IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2021.Doi Number

Remote Sensing Image Object Detection Based on Angle Classification

PENGFEI SHI[1], (Member, IEEE), ZHONGXIN ZHAO[3], XINNAN FAN[1], XIJUN YAN[3], WEI YAN[1], AND YUANXUE XIN[1,2]

[1]College of Internet of Things Engineering, Hohai University, Changzhou 213022, China
[2]National Mobile Communications Research Laboratory, Southeast University, Nanjing 211189, China
[3]School of Computer and Information, Hohai University, Nanjing 211100, China

Corresponding author: Xinnan Fan (e-mail: fanxn@hhuc.edu.cn).

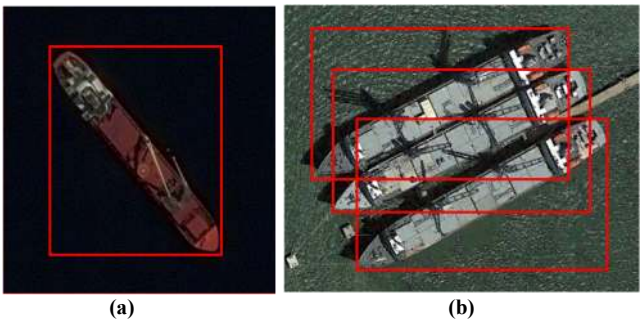
This work was supported in part by the National Natural Science Foundation of China (NSFC) under grant No. 61801169, in part by the Applied Basic Research Programs of Changzhou (CJ20200061), in part by the Fundamental Research Funds for the Central Universities (B210202087), and in part by the open research fund of National Mobile Communications Research Laboratory, Southeast University (No.2020D15).

ABSTRACT Arbitrarily-oriented object detection is a challenging task. Since the object orientation in remote sensing images is arbitrary, using horizontal bounding boxes will lead to low detection accuracy. Existing regression-based rotation detectors can lead to the problem of boundary discontinuity. In this paper, we propose a remote sensing image object detection method based on angle classification that uses rotation detection bounding boxes with angle information to detect objects. Specifically, we incorporate the neural architecture search framework with feature pyramid network (NAS-FPN) module in a dense detector (RetinaNet) and use a binary encoding method in angle classification. This method reduces the background influence, so that there is almost no overlap between detection boxes. Based on the angles of the detection boxes, we can infer the information of the motion direction of the target and further determine the motion trajectory of the target. We conducted ablation experiments on a large publicly available dataset for object detection in an aerial imagery (DOTA) dataset to verify the effectiveness of each module in the method and compared the method with several other detection methods. The experimental results demonstrate the effectiveness of our method.

INDEX TERMS Remote sensing image, Angel classification, Rotation detection frame, Object detection, Deep learning.

I. INTRODUCTION

Objection detection is a fundamental task in computer vision, and many researchers have applied the horizontal bounding boxes to locate objects in images. The use of horizontal bounding boxes can make the representation of candidate regions more concise and intuitive. In many methods based on deep learning [1]-[5], a large number of labeled samples are often needed to train the object detector model, and using an axis-parallel labeling frame can greatly improve the efficiency of labeling, to quickly obtain a large number of labeled samples quickly. In addition, the horizontal bounding boxes involve fewer parameters, simplifying the training process of the detection model. Therefore, in most object detection methods, a horizontal bounding box is used to represent the approximate range of the target in remote sensing images, as shown in Fig. 1.



(a) (b)

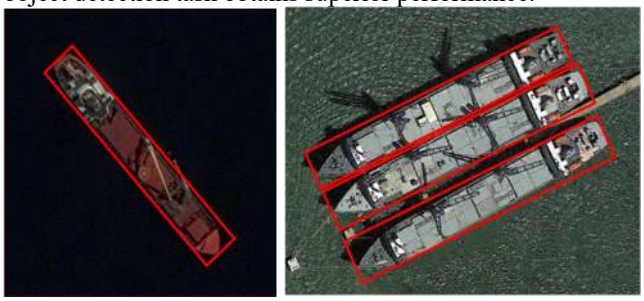
FIGURE 1. Horizontal detection frame.

