Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.049710

ARTICLE

Probability-Enhanced Anchor-Free Detector for Remote-Sensing Object Detection

Chengcheng Fan[1,2,*] **and Zhiruo Fang**[3]

[1]Innovation Academy for Microsatellites of CAS, Shanghai, 201210, China

[2]Shanghai Engineering Center for Microsatellites, Shanghai, 201210, China

[3]College of Artificial Intelligence, Nanjing Agricultural University, Nanjing, 210095, China

*Corresponding Author: Chengcheng Fan. Email: fancc@microsate.com

Received: 16 January 2024 Accepted: 29 April 2024 Published: 20 June 2024

ABSTRACT

Anchor-free object-detection methods achieve a significant advancement in field of computer vision, particularly in the realm of real-time inferences. However, in remote sensing object detection, anchor-free methods often lack of capability in separating the foreground and background. This paper proposes an anchor-free method named probability-enhanced anchor-free detector (ProEnDet) for remote sensing object detection. First, a weighted bidirectional feature pyramid is used for feature extraction. Second, we introduce probability enhancement to strengthen the classification of the object's foreground and background. The detector uses the logarithm likelihood as the final score to improve the classification of the foreground and background of the object. ProEnDet is verified using the DIOR and NWPU-VHR-10 datasets. The experiment achieved mean average precisions of 61.4 and 69.0 on the DIOR dataset and NWPU-VHR-10 dataset, respectively. ProEnDet achieves a speed of 32.4 FPS on the DIOR dataset, which satisfies the real-time requirements for remote-sensing object detection.

KEYWORDS

Object detection; anchor-free detector; probabilistic; fully convolutional neural network; remote sensing

1 Introduction

Remote-sensing object detection has a wide range of applications in environmental monitoring, urban planning, agriculture, and other fields. It encompasses both two-stage and one-stage approaches. Two-stage detection approaches include Region-based Convolutional Neural Networks (RCNNs) [1–4]. In the Faster-RCNN [3], the region proposal network (RPN) is used to generate prior boxes that are crucial for object detection. An excessive number of preselected boxes reduces the speed and efficiency of the detector. Faster-RCNN needs to use a selective search algorithm to determine the candidate region, which is very time consuming. Not every candidate box contains the corresponding feature, causing unnecessary computational overhead for the model. Conversely, in one-stage object detection approaches, both the probability of the target classes and location are predicted simultaneously. The YOLO series [5–9] transforms the detection problem into a regression problem. The one-stage object detection approaches generally have lower accuracy than the two-stage approaches. Specifically, the

one-stage approaches encounter certain difficulties in accurately regressing the foreground, which can affect the overall detection performance.

Anchor-based approaches, which are rooted in the employment of predefined anchor boxes, have been widely applied in remote-sensing object detection. The anchor-based approaches generate dense anchor boxes, enabling a network to directly classify targets and regress bounding boxes. However, anchor-based approaches need to carefully set numerous hyperparameters, such as scale and aspect ratio, which significantly impact the detection performance. In addition, the anchor-based approaches generate many redundant boxes, thereby increasing the computation and memory consumption. Therefore, in practical scenarios where detection efficiency is required, anchor-free approaches are more advantageous. Anchor-free method, for example, CornerNet [10] characterizes the bounding box by predicting two key points located at the upper left and lower right corners without requiring presenting anchor boxes. ExtremeNet [11] identifies four extreme points, namely topmost, bottommost, leftmost, and rightmost points of the object's boundary. This approach offers a robust and accurate means of detecting objects. CenterNet [12] represents an object bounding box by predicting the central point of the object and its corresponding distance to the boundaries of the box. FCOS [13] applies a fully convolutional network (FCN) to object detection, enabling the pixel-wise prediction of center point and its distances to the bordering pixels.

However, anchor-free based approach often misinterprets the background or noise of a target, leading to false detection. In CenterNet [12], the backbone adopts hourglass [14] and Deep Layer Aggregation (DLA) [15] networks. The main reason for using these backbone networks is that they have powerful feature extraction capabilities, which can effectively discriminant the background and noise. CenterNet2 [16] is a further improvement and optimization of CenterNet. It proposes a probabilistic Two-Stage detector that combines the advantages of both One-Stage and Two-Stage detectors. In the Two-Stage detector, the first stage infers the object-background likelihood, and the second stage obtains the specific classification score. The CenterNet2 model extracts region-level features and performs classification, and the two stages are trained together to maximize the probabilistic accuracy of the prediction. It enables CenterNet2 to achieve further performance improvements on the COCO dataset.

In remote sensing image object detection, the speed of training and the real-time inference are crucial. TTF [17] uses elliptical Gaussian kernel instead of the circular Gaussian kernel in CenterNet to make the training process more reasonable. Anchor-free design helps to reduce the amount of computation while improving the detection accuracy. For example, FCOS [13] detects objects with fractions that have four distances and centres. The foreground and background are classified according to their positions on a fused feature map. ATSS [18] further improves FCOS by changing the definitions of the foreground and background. To optimize the inference speed, some improved methods try to use a single-level detection architecture to reduce the computational burden caused by the multi-level detection architecture. The AF-EMS detector [19] introduces a modified CenterNet model with a splicing strategy to achieve better accuracy. A simple mathematical [20] anchor-free method detects arbitrary direction objects which contain small objects.

Remote sensing image often encompasses different categories, which often intricate overlaps and interdependencies. For example, the DIOR dataset [21] has categories such as 'harbour,' 'ship', 'vehicle,' and 'overpass' (Fig. 1). In this dataset, ship and harbour are different but related target categories, and they have positional, functional, and dependent relationships between them. In an image, a ship is often docked in a harbour with visually adjacent and overlapping sections that are challenging to detect. As small objects in DIOR, vehicles often appear together with overpasses, bridges, and other

objects. Challenges are encountered in separation between a vehicle's front and background. Remote-sensing object detection encounters various challenges owing to the intricacies of natural backgrounds, diverse interclass characteristics with inherent similarities, and a wide range of object scales. Accurately interpreting the probabilities of foreground and background regions can significantly strengthen the classification of the object's foreground and background. This paper proposes an anchor-free detector called probability-enhanced anchor-free detector (ProEnDet) for remote-sensing object detection. An FCN is introduced and combined with a bidirectional feature pyramid as the neck to generate a fusion feature map. To bolster the discrimination between the foreground and background in remote-sensing object detection, we employed probabilistic enhancement techniques to refine the precision of the anchor-free detector. ProEnDet takes an anchor-free detection approach having the speed of the one-stage method and achieving the accuracy of the two-stage detector through probabilistic reinforced classification. Considering the aforementioned challenges, this paper presents an anchor-free detector for remote-sensing images. The contributions of this paper are as follows: (1) This paper proposes an anchor-free detector which utilizes an FCN to improve the distinction between the foreground and background using remote-sensing images with complex scenes. (2) The neck of the architecture adopts a repeated weighted bidirectional feature pyramid network (BiFPN), which upgrades the detector to multiple scales. (3) A probabilistic enhancing component is used to augment the prediction probability. A likelihood score is introduced to improve the detection accuracy of prospects at different feature levels.

