Remote Sensing Image Object Detection Method Based On Improved YOLOv3

Zhiyuan Lu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: luzhiyuan@st.xatu.edu.cn

Bailin Liu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 498194312@qq.com

***Abstract*—In order to solve the problem that irregular targets and dense targets are difficult to be detected in optical remote sensing images, this paper improved the YOLOV3 model Firstly, in order to further combine the feature information of different scales, the PaNet structure is introduced into the FPN part of the original YOLOv3, and the obtained effective feature layer is continued to be extracted for a round of feature. The feature is not only up-sampled to achieve feature fusion, but also down-sampled again to achieve enhanced feature fusion SimOTA method is introduced to dynamically match positive samples and set different positive sample numbers for different targets, which not only improves the speed of the algorithm, but also reduces the extra hyperparameters Experimental verification using richer DOIR data sets shows that the detection ability of the improved algorithm is significantly improved. Compared with the original YOLOv3, its mAP improves by 15.1 points, among which the detection accuracy of dense small targets is improved the most.**

Keywords-Remote Sensing Image; Object Detection; Yolov3; DOIR Dataset

I. INTRODUCTION

Object detection is an important computer vision task that deals with detecting instances of visual objects of a certain class (such as humans, animals, or cars) in digital images. The objective of object detection is to develop computational models and techniques that provide one of the most basic pieces of information needed by computer vision applications: What objects are where? [1]. Up to now, many excellent object detection algorithms have been proposed to classify them according to time. Object detection methods can be divided into traditional object detection method and object detection method based on deep learning. The traditional object detection methods include HOG [2] detector and DPM [3] detector Object detection methods based on deep learning can be divided into two-stage detection methods and single-stage detection methods. Typical representatives of two-stage object detection methods are RCNN [4], Fast RCNN [5] and Faster RCNN [6] and typical representatives of single-stage target detection methods are YOLO [7][8],[9]. SSD [10] and RetinaNet [11] etc.

In the early days when remote sensing technology was born, remote sensing image classification and recognition adopted manual visual and manual marking methods [12]. With the development of deep learning technology, more and more excellent object detection algorithms are applied to remote sensing image object detection. But due to the optical remote sensing image of object and the natural scene in the image difference is very big, mainly embodied in the following aspects: the target scale diversification, different categories of object scale is large, the same category object scales under different shoot height difference is bigger also Image of the upper vertical view shows only the object. Small object with irregular shape object accounts for more than, in the background of more complex scenarios, target and background is easy to confuse Target due to the cause of the shooting Angle in the direction of the arbitrary to these problems such as optical remote sensing image object detection and recognition of enormous challenge [13]. Therefore,

DOI: 10.2478/ijanmc-2022-0031

the problem of remote sensing image object detection is a very challenging topic.

In view of the above difficulties in remote sensing image target detection, an improved algorithm based on YOLOv3 is proposed for remote sensing image object detection. By introducing PaNet [14] method to enhance feature fusion, features of different scales are further fused to deal with the problem of target scale diversity and large number of small objects. SimOTA [15] method was introduced to carry out more precise screening of positive samples with less computation and parameter number. In order to improve the detection accuracy of the original YOLOv3 algorithm detection speed.

