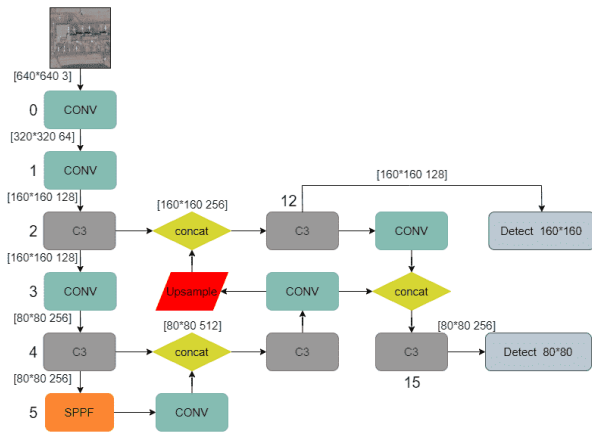


FIGURE 4. The improved model structure.

while improving the model’s recognition accuracy for aircraft targets with complex backgrounds. CA uses the tensor  $M = [m_1, m_2, \dots, m_c] \in R^{C \times H \times W}$  to embed coordinate information and generate coordinate attention, and outputs the tensor  $N = [n_1, n_2, \dots, n_c] \in R^{C \times H \times W}$  with enhanced representation ability. It can not only capture cross-channel information but also capture direction perception and position perception information, which can help the model locate and identify aircraft targets more accurately. Therefore, before the model performs feature fusion, we draw on the idea of a residual network and integrate CA into the model, as shown in Fig.5. It can not only fuse the original feature information in Backbone but also pay more attention to dense aircraft targets and aircraft targets in complex backgrounds so that the detection accuracy is further improved. Moreover, it is more accurate for the detection of dense aircraft targets and complex backgrounds, the detection accuracy increases from 0.922 to 0.928, and the number of parameters hardly increases.

The improved model uses two detection heads to detect aircraft targets in remote sensing images. The attention to aircraft targets is shown in Fig.6. The model can better extract the features of aircraft targets in remote sensing images and can pay attention to a larger area for dense targets, and reduce a lot of semantic loss so that the model is reduced by the influence of high-level feature maps.

C. DILATED CONVOLUTION AND RESIDUAL

Dilated convolution [19] was originally proposed to solve the problem of image segmentation, specifically to expand the receptive field to obtain denser data and aggregate multi-scale context information. Dilated convolution eliminates down-sampling and up-sampling in the network and expands the receptive field by expanding the convolution kernel. However, in the YOLOv5 algorithm, to detect targets of different scales, it is necessary to down-sample through convolution operations to reduce the feature map to increase the receptive field, and then up-sample to restore the image size. The process of shrinking and enlarging the feature map

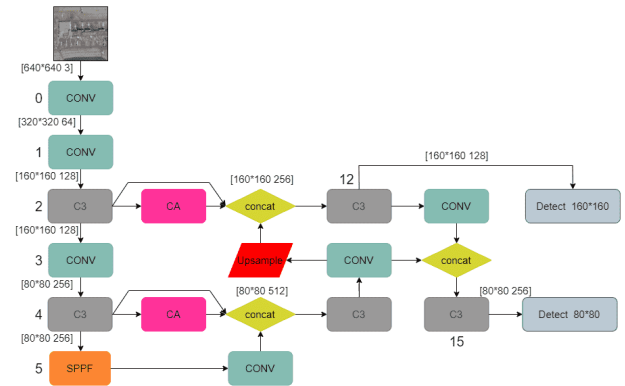


FIGURE 5. Yolo-extract model structure diagram.

results in a loss of accuracy, and the position information of many aircraft targets will be lost, resulting in low accuracy of aircraft target detection. After the above improvement of the model structure, the semantic loss of the model is greatly reduced and the detection accuracy is improved. However, the semantic information of the upper layer feature extractor is relatively less, and the location information is rich. Therefore, to balance the contradiction between multi-scale target detection and accuracy loss, this paper redesigns a new structure that combines mixed dilated convolution and residual [20] to expand the receptive field of the shallow feature extractor in Backbone, which can extract more target features while passing shallow rich location information. It makes up for the disadvantage that the shallow feature extractor in the CNN network has rich location information but insufficient semantic information, and does not suffer from the grid effect caused by single dilated convolution. At the same time, The model can not only extract the detailed features of the target, but also detect the target with different scales. Its structure is shown in Fig.7.

The above structure is applied to the Backbone of YOLOv5, and the model structure is shown in Fig.8.