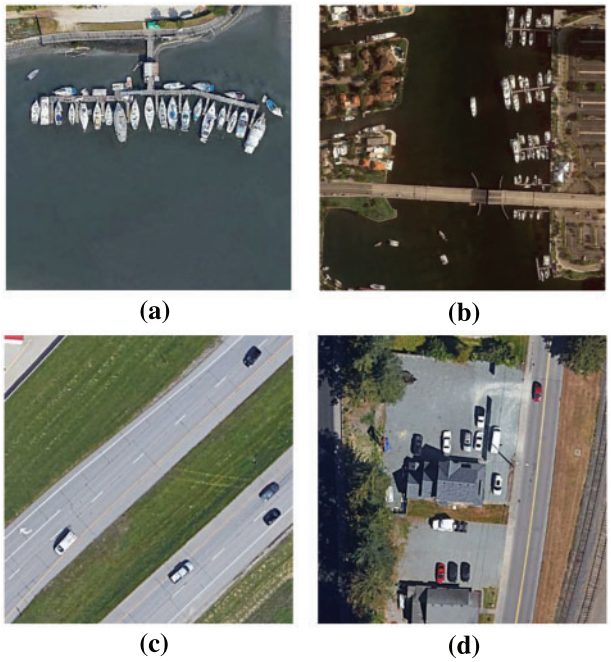

Figure 1: Images in the DIOR dataset: (a) and (b) 'ship' and 'harbour', (c) and (d) 'vehicle' and 'overpass'

2 Probability-Enhanced Anchor-Free Detector

2.1 Over-All Architecture

ProEnDet consists of a backbone, detection neck, detection heads, and probabilistically enhanced components. An overview of the architecture is shown in Fig. 2.

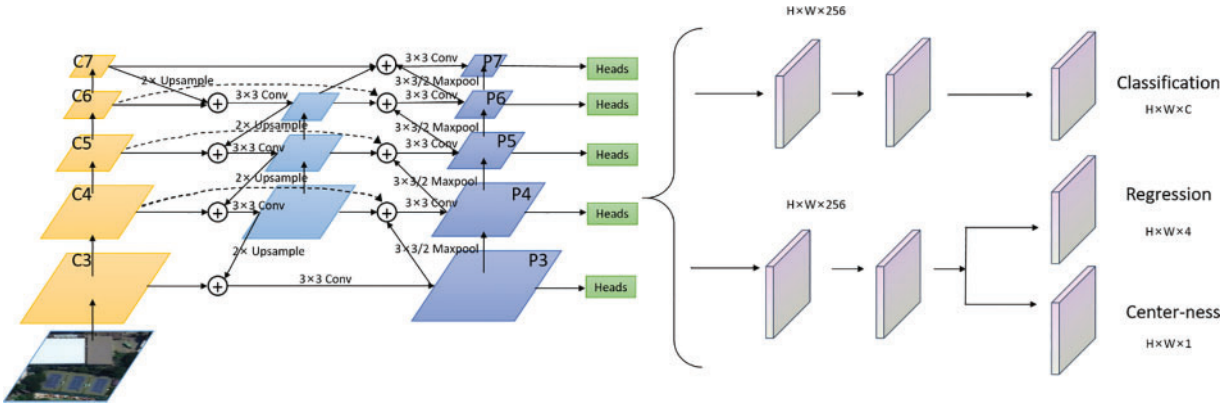

Figure 2: Framework of ProEnDet

The primary objective of ProEnDet is to attain comparable performance in remote-sensing tasks based on the anchor-free object detection method. As the backbone, C3–C5 denote the feature maps. A repeated-weighted bidirectional feature pyramid network is employed as the neck, in which different weights are assigned to each layer for fusion, thereby enabling the network to prioritise attention towards crucial layers. A traditional FPN is achieved by integrating the features of different levels when addressing targets of different scales. Based on this, the BiFPN introduces a weighted fusion of features across distinct levels, enabling the network to handle objects of varying scales more effectively. A detection head comprises three distinct branches: Classification, regression, and center-ness. The regression and center-ness diverged into two separate branches from the same initial branch. Each branch undergoes a sequence of four Conv2d, GN, and ReLU modules before passing through a convolutional layer with a kernel size of 3×3 and step size of 1, ultimately yielding the final prediction outcome. Between the feature extraction and prediction heads, we introduce probability enhancement to strengthen the classification of the object foreground and background, and the class score is added to the final prediction score. ProEnDet is an anchor-free detection approach that enhances the performance of object detection in remote-sensing scenarios by integrating a robust backbone and probabilistic augmentation. Additionally, the model achieves lightweight and real-time performance.

2.2 Repeated Weighted Bidirectional Feature Pyramid Network

In the network architecture of ProEnDet, C3–C5 denote feature maps. Feature levels P3–P7 are used for the final prediction. The neck of the architecture adopts a repeated weighted BiFPN [22]. The neck uses three to seven feature levels {P3, P4, P5, P6, and P7}. At the neck, a bidirectional feature fusion approach encompassing both top-down and bottom-up flows is implemented, facilitating the seamless integration of diverse feature representations. The pyramid treats each top-down and bottom-up pathway as a distinct layer in the feature network and achieves higher-level feature fusion by repeating the same layer multiple times. All computations are performed with an input size of 800×1024 as an exemplary case. Repeated blocks with better accuracy and efficiency trade-offs are

achieved by fusing the feature maps sampled layer-by-layer. By introducing horizontal and vertical connections, the features of different scales can be better fused and utilised.

The pyramid network enables the detector to detect targets of various scales. Four-layer classification and regression branches are applied at all scale levels to generate heat maps and bounding box regression maps. The detection head comprises three distinct branches. The classification branch facilitates the prediction of the object categories, the regression branch localises the positional coordinates, and the center-ness branch assesses the centrality of the anchor relative to the detection box. Within the detection head, both the classification and regression branches are enhanced through a sequence of four convolutional layers.

In ProEnDet, the backbone and neck are treated as stages to produce class-independent scores. The final detection result combines the detector output with a probability enhancement score. Class and box networks process the fused features, and a class score is converted into the final prediction score, which is enhanced with probability. The likelihood score of the foreground is added to enhance the foreground and background separation effects of the detector. The proposed method uses a centre-based design, which improves performance.

2.3 Probabilistic Enhancement

In ProEnDet, the regression box obtains a score that determines whether an object belongs to either the foreground or background. The probabilities of the jointly inferred locations and categories in a frame are obtained through the backbone and neck in the first stage. ProEnDet extracts and classifies the region-level features. The output of the feature levels predicts a reliable target probability for each proposed box using an optimised anchor-free box detector. ProEnDet makes the object a keypoint located at its centre and then returns to the box argument. A probabilistic two-stage enhancing component is used for the prediction probability augmentation. The traditional two-stage algorithm only needs to judge whether the candidate box is foreground or background in the first stage, while the proposed method not only needs to judge whether the candidate box is background or foreground in the first stage, but also gives the corresponding possibilities, that is, to score these possibilities. By using a one-stage network, we reduce the number and quality of candidate boxes generated in the first stage, which reduces the amount of calculation for regression prediction in the second-stage model.

The backbone network passes the features to the head to obtain the class probability score, prediction and class-independent confidence prediction of each layer feature map, respectively. If only use the network as the first stage to generate proposals, we only have class-independent confidence. The one-stage detector alone then has no category independent confidence. If it is the first stage of the two-stage probability detection, it includes the positive and negative sample loss of confidence.