However, objects in aerial images are often arbitrarily oriented. Therefore, the use of horizontal bounding boxes to detect objects [6]-[9], will give rise to several problems. First this type of object detection frame often contains many background areas. As shown in Fig. 1(a), approximately 60%

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107358, IEEE Access

of the pixel area in the figure belongs to the background area. The presence of too many background regions in the detection frame not only increases the difficulty of the classification task, but also leads to the problem of inaccurate representation of the target range. Second, the horizontal bounding boxes will lead to strong overlap between detection frames, as shown in Fig. 1(b), reducing detection accuracy. Finally, since the objects in images such as aircraft, ships, and vehicles contain motion direction information, information regarding the direction of motion of the targets cannot be obtained if the horizontal bounding boxes are used.

The above three problems can be effectively solved by using a rotation detection frame with angle information, as shown in Fig. 2. First, rotation detection can precisely locate the targets in the images, and the bounding boxes contain almost no background area, thus greatly reducing the influence of background on object classification. Second, there is almost no overlap between the rotating detection frames, so that the objects contained in the frames can be more clearly identified. Finally, the motion direction information of the object can be roughly obtained from the rotating detection frame, so that the motion trajectory of the object can be judged. In summary, the use of rotation detection with angle information in remote sensing image object detection task obtains superior performance.



(a) (b)

FIGURE 2. Rotation detection frame.

A. RELATED WORK

Most of the classical target detection methods use horizontal detection frames. Object detection methods using deep learning can be broadly classified into two categories: two-stage detectors and single-stage detectors. Two-stage detectors first extract candidate regions from images, and then predict and classify objects within the candidate regions.R-CNN [1] applies complex neural networks to the target detection task, followed by Fast-RCNN [2] and Faster-RCNN [3] that, are faster and have higher detection speed. Single-stage detectors predict the bounding box and class probability of the parent with only one evaluation of the image. In the field of single-stage detectors, the representative methods include SSD [10] and YOLO [11]. Single-stage methods have higher detection speed than two-stage methods. However, the objects in remote sensing images have the characteristics of small size, large scale difference and diverse directions, so that horizontal detection frame cannot detect the target accurately.

Rotation detectors are used in a wide range of applications in aerial images and scene texts. [12,13] use rotation detection frames to detect ships in aerial images. In recent years, deep learning techniques have developed rapidly, and many researchers have applied these techniques to target detection in remote sensing images. [14] adds the angle regression function into the detector to achieve the detection of arbitrary angle objects, and [15] improves the two-stage detection algorithm to generate the rotation bracketing box using regression to improve the detection accuracy. The scenes in remote sensing images are generally complex, with a large number of objects and uncertain angles. To solve these problems, some robust algorithms have been developed, such as some current state-of-the-art algorithms SCRDet [16], and ROI-Transformer [17]. However, most of the abovementioned algorithms have boundary problems due to the regression approach [18,19]. In this paper, we propose to avoid the boundary problem by using classification algorithms instead of regression algorithms.

B. CONTRIBUTION

In this paper, we hope to find a method to avoid the boundary problem and at the same time be able to improve the accuracy of object detection. Specifically, we propose an object detection algorithm for remote sensing images based on angle classification. The method uses a deep residual network to extract features in remote sensing images, employs the long-edge definition method and uses a neural structure search-based feature pyramid network (NAS-FPN) [20] for fusion of feature maps at different scales. The long-edge definition method is then used to represent the rotation detection frame and the binary-valued coded labeling technique from the dense coded labeling technique is utilized in the detection frame regression task. This technique transforms the angle regression problem into an angle classification problem, which can avoid the problem of boundary discontinuity that exists in the long-edge definition method. The main contributions of this paper are as follows.

(1)We use the IoU smooth L1 loss function on the angle-based classification method in the regression loss part of calculating the bounding box, which measures the intersection ratio between the predicted and true value boxes. We validate the effectiveness of the network on a large publicly available dataset for object detection in an aerial imagery (DOTA) dataset and the detection accuracy of the network is better than that of some current remote sensing image based target detection methods.