As shown in Fig.8, R6 has six groups of residual structures fused with dilated convolution, and the receptive field of the second layer is expanded from 3 in YOLOv5 to 3, 7, and 15. Since the receptive field of the fourth layer has been expanded compared with that of the second layer, a smaller expansion rate is used to expand the receptive field from 3 to 3, 7, and 11. The receptive field is calculated as follows:

$$F_k = F_{k-1} + \left\{ (f_k - 1) \times \prod_i^{k-1} S_i \right\} \tag{1}$$

$F_k$  represents the receptive field of the  $k$ th layer,  $F_{k-1}$  represents the receptive field of the  $(k - 1)$ th layer,  $f_k$  represents the size of the convolution kernel of the  $k$ th layer, and  $S_i$  represents the step size of the  $i$ th layer. The receptive field changes of the second and fourth layers in the model are shown in Fig.9.

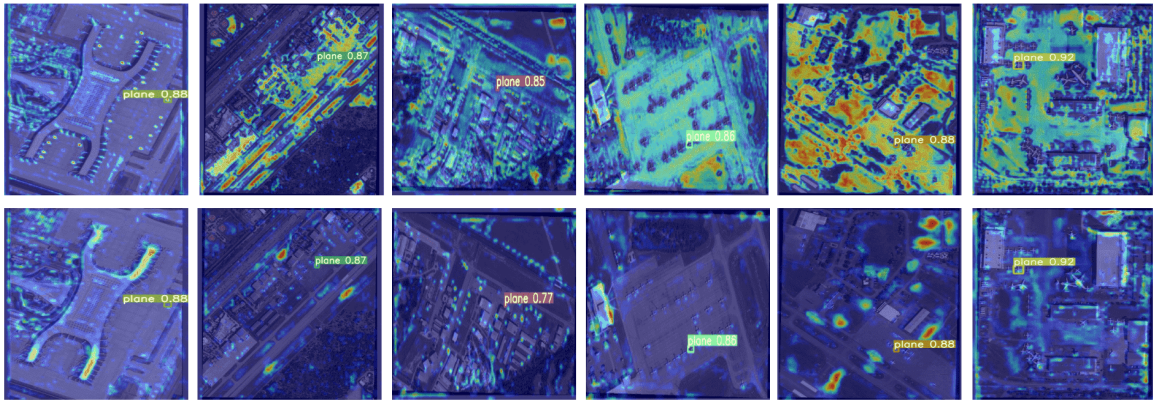


FIGURE 6. The attention of the two detection heads to the targets.

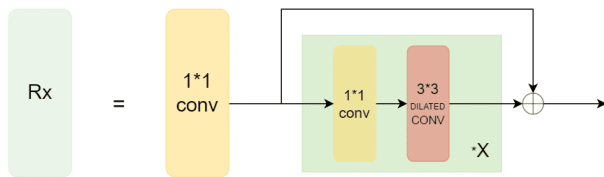
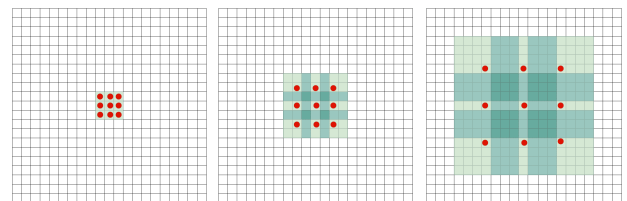
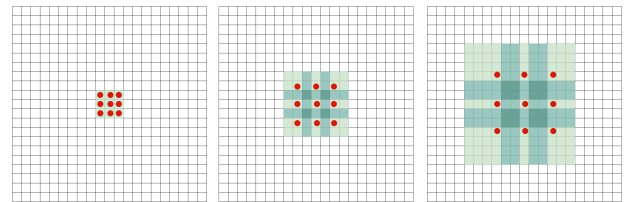


FIGURE 7. The module that fuses the mixed dilated convolution and residual structures is named Rx, and x represents the use of x groups of bottleneck residual structures. For the input feature map, first reduce the dimension through  $1 \times 1$  convolution, and then input it into the residual structure of the x group. In each group, the dimension is reduced again through  $1 \times 1$  convolution, and then through  $3 \times 3$  dilated convolution is used for feature extraction and dimension enhancement, and finally fusion.



(a) Receptive fields with dilation rates of 1, 2, and 4.



(b) Receptive fields with dilation rates of 1, 2, and 3.

FIGURE 9. Receptive field after using dilated convolution.