As shown in Fig. 3, the class-agnostic $P(O_k)$ used in the probabilistic component adds a class score to the final prediction score. Anchor-free detectors, combined with probabilistic augmentation components, require strong backbones. Stages P3–P7 are used as primary and secondary detectors. In the proposed method, the anchor-free detector enhances the prediction to ensure an accurate object likelihood for probabilistic augmentation. The detector must predict the exact object likelihood to obtain an overall detection score. Subsequently, it combines them with the inference. These likelihoods are combined with the classification scores to produce a probability score for the final detection. At inference, the score of the second stage is multiplied with the score of the first stage. The head process the proposals, adding categories and regression boxes. After pooling, it goes to the first head for classification and regression, and then sends proposals to the next head to repeat until all heads

have gone through the above steps. The predicted category score for each head is then used to interact with the probability score.

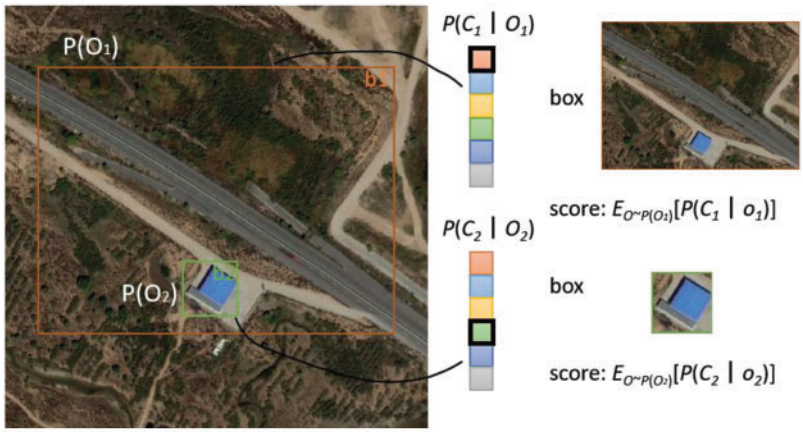

Figure 3: Illustration of probability scores

In ProEnDet, the detected object produces K bounding boxes. The bounding boxes follow the class distribution as $b_1, b_2, \ldots, b_k$. The probabilistic component in the proposed methodology emphasises class distribution while maintaining an unaltered bounding box regression. For the detector, $O_k = 1$ represents the foreground and $O_k = 0$ represents the background in the first part. $\mathrm{P}(C_k|O_k)$ represents conditional categorical classification. Thus, the joint class distribution of the anchor-free detection model is $\mathrm{P}(C_k) = \sum_o \mathrm{P}(C_k|O_k = o)\,\mathrm{P}(O_k = o)$.

For the maximized target classification and background classification, respectively,

$$\log \mathrm{P}(C_k) = \log \mathrm{P}(C_k|O_k = 1) + \log \mathrm{P}(O_k = 1) \tag{1}$$

$$\log \mathrm{P}(bg) = \log \mathrm{P}(\mathrm{bg}|O_k = 1)\,\mathrm{P}(O_k = 1) + \log \mathrm{P}(O_k = 0) \tag{2}$$

Probability estimation is combined with the gradient calculation to obtain an accurate evaluation score of the predicted class. Numerous evaluation metrics are generated here, which slows down the training; however, subsequently, a joint optimisation of the two lower bounds is used.

$$\log(\alpha x_1 + (1-\alpha)\,x_2) \geq \alpha \log(x_1) + (1-\alpha)\log(x_2) \tag{3}$$

From Jensen's inequality, in the inequality, $\alpha = \mathrm{P}(O_k = 1)$, $x_1 = \mathrm{P}(\mathrm{bg}|O_k = 1)$, $x_2 = 1$.

Training has two lower bounds to optimise speed. The first lower bound maximises the background log-likelihood of high-scoring objects during training. In the inequality, $\mathrm{P}(O_k = 1) \to 0$ and $\log \mathrm{P}(\mathrm{bg}|O_k = 1) \to 1$.

$$\log \mathrm{P}(bg) \geq \mathrm{P}(O_k = 1) \log \mathrm{P}(\mathrm{bg}|O_k = 1) \tag{4}$$

In the first stage, $\mathrm{P}(\mathrm{bg}|O_k = 1)\,\mathrm{P}(O_k = 1) > 0$. The log function is monotonic. Therefore, in practice, the joint optimisation of both lower bounds leads to better results.

$$\log \mathrm{P}(bg) \geq \log(\mathrm{P}(O_k = 0)) \tag{5}$$

With the joint optimisation of the two lower bounds, the RPN in two-stage detection is replaced by positive labels on the detected objects and negative labels at other locations.

ProEnDet extracts and classifies the region-level features. In addition, ProEnDet focuses on the class distribution. In the probabilistic form, the classification score is multiplied by the class-agnostic detection score. Using an effective one-stage detector, reliable target probability can be predicted for each proposed box. This enables the detector to better distinguish between the foreground and background. This enhances the precision of remote-sensing object detection in complex environments and diverse backgrounds. Additionally, joint optimisation does not cause redundant computations.

2.4 Loss Function and Other Details

Three types of head detection methods exist: Classification, regression, and center-ness. Anchor-free detection regress ranges with bounding boxes at each level. Heads are shared at different levels. Regression and center-ness are two components of the same branch. Each branch first passes through a composing module, which includes four Conv2d, a GN, and a ReLU. The input is passed through a convolutional layer with a kernel size of 3×3 and a stride of 1 to obtain the final prediction result. For the classification branch, 80 score parameters and an additional class score for probabilistic augmentation component inference are anticipated at each position of the predicted feature map. Within the regression branch, the prediction feature map at each position undergoes the prediction of four distance parameters, enabling precise localization of objects, representing the upward, downward, left, and right positions. For the center-ness branch, a parameter is predicted at each position on the prediction map. Center-ness reflects the closeness of the point to the target centre, and it ranges from 0 to 1. In the post-processing part of the network, high-quality anchor boxes are screened.

$$\mathrm{L}\left(\{p_{x,y}\},\{t_{x,y}\}\right)=\frac{1}{N_{pos}}\sum_{x,y}L_{cls}\left(p_{x,y},c^{*}_{x,y}\right)+\frac{\lambda}{N_{pos}}\sum_{x,y}1_{c^{*}_{x,y}>0}L_{reg}\left(t_{x,y},t^{*}_{x,y}\right) \tag{6}$$

The loss function of ProEnDet is given by Eq. (6). L_{cls} represents the focal loss, L_{reg} represents the intersection-over-union (IoU) loss, N_{pos} represents the number of positive samples. λ represents balancing weights of L_{reg}. $p_{x,y}$ represents the score for each class predicted at point (x, y) of the feature map. $c^{*}_{x,y}$ represents the true class label at point (x, y) in the feature map. $1_{c^{*}_{x,y}>0}$ if point (x, y) of the feature map is matched as a positive sample, and 0 otherwise. $t_{x,y}$ represents the predicted object bounding box information at the point, and $t^{*}_{x,y}$ represents the corresponding real bounding box information.