(2)We use a rotation detector based on angle classification to avoid the boundary discontinuity problem that occurs with parametric regression methods, and we use a binary encoding tag-based encoding method for angle classification, which has a shorter encoding length compared to other encoding methods and can improve the model efficiency.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107358, IEEE Access

II. OBJECT DETECTION METHOD BASED ON ANGLE CLASSIFICATION

A. METHODOLODY ARCHITECTURE

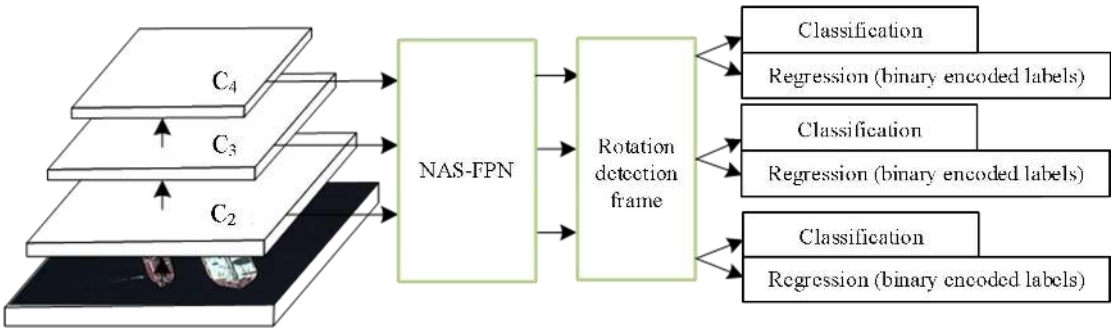


FIGURE 3. Block diagram of the overall network.

The proposed rotation object detector framework is presented in Fig 3. Our network is based on the RetinaNet framework. The feature maps labeled C2, C3, and C4 in the figure are extracted by the deep convolutional neural network. The overall steps of the method are as follows: first, the feature extraction network is used to extract the features in the remote sensing images, and the NAS-FPN is used to fuse the extracted features to obtain the feature maps at different scales. Then, we use the long-edge definition method to represent the rotation detection frame, and the binary encoding labeling technique is used to transform the angle regression problem into an angle classification problem in the frame regression task. Some of the important structures in the method are described in detail below.

The backbone network RetinaNet used in this paper is an end-to-end target detection algorithm, and based on this network, we replace the obsolete parts of the RetinaNet network with new techniques that do not harm the end-to-end learning approach. We apply this method to the rotation detection task, and from the experimental results, the method in this paper performs well in all 15 classes of images in the DOTA dataset, and has the best overall performance. Moreover, this paper verifies the effectiveness of each module of the network by ablation experiments.

B. NAS-FPN MODULE

In recent years, deep learning has been widely used in a variety of automated tasks. The success of deep learning relies heavily on the powerful learning capability of the technique, the huge amount of data, and the evolving hardware capabilities. The most critical task in deep learning techniques is the design of the neural structure, such as designing the number of layers of the network. The design of neural architecture is also known as neural architecture search (NAS) [21]. Most of the NAS still relies on manual analysis that does not guarantee the stability of the neural structure. To address this problem, researchers have started to focus on the study of neural architecture search that can learn autonomously.

A feature pyramid network can fuse feature maps at different scales, but the network focuses too much on low-level features and neglects the optimization of high-level features, leadings to a decrease in the detection accuracy of large objects. Moreover, the network is based on manual design, and since the number of combinations of feature maps at different scales increases exponentially with the number of the layers in the network, the manual design approach will lead to a huge design space, making the performance of the feature pyramid network not necessarily optimal. To obtain a feature pyramid network with better performance and more variability, Ghaisi *et al.* combined the idea of cross-layer connectivity to find a feature pyramid network structure with optimal performance in a deterministic search space. The structure is called NAS-FPN.