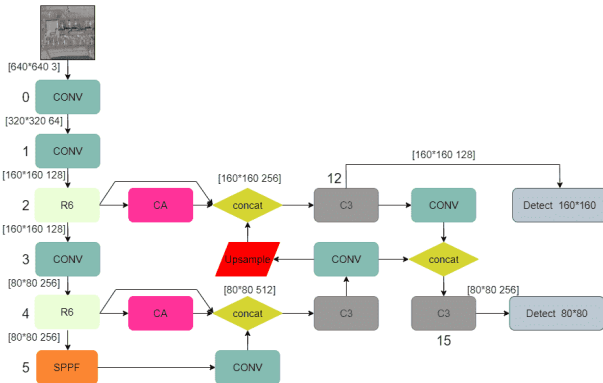


FIGURE 8. R6 is used in the Backbone of YOLOv5, the dilated convolution with expansion rates of 1, 2, 4, 1, 2, 4 is used in the second layer, the dilated convolution with the expansion rate of 1, 2, 3, 1, 2, 3 is used in the fourth layer.

D. FOCAL- $\alpha$  EIou LOSS

In object detection, bounding box regression is a key step to determine the performance of object localization. In the YOLOv5 model, CIoU Loss is used as the loss function of bounding box regression. Although CIoU [21] increases the Loss for the scale of the predicted box based on DIoU [21], that is, further increases the penalty term of the ratio of length to the width between the predicted box and the real box,

as shown in the formulas (2), (3):

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \left( \frac{v}{(1 - IoU) + v} \right) v \quad (2)$$

$$v = \frac{4}{\pi^2} \left( \arctan \frac{w^{gt}}{h^{gt}} - \arctan \frac{w}{h} \right)^2 \quad (3)$$

$b$  and  $b^{gt}$  represent the center point of the predicted box and the real box, and  $\rho$  represents the Euclidean distance between the predicted box and the center point of the real box.  $c$  represents the diagonal distance of the smallest closure area that can contain both the predicted box and the real box,  $w^{gt}$  and  $h^{gt}$  are the width and height of the real box, and  $w$  and  $h$  represent the width and height of the predicted box. However, CIoU describes the relative value of the aspect ratio, the width and height of the prediction box cannot be increased or decreased at the same time, and the balance of difficult and easy samples is not considered. EIou [22] separated the ratio of length to width on the basis of CIoU, and clearly measured the difference of three geometric factors, namely overlapping area, center point, and side length. Focal Loss [23] was



FIGURE 10. The same situation of IoU.

also introduced to solve the problem of imbalanced difficult and easy samples. The calculation method is shown in formulas (4), (5):

$$L_{EIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \frac{\rho^2(w, w^{gt})}{c_w^2} + \frac{\rho^2(h, h^{gt})}{c_h^2} \quad (4)$$

$$L_{FocalEIoU} = IoU^\gamma L_{EIoU} \quad (5)$$

Although EIoU fully considers various geometric characteristics between the prediction frame and the real frame, for remote sensing images, the aircraft target is not only small in scale and high in density, it will lead to the same situation of IoU, is shown in Fig.10. Resulting in no improvement of the convergence speed of the model.

Therefore, in this paper,  $\alpha$ -IoU [24] is combined with Focal EIoU, and the Euclidean distance between the calculated prediction box and the center point of the real box in the Focal EIoU Loss is introduced into a larger power transformation ( $\alpha = 3$ ), to The length and width introduce a smaller power transformation ( $\beta = 2$ ), named Focal- $\alpha$  EIoU Loss. This loss not only retains the advantages of Focal EIoU Loss but also improves the loss of High IoU target and the accuracy of gradient adaptive weighted bbox regression, providing stronger robustness for small target datasets. The calculation method is shown in formula (6):

$$L_{EIoU} = IoU^\gamma \left( 1 - IoU + \frac{\rho^{2\alpha}(b, b^{gt})}{c^2} + \frac{\rho^{2\beta}(w, w^{gt})}{c_w^2} + \frac{\rho^{2\beta}(h, h^{gt})}{c_h^2} \right) \quad (6)$$

$c_w$  and  $c_h$  respectively represent the width and length of the closure region formed by the predicted box and the ground real box.

### III. EXPERIMENTAL RESULTS AND ANALYSIS

#### A. EXPERIMENTAL DATA

The remote sensing images studied in this paper come from the aircraft target images in the DOTA dataset, including

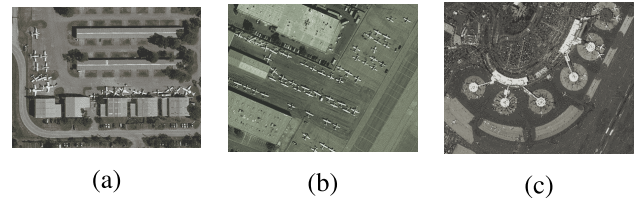


FIGURE 11. (a) dense target, occlusion target; (b) dense target; (c) complex background, small target, sparse target.