For the classification branch, at each position on the predicted feature map, score parameters are estimated. Concurrently, within the regression branch, four distance parameters are predicted for each location on the prediction feature map, facilitating accurate localization of objects. If, for a point on the predicted feature map, the coordinates mapped onto the original image are (c_x, c_y), and the step distance of the feature map from the original image is s, then the network estimates the coordinates of the bounding box using points using (7)–(10), respectively.

$$x_{min}=c_x-l\cdot s \tag{7}$$

$$y_{min}=c_y-t\cdot s \tag{8}$$

$$x_{max}=c_x+r\cdot s \tag{9}$$

$$y_{max}=c_y+b\cdot s \tag{10}$$

For the center-ness branch, one parameter is predicted at each position of the prediction feature map, which reflects the closeness of the point on the feature map to the target centre, and its range is between 0 and 1. The closer the point is to the target centre, the closer the *centerness* is to 1. Positive

samples are considered for the calculations.

$$centerness^{*} = \sqrt{\frac{\min\left(l^{*}, r^{*}\right)}{\max\left(l^{*}, r^{*}\right)} \times \frac{\min\left(t^{*}, b^{*}\right)}{\max\left(t^{*}, b^{*}\right)}} \tag{11}$$

When screening high-quality bounding boxes in the post-processing stage of the network, the bounding boxes are ranked based on the acquired results, and those with higher scores are retained. The bounding boxes with low target class scores and the predicted point are far from the target centre and are filtered out, and the high-quality bounding boxes are retained.

Based on the results, the anchor boxes with higher scores are retained when anchor boxes are sorted. Boxes with low target class scores and prediction points far from the target centre are filtered out, and high-quality anchor boxes are retained. The probability enhancement score provides a certain enhancement in the prediction, which improves the accuracy of the prospect on different feature levels.

3 Experiment

3.1 Datasets

To validate the efficacy of our model, we conduct extensive experiments on benchmark dataset including The DIOR dataset [21] and The NWPU-VHR-10 dataset [23]. The DIOR dataset [21] is collected from Google Earth satellite imagery by Google. It comprises 23,463 high-quality remote-sensing images encompassing 20 common object categories. The dataset contains 192,472 labelled object instances. The DIOR dataset is notable for its extensive size and diverse categories. DIOR images in the same category contain rich dimensional variations. The features of the DIOR dataset include a large image size, rich information in each image, in-class similarity, and out-of-class similarity with categories. In this study, the training and validation sets of the DIOR used in the experiment were set to 2:1. The NWPU-VHR-10 dataset [23] is collected from Google Earth satellite imagery and labelled by the Northwestern Polytechnical University (NWPU). The dataset contains a total of 3775 instances and ten classes. It has overlapping categories with the DIOR dataset. It contains 390 baseball diamonds, 757 airplanes, 124 bridges, 477 vehicles, 159 basketball courts, 302 ships, 224 harbours, 163 ground track fields, 655 storage tanks, and 524 tennis courts. As shown in Fig. 4, the dataset has positive and negative images. The NWPU-VHR-10 dataset has overlapping categories with DIOR, but it has smaller picture sizes and fewer quantities. The ratio of the training, test, and validation sets used in the experiment is 7:2:1. We used this dataset validation method for the secondary detector.

3.2 Experimental Details

For this investigation, the Ubuntu 16.04 LTS system served as the experimental platform. All model training were conducted on NVIDIA 1080Ti GPUs, equipped with 32 GB of RAM, and executed using Python 3.8. During the training process, parameter scheduling adhered to the guidelines established by Detectron2. Notably, all experimental outcomes were derived without any alterations or modifications. Figs. 5a–5l show visualisations of the test results on the DIOR dataset.

3.3 Evaluation Metrics

In the domain of object detection, the performance of the detector is assessed using the average precision (AP) for each category and the mean average precision (mAP) across all categories. The recall reflects the model's capability to detect an object, whereas the precision indicates the model's accuracy. The greater the precision, the more precise the model in predicting the object. The formulae

for recall and precision are as follows: Where TP, FP, FN, and TN denote true positive, false positive, false negative, and true negative, respectively.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{12}$$

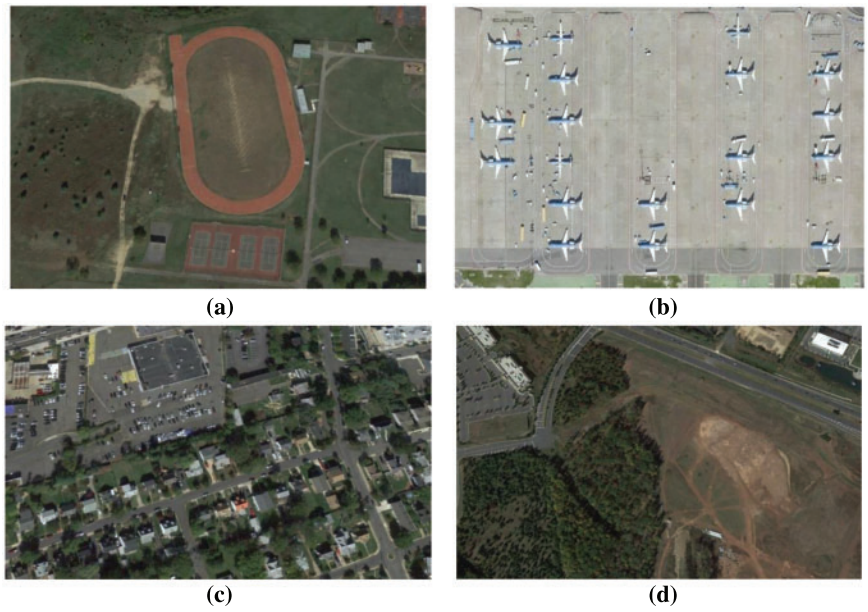

Figure 4: Positive and negative images in the NWPU-VHR-10 dataset

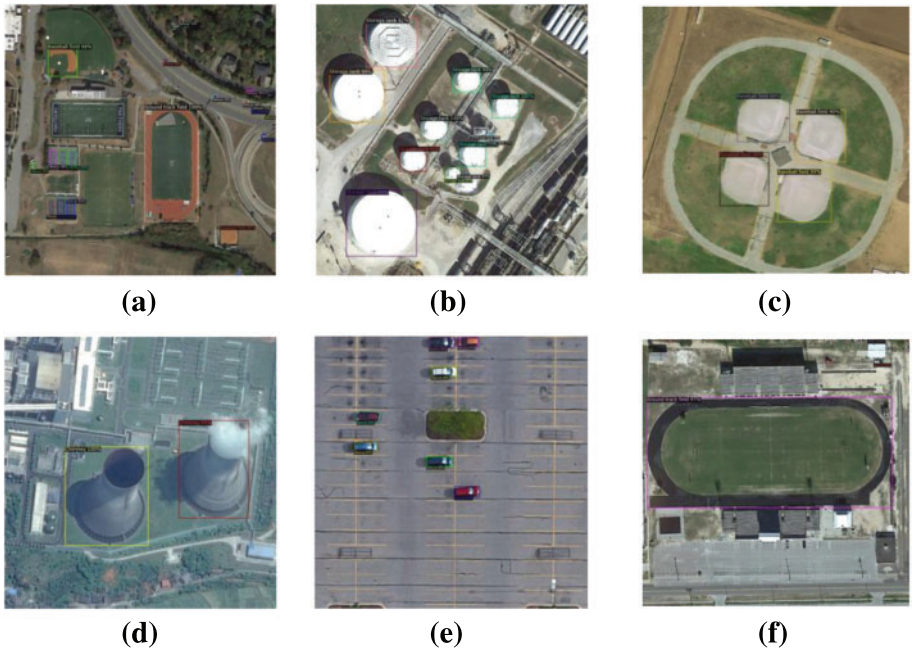

Figure 5: (Continued)