In NAS-FPN, the most important structure is the merged cell structure that consists of a collection of feature graph nodes, a pool of operations, and a search termination condition. Below, the search process of the feature graph is briefly described in the context of Fig. 4.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107358, IEEE Access

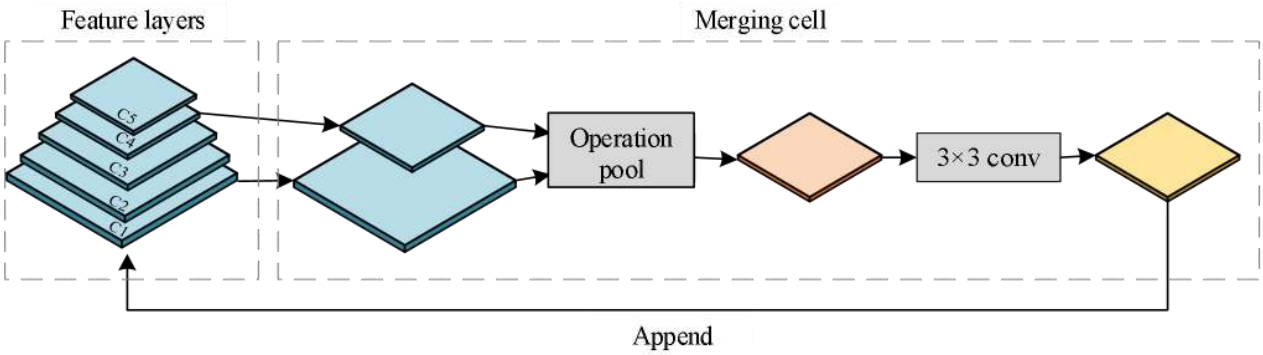


FIGURE 4. Structure of the merging unit.

(1) A feature map is randomly selected from the feature map node set as one of the inputs. The initial set of feature map nodes contains five scales of feature maps, denoted as {C1, C2, C3, C4, C5}.

(2) Randomly select another feature map from the feature map node set as another input.

(3) Select the resolution of the output feature map.

(4) Select an operation in the operation pool to operate on the feature map nodes selected in (1)(2) to produce a feature map with the same resolution as the output feature map and add this feature map to the feature map node collection for selection.

(5) The above steps are repeated cyclically. The termination condition of the search is to generate five feature pyramid networks with the same resolution as the initial feature map, which is denoted as {P1, P2, P3, P4, P5}.

Step (4) consists of two operations, namely sum and global pooling. The sum operation scales the smaller of the two input feature maps to the same size as the larger feature map, and then fuses the two feature maps by using pixel-by-pixel summing. The global pooling operation pools the smaller feature map, multiplies the larger feature map by pixel after the sigmoid operation, and then adds the obtained feature map with the smaller feature map for fusion. The feature pyramid network obtained by the NAS-FPN module achieves a certain improvement in the detection accuracy of the object detection method.

C. ROTATION DETECTION FRAME

There are three typical angle coding methods, including two five-parameter methods for different angle ranges [22] and an eight-parameter method [23]-[25]. The details are as follows:

(1) Five-parameter method with 90° angular range (OpenCV definition method): its schematic diagram is shown in Fig. 5. This definition method contains five parameters [*x, y, w, h, θ*]. Here, x and y are the center coordinates of the rotating frame, θ is the acute angle between the rotating frame and the x-axis, and the counterclockwise direction is specified as the negative angle, so that the angle range is [-90°,0); the width *w* of the rotating frame is the side where the rotating frame is located in the angle, and the height h of the rotating frame is the other side.

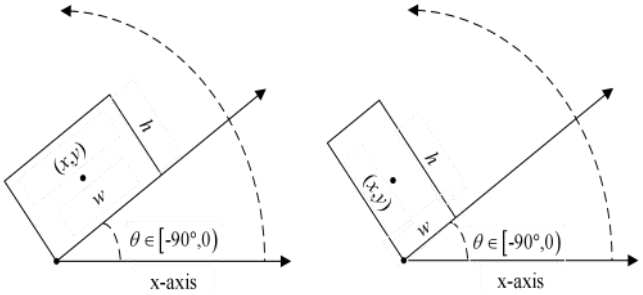


FIGURE 5. Five-parameter definition method for the 90° range.