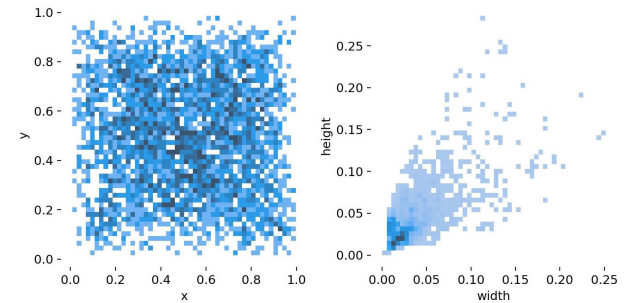


FIGURE 12. Target centroid position distribution and target size distribution.

300 remote sensing images and 7302 aircraft targets, and use Make Sense for data annotation. The image contains aircraft targets in complex backgrounds, dense small aircraft targets, and occluded aircraft targets. A typical remote sensing image in the dataset is shown in Fig. 11, Fig. 12 shows the distribution of centroid positions of aircraft targets in remote sensing images and the distribution of image sizes.

#### B. EVALUATION INDICATORS

In this paper, loss function curve, mAP, Recall, FPS (Frames Per Second), Parameters, and FLOPs (Floating Point Operations) are used as evaluation metrics to describe the performance of each network.

Precision represents the proportion of the number of targets that are correctly predicted among all targets and is calculated by the formula (7):

$$P = \frac{TP}{TP + FP} \quad (7)$$

Recall represents the number of correct samples detected in the prediction results, is calculated by the formula (8):

$$R = \frac{TP}{TP + FN} \quad (8)$$

But Precision and Recall are often a contradictory pair of performance metrics. Therefore, the mAP parameter is introduced to inherit two parameters to detect the algorithm performance of the network. It is calculated by the formula (9):

$$mAP = \frac{\sum_{k=1}^N P(k) \Delta R(k)}{C} \quad (9)$$

In the above formulas,  $TP$  represents the number of aircraft that are correctly predicted,  $FP$  represents the number of aircraft targets that are predicted but are actually backgrounds,

$FN$  represents the number of wrong predictions,  $C$  represents the total number of aircraft targets, and  $P(k)$  is the number of simultaneous recognition of  $k$  samples. The size of the correct rate,  $\Delta R(k)$  represents the change of Recall when the number of detection samples changes from  $(k - 1)$  to  $k$ .

**C. EXPERIMENTAL RESULTS AND ANALYSIS**

This paper mainly improves the YOLOv5 algorithm and obtains the YOLO-extract algorithm. To compare the YOLOv5 algorithm with the improved algorithm proposed in this paper and ensure the reliability of the experiment, all training and test data are trained in the same training environment, and the epoch of each training is guaranteed to be 300. After the training is completed, use the optimal model weights for testing.

In the training process of the model, the loss value of the model can directly reflect the convergence speed of the model and the accuracy of detection. There are confidence loss and localization loss in YOLO-extract. The confidence loss is used to calculate the confidence of the network, and the localization loss is used to calculate the gap between the predicted box and the true box. In this paper, the confidence loss and localization loss of the YOLOv5 model and the YOLO-extract model are compared, as shown in Fig.13.

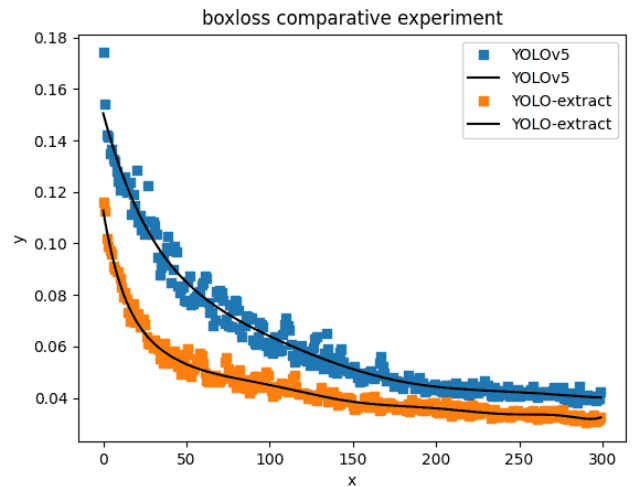
As shown in Fig.13, compared with the YOLOv5 model, both the confidence loss and the localization loss decrease with the increase of training times, and the loss value is closer to 0.