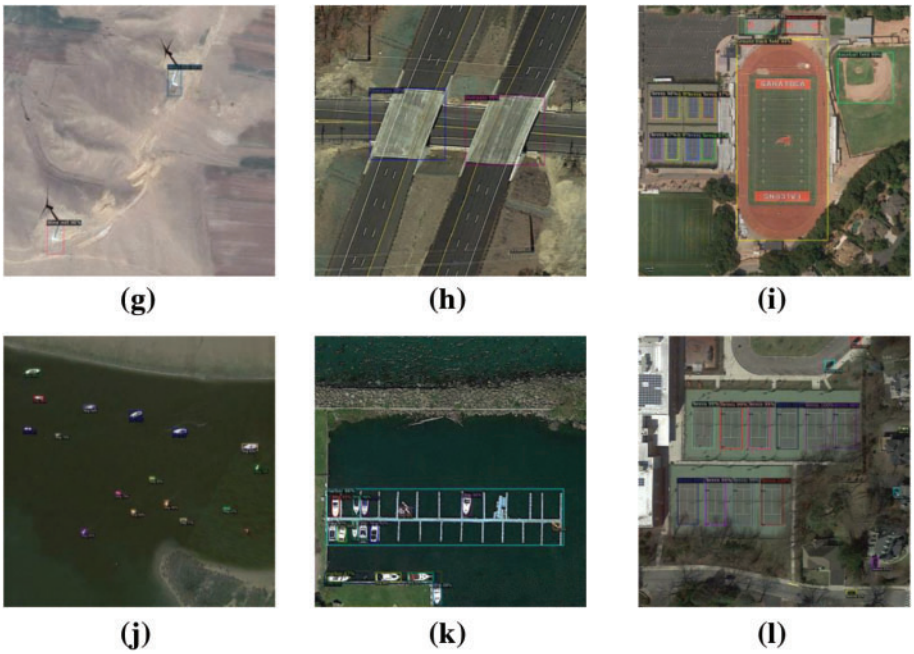

Figure 5: Visualization results of ProEnDet detection on the DIOR dataset

The precision of objects across various scales, specifically APS, APM, and APL, respectively denote the accuracy achieved for small, medium, and large-sized objects. The classification of object sizes is determined by their areas, with small objects measuring less than 322, medium-sized objects ranging from 322 to 962, and large objects exceeding 962. To assess the average precision under different Intersection over Union (IoU) thresholds, metrics such as AP50 and AP75 are employed, utilizing both the PASCAL VOC standard and a more stringent evaluation criterion.

$$AP = \int_0^1 P(R)\, dR \tag{13}$$

$$mAP = \frac{1}{N_{cls}} \sum_{i=1}^{N_{cls}} AP_i \tag{14}$$

Frames per second (FPS) is the metric to quantify the speed of each model. The parameter count represented the aggregate number of parameters that underwent training during the model training process. Floating-point operations (FLOPs), which serve as a measure of algorithm complexity, were employed as an indirect indicator of neural network model speed. In this study, the comprehensive evaluation encompassed metrics such as FPS, and FLOPs, Average Precision (AP), AP50, AP75, mAP, APS, APM, APL and model size.

3.4 Comparison Experiment

In this section, we compare the proposed detector with anchor-based and anchor-free methods, including anchor-free, two-stage, and one-stage methods. The two-stage method is the Faster-RCNN [3]. One-stage methods included SSD [24] and RetinaNet [25]. Anchor-free methods include CenterNet [12] and CenterNet2 [16]. The correspondence between the categories and objects in the DIOR is shown in Table 1.

Table 1 : Correspondence between categories and objects in the DIOR dataset

C1	Airplane	C11	Overpass
C2	Airport	C12	Stadium
C3	Bridge	C13	Train station
C4	Basketball court	C14	Storage tank
C5	Ship	C15	Ground track field
C6	Expressway toll station	C16	Tennis court
C7	Golf field	C17	Expressway service area
C8	Harbor	C18	Windmill
C9	Chimney	C19	Vehicle
C10	Dam	C20	Baseball field

In the comparison experiments, all hyperparameters were consistent in Detectron2 and all experiments are conducted on a single 1080Ti GPU. The input image and patch sizes are set to 800 × 800 as the original image size for the DIOR dataset. For a fair comparison, we conduct experiments using Faster-RCNN [3] and SSD [24]. RetinaNet [25] used ResNet-101 as the backbone, whereas the proposed method uses ResNet-50. During the training phase, the batch size is fixed at 8. All models are initialized with a learning rate of 0.01. The number of warm-up iterations is adjusted within a range of 600 to 1500. Additionally, the non-maximum suppression (NMS) threshold is set to 0.75 to enhance the accuracy of object detection. Table 2 presents the comparison results of the model details, including the model size, reference time, and FPS. Table 3 shows the comparison results for the DIOR dataset.

Table 2: Comparison experiments of precision on the DIOR dataset

	Backbone	mAP	APS	APM	APL	AP50	AP75
Faster-RCNN	VGG-16	48.83	4.42	27.78	41.56	45.91	26.54
SSD	VGG-16	58.49	6.65	25.48	42.54	43.12	25.84
RetinaNet	ResNet-101	60.35	6.54	32.47	50.46	39.14	31.39
CenterNet	ResNet-50	52.90	5.27	34.14	66.62	63.52	57.64
CenterNet2	ResNet-50	54.52	8.32	37.43	72.01	66.54	53.61
Ours	ResNet-50	61.39	13.34	41.57	78.69	73.95	64.18

Table 3: Performance evaluation of different methods on the DIOR dataset. The model size, runtime and inference speed compare with SOTA model in the field of one-stage detector and anchor-free detector

Methods	Backbone	Size (M)	Time (ms)	FPS
Faster-RCNN	VGG-16	65.3	/	/
SSD	VGG-16	44.8	52	19.2

(Continued)

Table 3 (continued)

Methods	Backbone	Size (M)	Time (ms)	FPS
RetinaNet	ResNet-101	45.9	102.5	9.8
CenterNet	ResNet-50	50.7	44	19.3
CenterNet2	ResNet-50	47.4	40	22.4
Ours	ResNet-50	42.3	29	32.4

The proposed detector achieved the best performance with a mAP of 61.4. Additionally, the model size and FPS of the proposed detector were 42.3 M and 32.4 FPS, respectively. Speed and efficiency are important for the practical applications of remote-sensing target detection. The proposed detector has a significant improvement in the accuracy of small-object detection compared to the anchor-free model and other methods. Small object detection has always been a challenge on remote sensing public datasets. One-stage methods such as SSD and RetinaNet perform mediocre, and the original centerpoint detection method CenterNet can only achieve close to the accuracy of one-stage methods. CenterNet2, which introduces a two-stage model, improves the detection of small objects. In order to solve the problem of small objects, our proposed method uses the fully convolutional framework and probabilistic interpretation, and the experimental results show that the small object detection accuracy is further enhanced on this basis. In terms of model size and speed, anchor-based methods need to preset anchor boxes, resulting in model volume redundancy. The one-stage method is improved in terms of computation and has lightweight models. Our proposed method does not add computational burden to the center point detection method. At the same time, the reasoning speed is further improved.

Figs. 6a to 6e show the prediction results of Faster-RCNN, with FPN, which is the backbone of ResNet-101, RetinaNet, CenterNet and ProEnDet. The proposed approach reduced the number of incorrect predictions and exhibited better performance on objects with complex backgrounds. For large-scale objects, such as basketball fields, ProEnDet had a comparable result, particularly for small object vehicles between the three large objects. For harbours and ships, which overlap in the image, ProEnDet distinguished between these two categories, and small-object ships in the harbour can be accurately detected. Generally, ProEnDet achieves the detection of the SOTA model and better distinguishes the foreground and background based on the anchor-free method.