II. INTRODUCTION OF YOLOV3 ALGORITHM

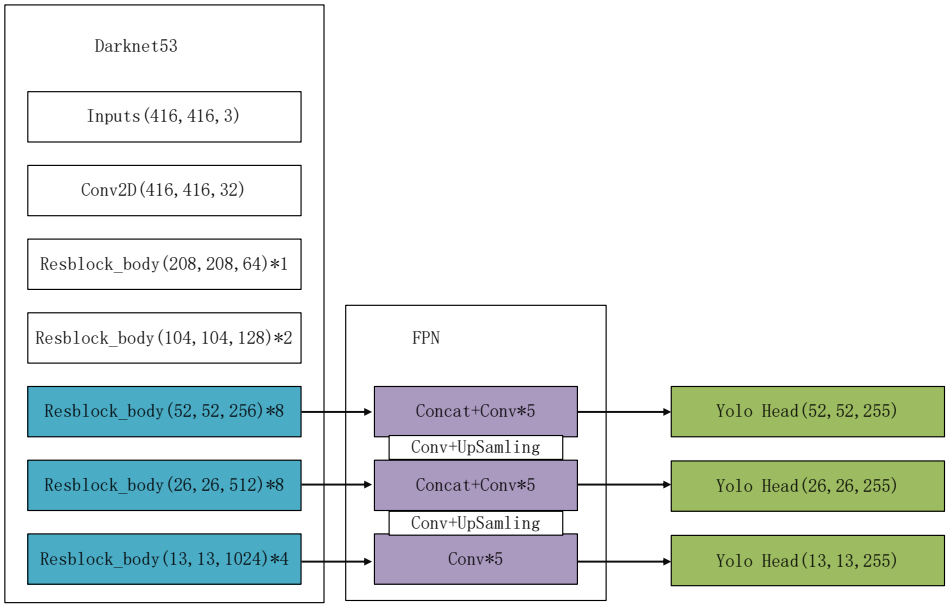

Figure1. YOLOv3 overall structure

YOLOv3, the third version of YOLO algorithm, was first proposed by Joseph Redmon, who creatively treated the object detection problem as a regression problem. Yolov3 network structure as shown in Fig. 1 It first adjust the image size is 416×416, the input to the network, after a series of convolution of image feature extraction, and then through a pyramid fusion of different scale layer characteristics, characteristics of results can be divided into three categories, respectively used to predict the small size targets The goal of medium size and the goal of large size.

A. Backbone Network of YOLOv3

The backbone network of YOLOv3 adopts Darknet-53 network, in which the average pooling layer, full connection layer and Softmax are removed, and is mainly composed of convolution and residual modules. An important feature of Darknet-53 is that Residual network is used. Residual convolution in Darknet-53 is carried out first with a convolution kernel size of 3×3, the convolution with step size 2 will compress the width and height of the input feature layer. At this time, a feature layer can be obtained. After the feature layer is named layer, the 1×1 convolution and 3×3 convolution are carried out for the feature layer, and the result is added to the layer, and then the residual structure is formed Through constant 1×1 convolution and 3×3 convolution and superposition of residual edge, can greatly deepened the characteristics of the residual network is easy to optimize, and to improve accuracy to by adding considerable depth. Its internal residual block uses the jump connection, easing in the depth of neural network to increase the depth of gradient disappeared.

Each convolution part of Darknet-53 uses the unique DarknetConv2D structure. L2 regularization is performed during each

convolution, and Batch Normalization is performed after completion of convolution with Leaky ReLU. The normal ReLU is to set all negative values to zero, Leaky ReLU gives all negative values a non-zero slope which can be mathematically expressed as:

$$y_i = \begin{cases} x_i & x_i \geq 0 \\ \frac{x_i}{a_i} & x_i < 0 \end{cases} \quad (1)$$

B. The features are constructed as predicted results

The process of obtaining prediction results from features can be divided into two parts: building FPN [16] feature pyramid to enhance feature extraction; Three effective feature layers were predicted by YOLO Head.

- FPN feature pyramid was constructed to enhance feature extraction.

In the feature utilization part, YOLOv3 extracts multiple feature layers for target detection. There are altogether three feature layers extracted. Three feature layers are located in different positions of main Darknet-53, namely middle layer, middle and lower layer, and bottom layer. The shapes are (52,52,256), (26,26,512), (13,13,1024). After obtaining three effective feature layers, the three effective feature layers can be used to construct the FPN layer. The construction method is as follows: the feature layer of 13×13×1024 is convolution processed for 5 times. After processing, YOLO Head is used to obtain the prediction results, and part of it is used for up sampling UpSampling2d and 26×26×512 features. The shape of the combined characteristic layer is (26,26,768). The convolution processing is performed again for 5 times in combination with the feature layer. After the processing, YOLO Head is used to obtain the prediction results, and part of up sampling UpSampling2d is used to combine with 52×52×256 feature layer. The shape of the combined feature layer is (52,52,384). Five convolution processes were carried out in combination with the feature layers. After the processing, the prediction result of feature pyramid was obtained by YOLO Head. Feature fusion of feature layers of different shapes was beneficial to extract better features.