(2) Five-parameter method with 180° angular range (long-side definition method): its schematic diagram is shown in Fig. 6. The definition method also contains five parameters [*x, y, w, h, θ*], *x* and *y* represent the center coordinates of the rotation frame. The difference between the two definition methods is that this definition method first specifies that the long side of the rotating frame is the height *h* and the short side is the width *w*. It also specifies that the counterclockwise direction is the negative angle and the clockwise direction is the positive angle, while the angle θ represents the angle between the height *h* and the x-axis of the rotating frame, and the angle range is [-90°,90°).

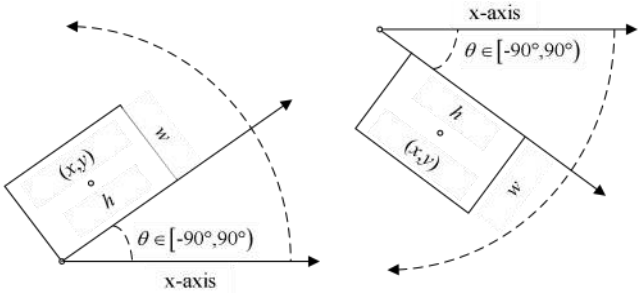


FIGURE 6. Five-parameter definition method for the 180° range.

(3) Eight-parameter method: The schematic diagram of this definition method is presented in Fig. 7, and shows that the definition method contains eight parameters [*a1, a2, b1,*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107358, IEEE Access

b2, c1, c2, d1, d2], and the point in the upper left corner of the definition is the starting point, and the remaining points are sorted counterclockwise.

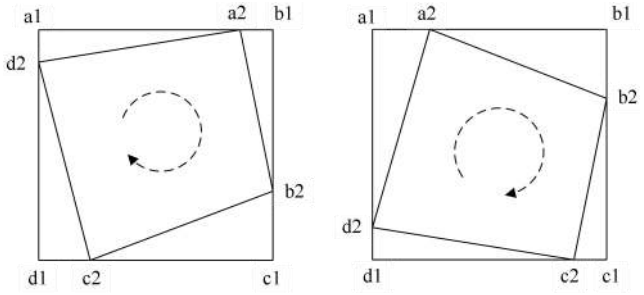


FIGURE 7. **Eight-parameter quadrilateral definition method.**

The representation of the rotating frame is not limited to the above three methods, but the representation of the rest of the rotating frame can be obtained by transforming the above three methods.

D. ANGLE CODING METHOD

Parametric regression is currently a popular method for rotation object detection. However, the parametric regression-based rotation detection method has some fundamental drawbacks. These methods often suffer the boundary discontinuity problem, leadings to inconsistent regression forms of the model at the boundary. The boundary discontinuity problem is mainly caused by the periodicity of angles and the exchangeability of edges. The periodicity of the angle and the commutativity of the edge will be explained in detail in the following section by combining the above three representation modes of the rotating frame.

(1) Five-parameter definition method for the 90° range: The boundary discontinuity problem of this rotating frame representation is sketched in Fig. 8.

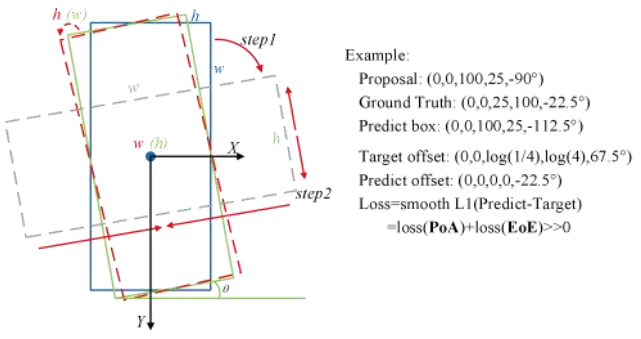


FIGURE 8. **Problems with the five-parameter definition method for the 90° range.**