This paper also compared the positioning losses of the YOLO-extract model before and after using Focal EIoU Loss and Focal- $\alpha$  EIoU Loss, as shown in Fig.14. In addition, the loss curves of the YOLO-extract model were compared with the FE-YOLO model proposed in [11] and the YOLOv5-Aircraft model proposed in [15], as shown in Fig.15.

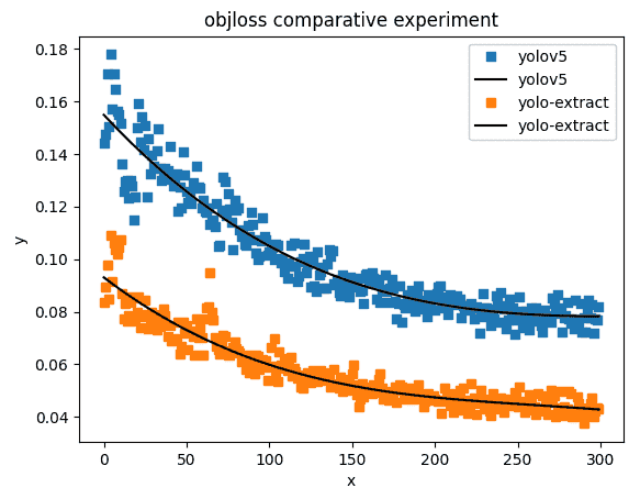
As shown in Fig.14, YOLO-eiou indicates that Focal EIoU Loss is used as the localization loss. The YOLO-extract model has a faster loss convergence speed and a smaller final loss value, which verifies the effectiveness of the improvement. As shown in Fig.15, the YOLO-extract model has a better convergence rate than existing methods and a lower loss value.

The mAP indicator is used to measure the quality of the detection algorithm. The larger the mAP, the higher the average detection accuracy and the better the performance of the algorithm. The Recall metric is used to represent the number of objects that the model can correctly detect. During the training process, the mAP value and Recall value of the validation set of the two algorithms during the training process is shown in Fig.16. The abscissa represents the number of training sessions, and the ordinate represents the mAP and Recall values.

As shown in Fig.16, After 300 rounds of training, the mAP of the YOLO-extract model reaches 0.968. After 300 rounds of training for the YOLOv5 model, the mAP is only 0.874.



(a) the location loss of the YOLOv5 model and the YOLO-extract model.



(b) the confidence loss of the YOLOv5 model and the YOLO-extract model.

**FIGURE 13. Loss comparison of YOLOv5 and YOLO-extract.**

**TABLE 1. Ablation experiment.**

Structure Improvement	Dilated Convolution and Residual	Focal- $\alpha$ EIoU Loss	mAP	Parameters	FLOPs
✓	-	-	0.928	35.518M	62.4G
-	✓	-	0.914	46.159M	107.9G
-	-	✓	0.893	46.117M	107.8G
-	-	-	0.878	46.117M	107.8G
✓	✓	✓	0.959	35.591M	62.5G

After 300 rounds of training, the recall of the YOLO-extract model reaches 0.925. After 300 rounds of training, the Recall of the YOLOv5 model is only 0.794.

During the training process, to better extract features and distinguish the background from the target, the confidence is set to be small, so the ablation experiment is conducted in this paper under the condition of 0.6 confidence level. mAP, Parameters and FLOPs of the model were tested before and after adding structure optimization, dilated Convolution and Residual, and Focal- $\alpha$  EIoU Loss. As shown in Table 1.

As shown in Table 1, after the structure optimization of the YOLOv5 model, the calculation amount and parameter

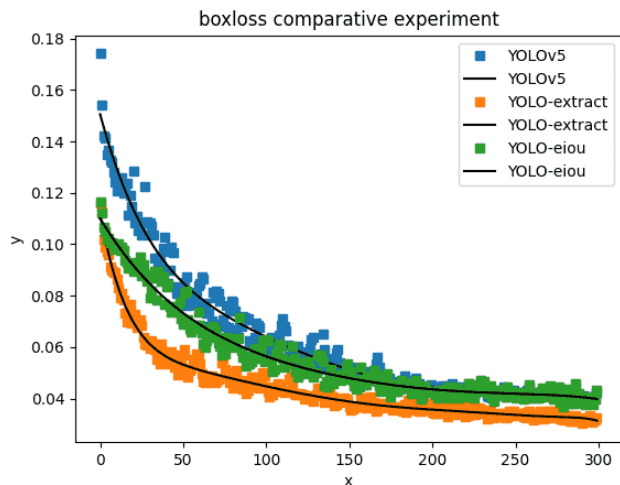


FIGURE 14. Loss comparison of EIoU and Focal- $\alpha$  EIoU.