- The prediction results were obtained by YOLO Head.

Three enhanced features can be obtained by using FPN feature pyramid, whose shapes are (13,13,512), (26,26,256), (52,52,128), and then the feature layers of these three shapes are passed into YOLO Head to obtain the prediction result YOLO Head is essentially a 3×3 convolution plus a 1x1 convolution, 3×3 convolution is for feature integration, 1×1 convolution is for adjusting the number of channels.

The shape of the output layer is (13,13,75), (26,26,75), and (52,52,75). The last dimension is 75 because the graph is based on VOC dataset. It has 20 classes, and YOLOv3 is for each feature layer There are 3 prior boxes for each feature point, so the number of channels for the prediction result is 3×25. If coco training set is used, there are 80 kinds of classes, and the last dimension should be 255=3×85. Shape of the three feature layers is (13,13,255), (26,26,255), (52,52,255). The actual situation is that N×416×416 images are input, and three shapes of (N,13,13,255), (N,26,26,255) and (N,52,52,255) will be output after multi-layer operation, corresponding to each figure is divided into 13×13, 26×26 and 52×52 grid Position of 3 prior boxes on.

C. Decoding of prediction results

According to the second step, the prediction results of three feature layers can be obtained, and shapes are: (N,13,13,255), (N,26,26,255), (N,52,52,255), Each effective feature layer divides the whole image into grids corresponding to its length and width. For example, the feature layer (N,13,13,255) divides the whole image into 13×13 grids. Then several prior boxes are built from each grid center, these boxes are pre-configured boxes for the network, the network's predictions determine whether these boxes contain objects and what kind of object it is. Since each grid point has three prior boxes, the prediction result above can be reshape into: (N,13,13,85), (N,26,26,3,85), (N,52,52,3,85). Where 85 can be split into 4+1+80,

where 4 represents x_offset, y_offset, h and w the four, the parameters of the prior box 1 represents a priori whether the box contains objects, 80 represent the types of the a priori box, because coco got the 80 class, so there is 80 but this result does not correspond to the final forecasting box in the picture, you also need to decode can complete YOLOv3 decoding process is divided into two steps:

- Add x_offset and y_offset to each grid point, and the result is the center of the prediction box.
- Then, the width and height of the prediction frame can be calculated by combining the prior frame and h and w, so that the position of the whole prediction frame can be obtained. After obtaining the final prediction results, score sorting and non - maximum suppression screening should be carried out.