Our method yields comparable results for objects with complex backgrounds. As shown in Fig. 7, small dense targets are detected in the dataset, whereas background objects and overlapping small objects were distinguished. For typical small objects, such as vehicles (C19) and ships (C5), these two small targets achieve mAP of 60.1 and 43.7 in our method, respectively, which are improved compared with other methods. Ships and ports are a group of typical objects with inclusion relations, as shown in (a)–(d). The method achieves a comparable effect in the detection of dense ships for small targets and can detect overlapping ports at the same time. In (e), the motor vehicle and overpass or bridge have the same inclusion relationship, and the overpass and bridge are similar in characteristics. The proposed method can detect the correct object more accurately. In (f), the two types of objects have a large-scale difference, and the model does not miss the detection of small objects because of the large object. Table 4 presents the details of the comparison experiments for each category. The experimental data show that the accuracy of the anchor-free method is maintained for the detection of large and medium targets. Typical small targets like C1, C5, C18, C19, have comparable results in accuracy. C3,

C8, C10, and C11 similar class predictions are also improved. Average precision is improved by 6.94 to RetinaNet and 6.87 compared to CenterNet2.

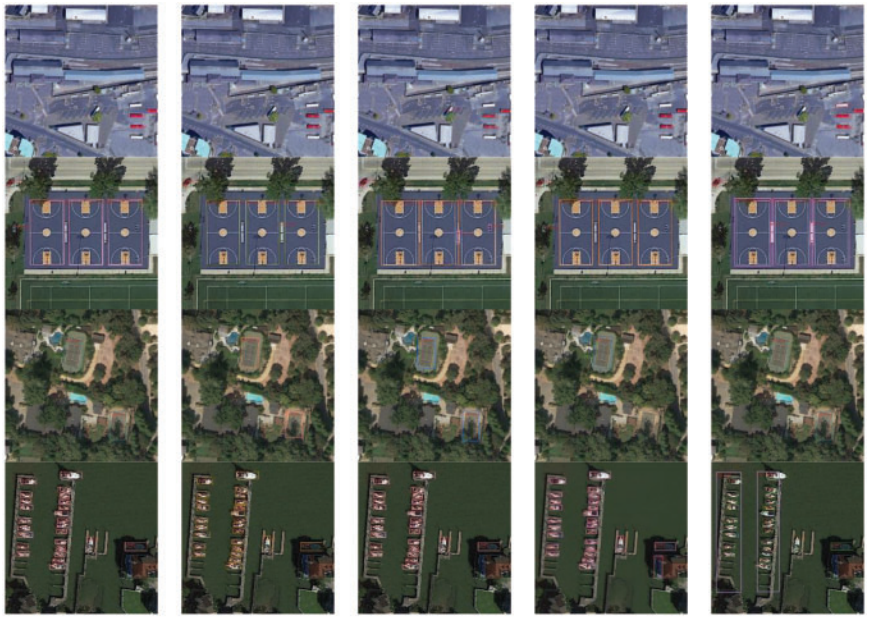

Figure 6: Prediction results on the DIOR dataset. (a) Faster-RCNN; (b) FPN; (c) RetinaNet; (d) CenterNet; (e) ProEnDet

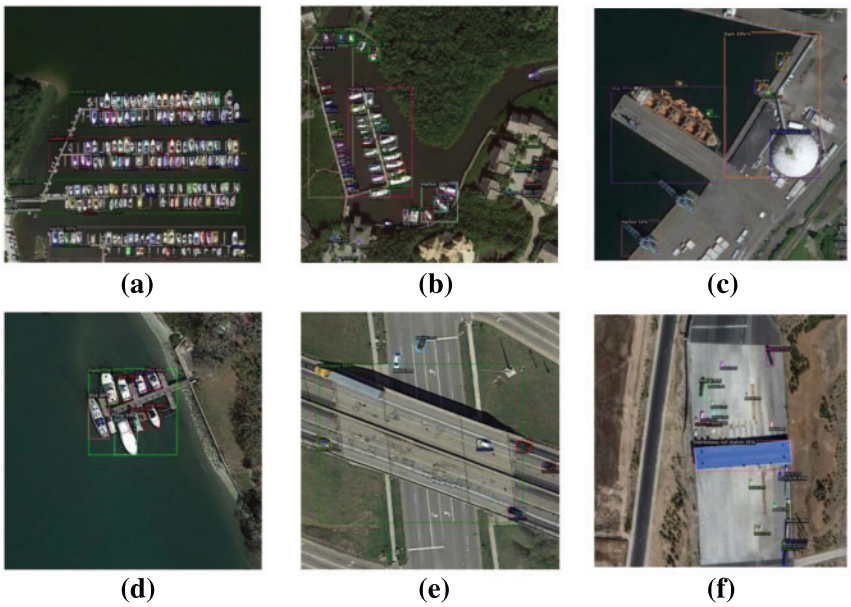

Figure 7: Visualization results of ProEnDet detection for categories with complex backgrounds or overlapping with related categories on the DIOR dataset

Table 4: APs of the comparison experiments. The category correspondence is shown in Table 2

Methods	Backbone	1	2	3	4	5	6	7	8
Faster-RCNN	VGG-16	43.67	49.35	28.09	36.22	27.74	55.25	55.60	30.23
SSD	VGG-16	72.73	54.80	28.93	39.83	33.89	44.34	52.03	30.11
FPN	ResNet-50	54.13	53.42	42.61	70.09	51.83	52.18	53.18	35.03
	ResNet-101	54.02	54.54	44.84	70.75	51.81	52.39	56.02	38.49
RetinaNet	ResNet-50	53.72	53.35	31.43	69.04	31.07	46.97	53.20	35.37
	ResNet-101	53.45	54.37	30.22	69.02	31.33	47.83	52.82	35.46
CenterNet	ResNet-50	67.09	51.71	25.75	36.39	39.67	52.40	51.93	32.04
CenterNet2	ResNet-50	68.12	52..02	26.71	35.73	40.84	49.98	56.26	34.74
Ours	ResNet-50	77.92	51.53	29.53	60.10	43.66	50.61	59.14	41.01
		9	10	11	12	13	14	15	16
Faster-RCNN	VGG-16	50.97	62.34	50.15	43.07	38.66	39.81	56.92	35.23
SSD	VGG-16	48.65	52.20	39.92	44.53	39.08	48.42	59.33	62.73
FPN	ResNet-50	73.03	57.51	40.02	57.01	36.48	53.54	55.58	80.22
	ResNet-101	72.55	60.08	42.23	68.31	39.51	53.58	56.86	79.87
RetinaNet	ResNet-50	51.14	44.56	41.84	56.67	33.75	52.09	55.37	48.59
	ResNet-101	50.24	44.72	42.35	56.31	33.32	52.33	57.84	48.31
CenterNet	ResNet-50	65.17	43.30	39.88	45.03	35.28	33.79	59.42	34.64
CenterNet2	ResNet-50	57.25	47.33	40.24	48.47	34.27	49.24	59.68	53.43
Ours	ResNet-50	75.91	42.30	43.21	73.14	35.99	66.11	62.27	78.90
		17	18	19	20	mAP			
Faster-RCNN	VGG-16	49.03	45.48	23.62	48.80	48.83			
SSD	VGG-16	32.82	34.52	38.91	32.91	58.49			
FPN	ResNet-50	47.72	40.87	43.12	63.30	54.19			
	ResNet-101	45.61	41.26	43.17	63.32	54.17			
RetinaNet	ResNet-50	37.92	41.31	29.15	78.08	53.95			
	ResNet-101	37.74	41.51	28.73	78.19	54.45			
CenterNet	ResNet-50	49.19	33.72	36.34	78.82	52.90			
CenterNet2	ResNet-50	40.91	35.90	35.99	78.33	54.52			
Ours	ResNet-50	52.08	36.76	39.85	80.17	61.39			