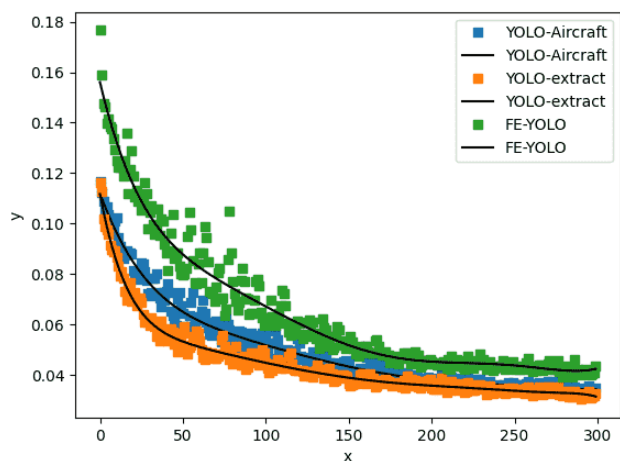
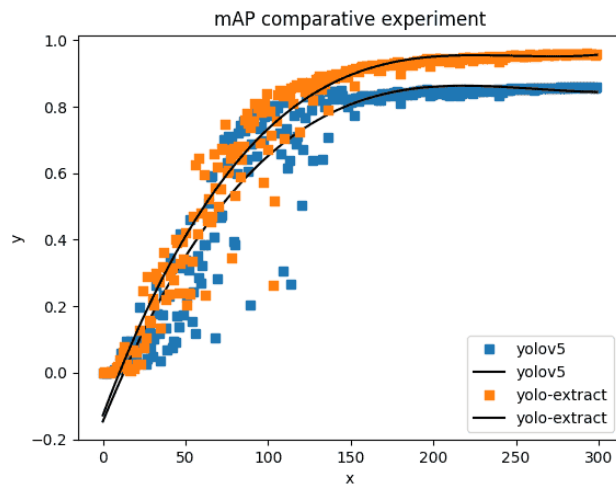


FIGURE 15. Loss comparison of different models.

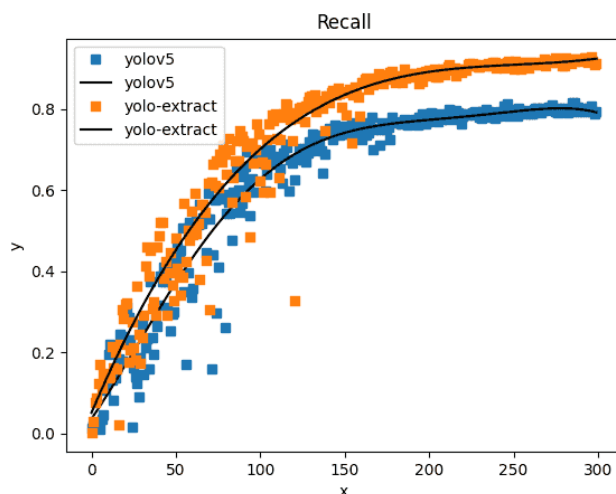
amount of the model is greatly reduced, and the Recall and mAP are significantly improved, indicating that the structure of the YOLO-extract model is more reasonable for the learning of small target features and the detection of small targets. After adding the redesigned dilated convolution and residual structure to the network, although the number of parameters and FLOPs are slightly increased, the mAP is improved. Finally, Focal- $\alpha$  EIoU Loss was used as location loss, and the detection accuracy could be improved without increasing the number of parameters and FLOPs.

In addition, this paper compares the mAP, Recall, Parameters, and FPS of YOLOv4, YOLOv4-tiny, YOLOv3, YOLOv3-SPP, YOLOv3-tiny, SSD, Faster RCNN, YOLOv5-Aircraft [15], FE-YOLO [11] and YOLOv7 [25] in the test environment: CPU-i5 8300H and GPU-1050Ti. as shown in Table 2.

As shown in Table 2, although the hardware devices in this article are limited, the YOLO-extract model is tested on devices with the same computing power, and it is better than the mainstream one-stage and two-stage algorithm in terms of mAP, Recall, Parameters, and FPS.



(a) the mAP comparison between the YOLOv5 model and the YOLO-extract model.



(b) the recall comparison between the YOLOv5 model and the YOLO-extract model.

FIGURE 16. mAP and Recall values of YOLOv5 model and YOLO-extract model.