III. IMPROVED YOLOV3

A. Strengthening feature fusion

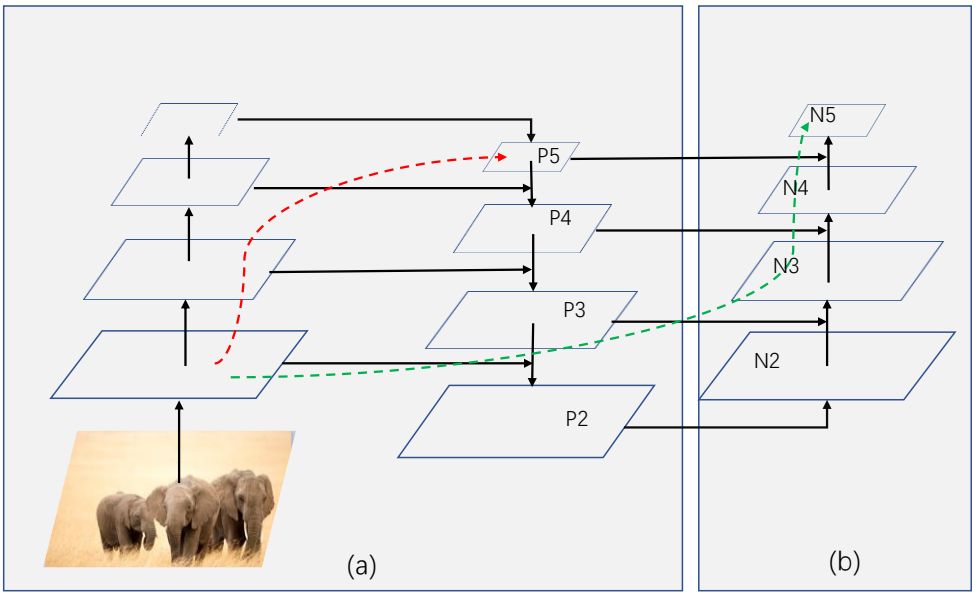

Figure2. PaNet overall structure

PaNet(Path Aggregation Network) is the addition of a bottom-up Path Augmentation to FPN. You can see in Figure 2 (a) the FPN is top-down route, through the lateral connection, transfer the strong semantic characteristics of the top down, only to enhance the characteristics of the pyramid semantic information For example, when the underlying characteristics to the P5 (red line), after very multilayer networks among them, at this point at the bottom of the target information is very fuzzy, so the FPN extensions, added a bottom-up route (green route, bottom ->P2-> N2~N5, where the path passes through less than 10 layers), as shown in Fig. 2 (b), thus compensating and reinforcing the positioning information.

By applying PaNet to YOLOv3, another round of down sampling is conducted on the basis of its FPN structure to Further strengthen feature fusion, as shown in Fig. 3, thus improving the accuracy of YOLOV3 for detecting dense small targets and multi-scale targets in remote sensing images.

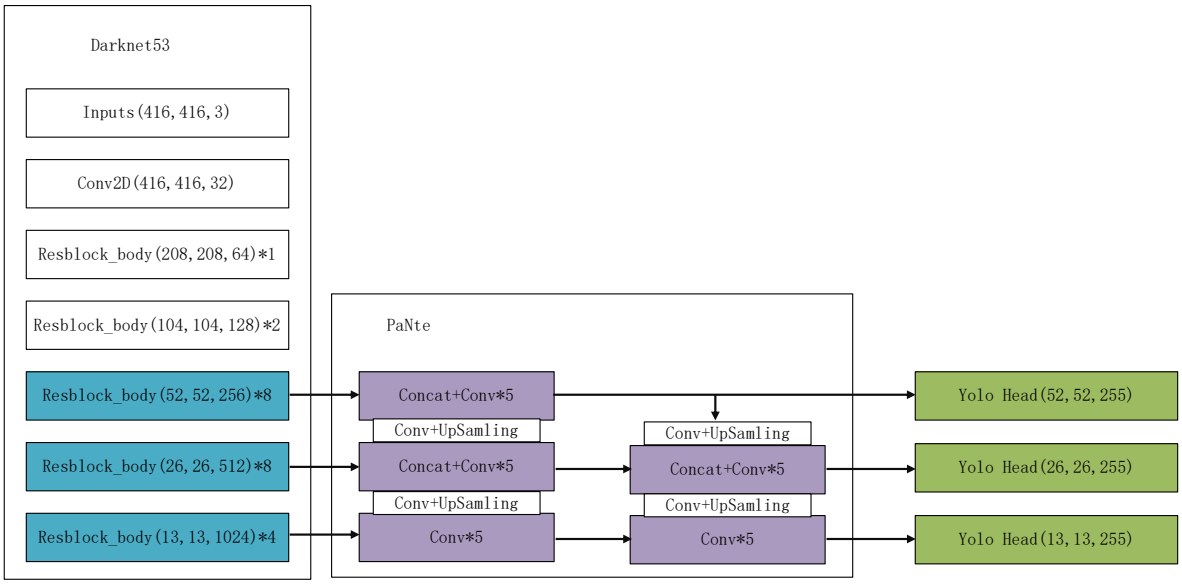

Figure3. YOLOv3 structure after introduction of PaNet

B. SimOTA method

SimOTA's function is to set different positive sample numbers for different targets, such as ants and watermelon. Traditional positive sample allocation schemes usually assign the same positive sample numbers to watermelon and ants in the same scene, so either ants have many low-quality positive samples, or watermelon only has one or two positive samples. It's not appropriate for either way of distribution. Therefore, SimOTA firstly calculates a cost matrix, which represents the cost relationship between each real box and each feature point. The cost matrix consists of three parts:

- The higher the degree of overlap between each real box and the prediction box of the current feature point, it means that this feature point has tried to fit the real box, so its Cost will be smaller.
- The higher the type prediction accuracy of each real box and the current feature point prediction box is, it also means that this feature point has tried to fit the real box, so its Cost will be smaller.
- Does the center of each real frame fall within a certain radius of the feature point? If the center of each real frame falls within a certain radius of the feature point, it means that the feature point should fit the real frame, so its Cost will be smaller.