The experiments are conducted using the NWPU-VHR-10 dataset under identical hardware and environmental conditions. The batch size was established as 8, while the input dimension of the image was standardized to 800 × 800. We used Faster-RCNN, SSD [25], RetinaNet, CenterNet and CenterNet2. The proposed model was trained for over 70 epochs until convergence. The experimental results are presented in Fig. 8 and Table 5. Small targets could be detected in complex scenes, and dense and overlapping small targets were rarely missed. Objects with large-scale spans in the same

scene could also be accurately predicted. Compared with other methods, the accuracy of the small target was improved, and the overall accuracy was improved by 4% compared with the SOTA model.

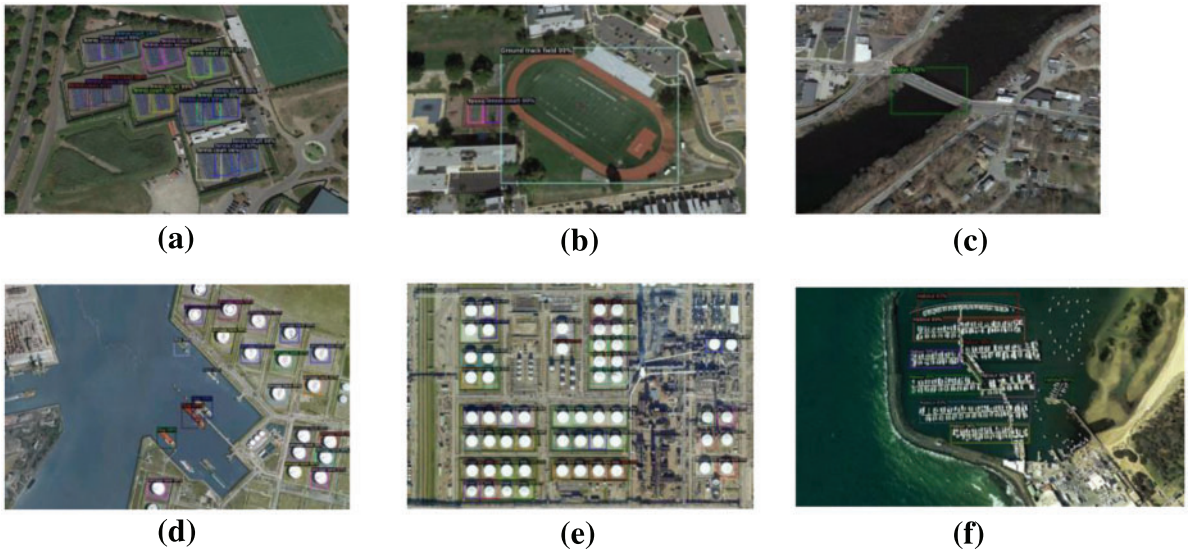

Figure 8: Visualization results of the NWPU-VHR-10 dataset of ProEnDet. (a) Including dense basketball courts. (b) Targets with large scale differences. (c) Targets with similar characteristics. (d) Dense small target ships and tanks. (e) Dense tanks and smaller vehicles. (f) Ports of inclusion relations and clustered small target ships

Table 5: APs of comparison experiments for the NWPU-VHR-10 dataset

Methods	Backbone	mAP	AP50	AP75	APS	APM	APL
Faster-RCNN	VGG-16	57.43	91.78	62.81	40.97	56.69	61.16
SSD	VGG-16	61.37	90.48	59.03	39.18	53.14	59.54
RetinaNet	ResNet-101	65.51	89.45	57.53	41.43	57.16	65.83
CenterNet	ResNet-50	63.27	91.34	58.77	43.18	59.83	64.32
CenterNet2	ResNet-50	64.03	90.86	58.32	43.45	61.36	64.56
Ours	ResNet-50	68.97	91.64	59.43	53.72	72.23	66.36

Compared with other methods, the proposed method demonstrated improvements in both accuracy and recall rate. The proposed detector achieved 35.9 FPS using an 1080Ti GPU, which was faster than other one-stage and anchor-free methods. Since this dataset and DIOR have duplicate classes and it is a small dataset, we use the same comparison model, and the proposed method obtains comparative results on this dataset. Similarly competitive results on other metrics with experiments on the DIOR. In the detection of small objects, the accuracy of the proposed method is better than CenterNet and the sequel.

3.5 Ablation Study

Table 6 lists the detailed parameters of the ablation experiment, FLOPs, and FPS for different necks. In practical applications, remote-sensing target detection must be performed in real-time and accurately. In many applications, models must be lightweight to perform downstream tasks and ensure portability. ProEnDet uses a repeated weighted bidirectional feature pyramid network, which more efficient integrates features at different scales more efficiently without additional computational effort. Compared with FPN [26] and PANet [27], the model volume did not increase, and real-time performance was achieved with the neck conducted in ProEnDet.

Table 6: Details of the ablation experiment on the DIOR dataset with different necks

Neck	Params (M)	FLOPs (G)	FPS
None	20.9	4.9	/
PANet	37.8	12.4	27.6
FPN	44.8	8.5	19.2
Ours	36.3	10.7	32.4

In the proposed method, an optimised feature extraction network is used. The results of the ablation experiments are listed in Table 7, where different backbones are replaced, including ResNet50 [28], DLA34 [15] and FCOS [13]. In the table, $\surd$ means the probabilistic interpretation is used. Based on the experimental results for different backbones, the effect of the full-convolution method combines with the bidirectional weighted pyramid was the best in practical applications. Probabilistic interpretation improved performance. Using the same architecture, the detection accuracy of small targets improves by 3% with probabilistic interpretation.

Table 7: Details of ablation experiment on the DIOR dataset. 'Pro' represents the probabilistic interpretation, and '$\surd$' represents the use of the probabilistic interpretation in the model

Backbone	Pro	mAP	APS	APM	APL	AP50	AP75
ResNet-50		42.33	4.56	29.43	50.13	32.07	30.93
ResNet-50 + FPN		45.13	6.68	32.98	63.89	49.55	36.98
DLA-34		51.35	5.37	32.93	65.16	71.34	52.78
DLA-34 + FPN		52.90	5.27	34.14	66.62	63.52	57.64
DLA-34	$\surd$	57.83	8.13	36.87	70.72	66.32	58.98
FCOS		58.22	6.15	34.89	60.38	59.93	53.90
FCOS	$\surd$	59.97	6.97	37.12	61.34	62.17	55.61
FCOS + BiFPN		56.62	8.31	36.13	63.13	62.07	51.31
FCOS + BiFPN	$\surd$	64.13	11.83	40.94	76.35	73.32	63.59

Table 8 shows an ablation experiment with loss and probabilistic interpretation. The experiments are conducted in a model with two-stage probability augmentation, and Res50 is used for the backbone network. A stricter IoU threshold is used in the experiments with the focal loss (loss) and multiplied by

the probabilities of the first and second stages (pro). All results use standard Res50-1x with multi-scale training.