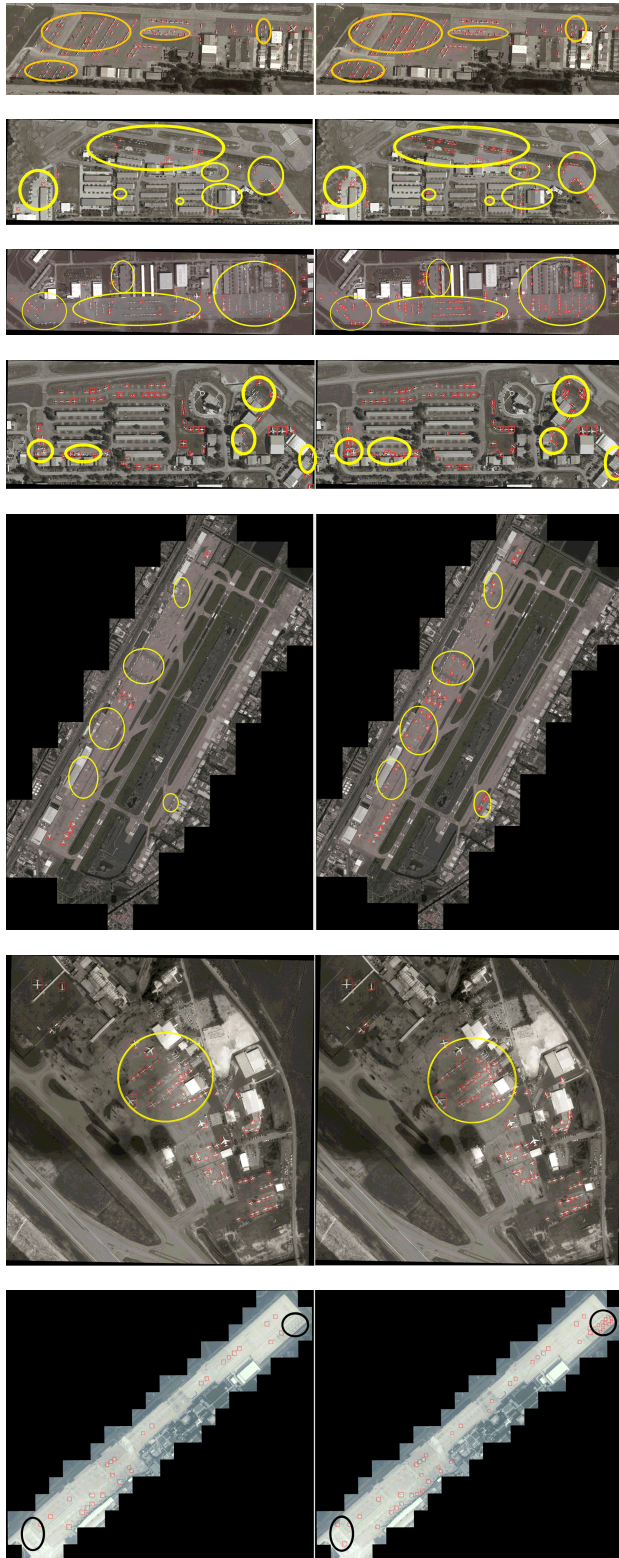
TABLE 2. Comparison of different algorithms.

Model	mAP	Recall	Parameters	FPS
YOLOv4	0.854	0.753	64.363M	5.215
YOLOv4-tiny	0.807	0.726	59.18M	15.524
YOLOv3	0.903	0.809	61.495M	2.778
YOLOv3-SPP	0.91	0.824	62.546M	4.272
YOLOv3-tiny	0.776	0.558	86.666M	12.821
SSD	0.784	0.793	-	15.873
Faster RCNN	0.749	0.786	-	1.934
YOLOv5	0.878	0.794	46.117M	11.905
YOLOv7	0.884	0.8	36.479M	19.361
FE-YOLO	0.883	0.878	50.154M	11.544
YOLOv5-Aircraft	0.903	0.897	52.632M	13.292
YOLO-extract	0.959	0.925	35.591M	32.258

To further verify the effectiveness of YOLO-extract, this paper compares the detection results of the YOLOv5 algorithm with a confidence level of 0.6. As shown in Fig.17.

As shown in Fig.17, the left side is the detection result of YOLOv5 model, and the right side is the detection result





**FIGURE 17.** Comparison of YOLOv5 and YOLO-Extract detection results.

of YOLO-Extract model. Lines 1 to 3 compare the detection effect of dense targets and small targets, the fourth line compares the detection effect of small targets and blocked targets, and the fifth line shows the detection effect of dense targets

and small targets against a complex background. The sixth and seventh lines represent the image blur detection effect caused by weather reasons such as clouds and fog. The yellow and black circles in the figure are the detection results with obvious contrast. The results show that the YOLO-Extract algorithm can achieve better detection accuracy under various conditions.