The process application of SimOTA is as follows:

- Calculate the coincidence degree between each real box and the current feature point prediction box.
- Calculate the IOU of the ten prediction frames with the highest coincidence degree and the real frame to get the K of each real frame, which means that each real frame has K feature points corresponding to it.
- Calculate the type prediction accuracy of each real box and the current feature point prediction box.
- Judge whether the center of the real box falls within a certain radius of the feature point.
- Compute the cost matrix.
- The k points with the lowest cost are taken as positive samples of the real box.

IV. EXPERIMENTAL RESULTS ANS ANALYSIS

A. Experiment preparation

All experiments were carried out on a Linux operating system computer equipped with AMD

R9-5900HX CPU, Nvidia 2080Ti GPU, 16G video memory and 16G memory. The remote sensing image dataset used in the experiment is DOIR dataset, which contains 23463 remote sensing images and 192472 object instances. These instances are manually marked as boundary boxes of axial pairs, covering 20 common object categories. The size of data set images is 800×800 pixels, and the spatial resolution is 0.5m~30m Among them, the training set accounted for 1/2, the verification set accounted for 1/6, and the test set accounted for 1/3.

B. Evaluation indicator

The evaluation indicators of the experiment include precision(P), recall (R), average precision (AP), and mean Average Precision of all categories(mAP). The calculation formula of each index is as follows:

$$\mathrm{R} = \frac{TP}{TP + FN} \tag{2}$$

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$AP = \int_0^1 P_{smooth}(r) d_r \tag{4}$$

$$mAP = \frac{\sum_{j=1}^{K} AP_i}{K} \tag{5}$$

Where TP, FP, TN and FN represent positive samples with correct predictions and positive samples but wrong predictions and negative samples with correct predictions and negative samples with wrong predictions. $P_{smooth}(r)$ is a smooth P-R curve.

C. Comparison of experiment results

The comparison results of YOLOv3 and the improved YOLOv3 in this paper are shown in TABLEI. It can be seen that the detection accuracy of the improved YOLOv3 for small objects has been significantly improved, and the mAP has increased by 15.1 points. In terms of single category, the improved method is applied to the Wind mill, Airplane, Tennis Court, Expressway service area, Baseball fields and other types of small object detection performed better. The best-performing ballpark had an accuracy of 98.2 percent.

TABLE I. DETECTION ACCURACY OF EACH CATEGORY ON DOIR DATASET

c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
Airplane	Airport	Baseball field	Basketball court	Bridge	Chimney	Dam	Expressway service area	Expressway toll station	Golf course
c11	c12	c13	c14	c15	c16	c17	c18	c19	c20
Ground track field	Harbor	Overpass	Ship	Stadium	Storage tank	Tennis court	Train station	Vechile	Wind mill

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11	c12	c13	c14	c15	c16	c17	c18	c19	c20	mAP
YOLOv3	90.9	69.5	81.7	78.6	61.2	69.7	66.9	88.6	74.4	61.1	89.1	44.9	49.7	90.4	70.6	68.7	87.3	59.4	68.3	78.7	72.5
改进 YOLOv3	96.7	95.1	98.2	85.6	71.7	95.4	79	96.1	91.3	90.1	92.4	68.2	74.2	92.5	90.8	80.5	96.7	77	84.9	96	87.6

P-R curves of each category are shown in Fig. 4, and the detection effect of the algorithm in this paper on small targets in remote sensing images is shown in Fig. 5, from which it can be seen that the improved YOLOv3 has a better detection effect on small targets.