Table 8: Details of ablation experiment on the DIOR dataset. A stricter IoU threshold with focal loss

P3–P7	Loss	Pro	mAP	Runtime
			34.2	52 ms
√			38.2	46 ms
√		√	38.6	45 ms
√	√		38.5	45 ms
√	√	√	41.3	42 ms

Both parts of the model are necessary. Multiplying the original RPN with the first-stage probability does not improve the accuracy. A strong network needs to be augmented by a reasonable probabilistic interpretation. We also add probabilistic interpretation to some two-stage detectors, such as Faster-RCNN and CascadeRCNN with RPN, to prove that use fewer proposals in the second stage, but does not improve accuracy. In more stringent IoU thresholds and focal loss such as RetinaNet, using two-segment probability improves accuracy. Overall, a strong backbone and suitable probabilistic hyperparameters can greatly improve the performance.

4 Conclusion

Most existing remote-sensing image-object detection methods rely on anchor boxes. However, the excessive use of anchor boxes introduces a considerable number of hyperparameters, which not only increases memory consumption but also results in redundant calculations. Furthermore, this approach can lead to critical problems such as an imbalance between positive and negative samples, causing further complications in the object detection process. The anchor-free method can avoid these problems and make more effective use of remote sensing image data, which enhances the performance. The anchor-free method has the advantage of being fast and has limitations in separating foreground and background. The proposed method enhances the performance of the detector by introducing a probability interpretation to the detector in one stage. The anchor-free method has the advantage of being fast and has limitations in separating foreground and background. The proposed method enhances the performance of the detector by introducing a probability interpretation to the detector in one stage. The interaction is inferred by the probability generated in the first stage and the likelihood of the head, which can better separate objects in complex natural scenes without increasing redundant calculations. The method has both the speed of one-stage detection and the accuracy of RPN effect in two-stage detection.

In this paper, we introduce ProEnDet, a detector designed for remote-sensing object detection. With the use of ProEnDet, the accuracy of detecting complex natural backgrounds is improved. We used an FCN combined with a bidirectional feature pyramid to enhance the foreground and background separation to enhance the feature extraction model. In addition, the probability component enhanced the prediction results. The experiments were verified on two public datasets (the DIOR and NWPU-VHR-10 datasets), and comparable results were achieved. By enhancing the probability of the object categories, the ability of the model to predict prospects was strengthened. The proposed method achieved comparable accuracy and speed for actual scenes. Proposed method combines the speed of

anchor-free detection and the accuracy of two-stage detection. It introduces probability interpretation for object detection in natural scenes and improves the separation effect of foreground and background for objects of different scales. However, when the target is occluded by other objects, its visible part may become incomplete, which increases the difficulty of detection. The center point detection method may not be able to effectively handle occlusion cases because it relies on the complete contour or features of the object for detection. Furthermore, the proposed method depends to some extent on the hyperparameter Settings and model design choices. There are still some limitations of the proposed method that need to be solved. Future research can improve and optimize these problems to further improve the performance and scope of the model. The anchor-free detection method is currently being applied in various fields. In the future, the detection with probabilistic interpretation can be combined with other frameworks to improve the effect of the model.

Acknowledgement: The authors would like to express their gratitude for the valuable feedback and suggestions provided by all the anonymous reviewers and the editorial team.

Funding Statement: This work was supported in part by the National Natural Science Foundation of China (42001408).

Author Contributions: The authors confirm contribution to the paper as follows: Study conception and design: C. Fan and Z. Fang; data collection: C. Fan; analysis and interpretation of results: C. Fan and Z. Fang; draft manuscript preparation: C. Fan and Z. Fang.

Availability of Data and Materials: In this study, we used public dataset DIOR and The NWPU-VHR-10, which can be downloaded from the website if needed (https://gcheng-nwpu.github.io/#Datasets).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *2017 IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 2980–2988. doi: 10.1109/ICCV.2017.322.

[2] R. Girshick, "Fast R-CNN," in *2015 IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, 2015, pp. 1440–1448. doi: 10.1109/ICCV.2015.169.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017. doi: 10.1109/TPAMI.2016.2577031.

[4] S. D. Khan, L. Alarabi, and S. Basalamah, "A unified deep learning framework of multi-scale detectors for geo-spatial object detection in high-resolution satellite images," *Arab. J. Sci. Eng.*, vol. 47, no. 8, pp. 9489–9504, 2022. doi: 10.1007/s13369-021-06288-x.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.

[6] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 6517–6525. doi: 10.1109/CVPR.2017.690.

[7] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.

[8] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.

[9] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, 2023, pp. 7464–7475. doi: 10.1109/CVPR52729.2023.00721.
[10] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," *Int. J. Comput. Vis.*, vol. 128, no. 3, pp. 642–656, 2020. doi: 10.1007/s11263-019-01204-1.
[11] X. Zhou, J. Zhuo, and P. Krähenbühl, "Bottom-up object detection by grouping extreme and center points," in *2019 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, 2019, pp. 850–859. doi: 10.1109/CVPR.2019.00094.
[12] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," arXiv preprint arXiv:1904.07850, 2019.
[13] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully," in *2019 IEEE Int. Conf. Comput. Vis. (ICCV)*, Seoul, Korea (South), 2019, pp. 9626–9635. doi: 10.1109/ICCV.2019.00972.
[14] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, 2016, pp. 483–499. doi: 10.1007/978-3-319-46484-8_29.
[15] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 2403–2412. doi: 10.1109/CVPR.2018.00255.
[16] X. Zhou, V. Koltun, and P. Krähenbühl, "Probabilistic two-stage detection," arXiv preprint arXiv:2103.07461, 2021.
[17] Z. Liu, T. Zheng, G. Xu, Z. Yang, H. Liu and D. Cai, "Training-time-friendly network for real-time object detection," in *Proc. AAAI Conf. Art. Intell.*, New York, USA, 2020, pp. 11685–11692. doi: 10.1609/aaai.v34i07.6838.
[18] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, 2020, pp. 9756–9765. doi: 10.1109/CVPR42600.2020.00978.
[19] J. Yan, L. Zhao, W. Diao, H. Wang, and X. Sun, "AF-EMS detector: Improve the multi-scale detection performance of the anchor-free detector," *Remote Sens.*, vol. 13, no. 2, pp. 160, 2021. doi: 10.3390/rs13020160.
[20] M. Wang, Q. Li, Y. Gu, and J. Pan, "Highly efficient anchor-free oriented small object detection for remote sensing images via periodic pseudo-domain," *Remote Sens.*, vol. 15, no. 15, pp. 3854, 2023. doi: 10.3390/rs15153854.
[21] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, 2020. doi: 10.1016/j.isprsjprs.2019.11.023.
[22] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *2020 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, 2020, pp. 10778–10787. doi: 10.1109/CVPR42600.2020.01079.
[23] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, 2016. doi: 10.1016/j.isprsjprs.2016.03.014.
[24] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, 2016, pp. 21–37. doi: 10.1007/978-3-319-46448-0_2.
[25] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, 1 Feb. 2020. doi: 10.1109/TPAMI.2018.2858826.
[26] T. Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 936–944. doi: 10.1109/CVPR.2017.106.
[27] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 8759–8768. doi: 10.1109/CVPR.2018.00913.
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.