#### IV. CONCLUSION

For the existing remote sensing image target detection algorithm, the detection accuracy and speed are low, besides, it is easily affected by the background of the images. Based on the characteristics of aircraft targets on optical remote sensing images, this paper optimizes the structure of the model, introduces dilated convolution to improve the feature extraction capability of the model for aircraft targets, and finally optimizes the convergence speed of the loss function plus block model, as well as improves the detection accuracy and detection speed of the model. Experiments show that the method in this paper can greatly improve the ability to overcome the interference of factors such as aircraft attitude and complex background. However, since remote sensing satellite images are easily affected by weather factors such as skylight conditions and clouds and fog, it is difficult to extract different types of aircraft target features in remote sensing images, and there are few data sets for aircraft types, so the detection of aircraft types cannot be completed. In subsequent experiments, relatively clear remote sensing images can be selected to learn and detect different types of aircraft features.

#### REFERENCES

- [1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
- [2] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
- [3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–15.
- [4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2016, pp. 21–37.
- [5] 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)*, Jun. 2016, pp. 779–788.
- [6] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.
- [7] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
- [8] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
- [9] D. Xu and Y. Wu, "Improved YOLO-V3 with DenseNet for multi-scale remote sensing target detection," *Sensors*, vol. 20, no. 15, p. 4276, 2020.
- [10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
- [11] D. Xu and Y. Wu, "FE-YOLO: A feature enhancement network for remote sensing target detection," *Remote Sens.*, vol. 13, no. 7, p. 1311, Mar. 2021.
- [12] S. H. Gao, M. M. Cheng, and K. Zhao, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

- [13] C. Cao, J. Wu, X. Zeng, Z. Feng, T. Wang, X. Yan, Z. Wu, Q. Wu, and Z. Huang, "Research on airplane and ship detection of aerial remote sensing images based on convolutional neural network," *Sensors*, vol. 20, no. 17, p. 4696, Aug. 2020.
- [14] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.
- [15] S. Luo, J. Yu, Y. Xi, and X. Liao, "Aircraft target detection in remote sensing images based on improved YOLOv5," *IEEE Access*, vol. 10, pp. 5184–5192, 2022.
- [16] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.
- [17] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," 2018, *arXiv:1805.10180*.
- [18] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 13713–13722.
- [19] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.
- [20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [21] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12993–13000.
- [22] Y.-F. Zhang, W. Ren, Z. Zhang, Z. Jia, L. Wang, and T. Tan, "Focal and efficient IOU loss for accurate bounding box regression," *Neurocomputing*, vol. 506, pp. 146–157, Sep. 2022.
- [23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
- [24] J. He, S. Erfani, X. Ma, J. Bailey, Y. Chi, and X.-S. Hua, " $\alpha$ -IoU: A family of power intersection over union losses for bounding box regression," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 20230–20242.
- [25] C.-Y. Wang, A. Bochkovskiy, and H.-Y. Mark Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.



**YUAN GAO** received the B.S. degree in software engineering from the Taiyuan University of Technology, Shanxi, China, in 2021. He is currently pursuing the master's degree with the Communication and Network Laboratory, Dalian University. His current research interest includes satellite remote image processing.



**QIANQIAN DU** received the B.S. degree in software engineering from the Taiyuan University of Technology, Shanxi, China, in 2021, where she is currently pursuing the master's degree with the College of Engineering Physics and Optoelectronics. Her current research interest includes satellite remote image processing.



**MENG CHEN** received the B.S. degree in automation specialty from the Nanhang Jincheng College, Nanjing, China, in 2021. He is currently pursuing the master's degree with the Communication and Network Laboratory, Dalian University. His current research interest includes satellite remote image processing.



**ZHIGUO LIU** received the Ph.D. degree from Northeastern University, China, in 2003. Since 2012, he has been a Professor with the Information Engineering College, Dalian University, China. He is currently a Distinguished Professor in Liaoning and enjoys special allowance from the State Council. His research interests include network architecture, protocol technology, and network management.



**WENQIANG LV** received the B.S. degree in software engineering from Inner Mongolia Agricultural University, Hohhot, China, in 2021. He is currently pursuing the master's degree with the Communication and Network Laboratory, Dalian University. His current research interests include routing and wavelength assignment in satellite optical networks.

...