In Fig. 8, the green box is the bounding box of the truth label that can be represented by a five-parameter definition of the 90° range as a value of [0,0,25,100, -22.5°], indicating that the width *w* is 25, the height *h* is 100, and the angle is -22.5°. The blue box is the proposed bounding box that can be expressed as [0,0,100,25, -90°] using the five-parameter definition method for the 90° range, indicating that the width *w* is 100, the height *h* is 25, and the angle is -90°. The red box is the prediction box, which can be expressed as [0,0,100,25, -112.5°], indicating that the width *w* is 100, the height *h* is 25, and the angle is -112.5°. This angle is the angle between the rotating box and the x-axis. It is observed from the figure, that the most ideal angle regression should be obtained by rotating the blue proposed box counterclockwise by 22.5° to obtain the red predicted box, at which the target offset is [0, 0, log(1/4), log(4),67.5°] and the predicted offset is [0, 0, 0, 0, 0, -22.5°]. The total loss is the difference between the predicted offset and the target offset after the smooth L1 function, and the total loss value is much larger than 0. From the above analysis, it is clear that the loss value of this rotating frame representation is not continuous, and the loss value at the boundary will increase suddenly. There are two main reasons for this phenomenon: first, the problem of the periodicity of the angle. Although angle rotation is a continuous process physically, the process has a large impact on the loss calculation. The second is the exchangeability of the edges. In the five-parameter definition method in the 90° range, the width *w* and height *h* may switch with each other, leading to a mismatch between the width and height of the proposed box, the true value box and the prediction box that give rise to a further increase in the loss value. To reduce the loss value, the network must adopt a more complex regression method, for example rotating the blue proposal box 67.5° clockwise and scaling the width *w* and height *h* However, this method will greatly increase the difficulty of angle regression.

(2) Five-parameter definition method for the 180° range: The boundary discontinuity problem of the representation of this rotated box is illustrated in Fig. 9.

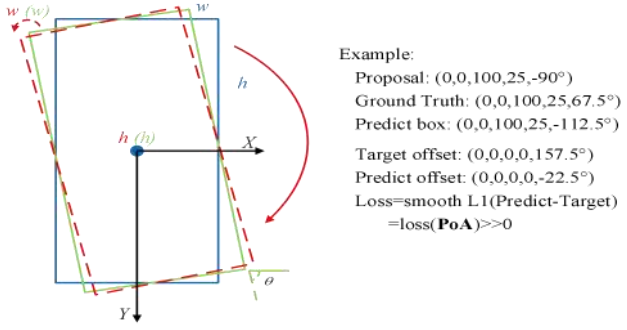


FIGURE 9. **Problems with the five-parameter definition method for the 180° range.**

In Fig. 9, the green box is also the bounding box of the truth label that can be represented by a five-parameter definition of the 180° range as a value of [0,0,100,25,67.5°], indicating that the width w is 100, the height h is 25, and the angle is 67.5°. The blue box is the proposed bounding box

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107358, IEEE Access

that can be expressed as [0,0,100, 25, -90°] using the five-parameter definition method for the 180º range, indicating that the width w is 100, the height h is 25, and the angle is -90°. The red box is the prediction box that can be expressed as [0,0,100, 25, -112.5°], indicating that the width w is 100, the height h is 25, and the angle is -112.5°. This angle represents the angle between the rotating box and the x-axis. It is observed from the figure that the optimal angle regression should be obtained by rotating the blue proposed box counterclockwise by 22.5° to obtain the red predicted box, the target offset is [0,0,0,0,157.5°] and the predict offset is [0,0,0,0, -22.5°]. It is obtained from the above analysis that the loss value of this rotated box representation also increases abruptly at the boundaries, but the only reason for this phenomenon is the periodicity of the angles and not the exchangeability of the edges, because this approach fixes the long and short sides of the rectangular box to be specified as w and h. Nevertheless, the loss in this method is still much larger than 0. To reduce the loss value, the network must use a more complex regression method, for example rotating the blue proposed box clockwise by 157.5°, but this method will also greatly increase the difficulty of the angle regression.

(3) Eight-parameter quadrilateral definition method: The problem of boundary discontinuity in the representation of the rotated box is shown in Fig. 10. The blue box is the proposed bounding box. If the red box is the truth label, after defining the distance and sorting the points according to the angle regression, the ideal is consistent with the actual angle regression as [(a1→a2), (b1→b2), (c1→c2), (d1→d2)]. When the green box is the truth label, the ideal and actual angle regressions are not consistent after the distance is defined and the points are sorted according to the angle regression. The ideal regression should be [(a1→b3), (b1→c3), (c1→d3), (d1→a3)], but the actual situation is [(a1→a3), (b1→b3), (c1→c3), (d1→d3)]. The problem also arises because of the existence of angular periodicity.