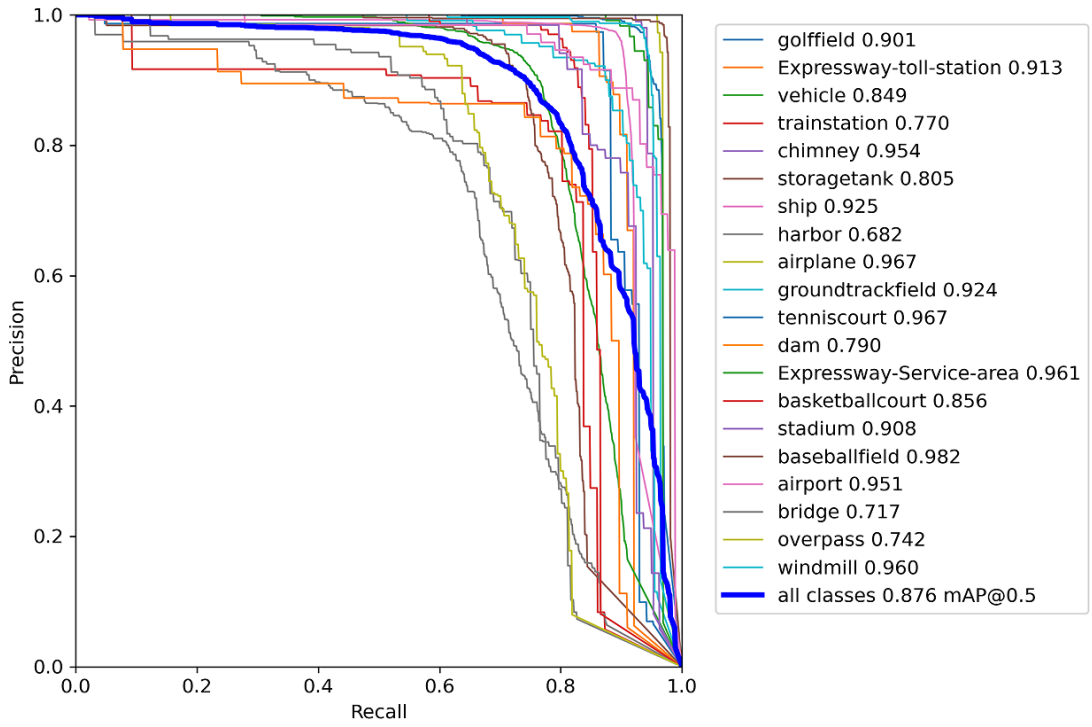

Figure4. P_R curve

(a) (b)

Figure5. The detection effect of improved YOLOv3

V. CONCLUSION

Based on remote sensing image inherent in the small object proportion is high, the object scale inconsistent problems, puts forward the improved YOLOv3 algorithm used for target detection in remote sensing image, the original structure of the characteristics of the pyramid module on the basis of round of sampling was conducted again, to strengthen the feature extraction and improved its ability to cope with multi-scale object In general, the improved YOLOv3 algorithm solves the problem of low detection accuracy of partially

dense small targets in previous algorithms, and improves mAP by 15.1 points.

REFERENCES

[1] Zou Z, Shi Z, Guo Y, et al. Object detection in 20 years: A survey[J]. arXiv preprint arXiv:1905.05055, 2019.

[2] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). Ieee, 2005, 1: 886-893.

[3] Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model[C]//2008 IEEE conference on computer vision and pattern recognition. Ieee, 2008: 1-8.

[4] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[5] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[6] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.

[7] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[8] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

[9] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[10] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.

[11] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[12] Tuia D, Persello C, Bruzzone L. Domain adaptation for the classification of remote sensing data: An overview of recent advances[J]. IEEE geoscience and remote sensing magazine, 2016, 4(2): 41-57.

[13] Li K, Wan G, Cheng G, et al. Object detection in optical remote sensing images: A survey and a new benchmark[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 159: 296-307.

[14] Liu S, Qi L, Qin H, et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8759-8768.

[15] Ge Z, Liu S, Li Z, et al. Ota: Optimal transport assignment for object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 303-312.

[16] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.