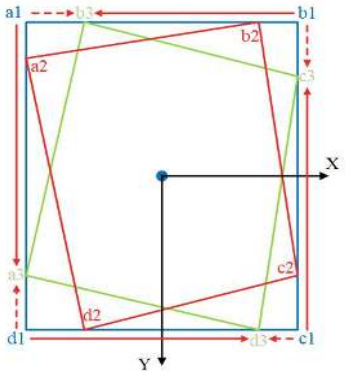


FIGURE 10. Problems with the eight-parameter quadrilateral definition method.

Rotation object detection methods based on angle regression have achieved good performance in various advanced vision tasks, and provide inspiration for many detection methods. However, these methods inevitably suffer from the boundary discontinuity problem that is usually caused by the angular periodicity and edge exchangeability in the five-parameter definition method and the angular point arrangement order in the eight-parameter definition method. The boundary discontinuity will cause problems such as the sudden increase in the loss value of the model at the boundary and the inconsistency of the regression form at the boundary and at the non-boundary. Although some special tricks are incorporated in many rotation object detection methods based on angle regression to alleviate the boundary discontinuity problem, these tricks increase the computational cost of the model and the difficulty of boundary prediction making these models unsuitable for the high-precision rotation object detection task, and reduce the detection accuracy of large aspect ratio objects. The boundary discontinuity problem that occurs with the rotation detection method based on angle regression usually arises because of the angular periodicity or corner ordering. and the root cause is not limited to a specific representation of the bounding box; therefore, to avoid the boundary discontinuity problem, we adopt the detection method based on angle classification.

Angle classification is to encode each angle. Each angle is considered a category, and the angle prediction problem is transformed into an angle classification problem. The commonly used angle encoding methods are shown in Fig. 11.

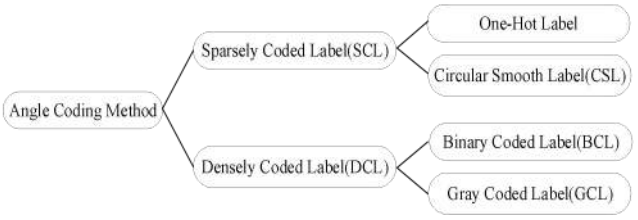


FIGURE 11. Angle coding methods.

There are two types of commonly used angle coding methods: sparse coded labels [19] and dense coded labels [26]. The sparse coding labels contain one-hot labels and circular smoothing labels (CSL), while the dense coding labels contain binary coding labels (BCL) and grayscale coding labels (GCL). It has been experimentally demonstrated that the object detection performance of the angle coding method based on binary coding labels is better than those of the other angle coding methods [26]. Other encoding methods require a longer number of bits for the encoding, while the binary encoding tag-based encoding method has a shorter encoding length compared to other methods, thus improving model efficiency. Therefore, the following section focuses on the binary encoding labeling process in dense encoding labels.

Table 1 shows the encoding process of the binary encoded tag, and Table 2 shows the decoding process of the binary encoded tag.

TABLE 1. Encoding process of binary encoded tags.

Input: angular range *AR*, discretized granularity ω.
Output: list containing all encoded binary labels *L*.
Initialized list *L*=[], Encoding length $n=\log_2(AR/\omega)$;
for *i* in *AR* do
result=bin(*i*,*n*)=bin(-round((*i*-90°)/ω)) # Convert angle *I* to *n-bit* binary *result*
L.append(*result*) # Add the binary number *result* to the list *L*
end for
return *L*

TABLE 2. Decoding process of binary encoded tags.

Input: list of predicted probabilities *p* at each position in the encoded binary numbers, discretized at granularity ω.
Output: predicted angle θ_{pred}.
Initialize the list *B*=[],the list of known predicted probabilities *p*;
for *j* in *p* do:
result=round(sigmoid(log(*j*/(1-*j*)))) #Convert each predicted probability *j* in list *p* to a binary number 0 or
B.append(*result*) # add the binary number *result* to the list *B*
θ_{pred}=90° -int(*B*)*ω # int(*B*) means convert list *B* to a decimal number
end for
return θ_{pred}

Below, we provide a specific example in order to illustrate the encoding and decoding process of binary encoded tags, as shown in Fig.12.

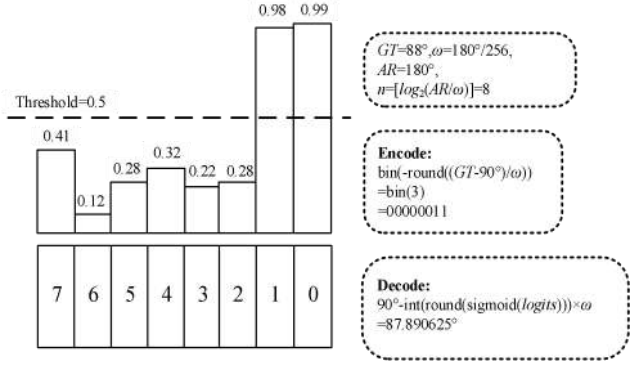


FIGURE 12. Binary encoding and decoding example.

In the encoding process, it is assumed that the angle GT=88° for the true value label box; and it is assumed that the angle size ω=180°/256≈0.703125° for each category because we use the five-parameter definition method of 180° range to represent the rotation box, then the angle range AR=180°; the encoding length $n=\log_2(AR/\omega)=\log_2256=8$, representing the use of an 8-bit binary number for encoding. The result of -round((*GT*-90°)/ω) is converted to a binary number and the final code is 00000011.

In the decoding process, the predicted probabilities at each position of an encoded binary number are assumed to be [0.41,0.12,0.28,0.32,0.22,0.28,0.98,0.99]. Each probability was rounded by round(sigmoid(log(*j*/(1-*j*)))) to obtain a binary list [0,0,0,0,0,0,1,1] that was converted to a predicted angle of 87.890625° and rounded to 88°.

E. LOSS FUNCTION

We use a multitask loss function to describe the difference between the true value and the predicted value. The multitask loss function contains three components: the regression loss of the bounding box, the classification loss of the angle, and the classification loss of the category, as descried in Eq 1.

$$L=\frac{\lambda_1}{N}\sum_{n=1}^{N}obj_n\sum_{j\in\{x,y,w,h\}}\frac{L_{reg}(v'_{nj},v_{nj})}{|L_{reg}(v'_{nj},v_{nj})|}|-\log(IoU)|$$
$$+\frac{\lambda_2}{N}\sum_{n=1}^{N}obj_nL_{bcl}(\theta_{gt},logits) \quad (1)$$
$$+\frac{\lambda_3}{N}\sum_{n=1}^{N}L_{cls}(p_n,t_n)$$

λ_1 is the weight coefficient, *N* represents the number of proposal boxes, obj_n is the binary value, obj_n=0 for background and obj_n=1 no regression for background, *x, y, w, h* is the center coordinates, width and height of the proposal box, *v ' nj* represents the prediction vector of *x , y, w, h; v ' nj* is the truth vector of *x, y, w, h*, as expressed by Eqs 2 and 3; $L_{reg}(v'_{nj},v_{nj})$ is calculated using the smooth L1 function, and *IoU* is the intersection ratio between the prediction frame and the truth frame.

In the regression loss part of the bounding box, the *IoU* smooth L1 function is used to calculate the loss value to further eliminate the discontinuity problem at the boundaries. In the categorical loss of angle part, λ_2 is the weight coefficient; *N* represents the number of proposed boxes; obj_n also represents a binary value, and $L_{cls}(p_n,t_n)$ is calculated by binary coded label loss function, as expressed by Eq 4.

In the classification loss part of the category, λ_3 is the weight coefficient; *N* represents the number of proposal frames, p_n represents the predicted probability distribution of each category, t_n represents the true value label, and $L_{cls}(p_n,t_n)$ is specifically calculated using the focus loss function.

In this paper, the hyperparameters λ1, λ2, and λ3 of the three components of the loss function are taken as 4, 1, and 2, respectively. The values of the three weight coefficients are derived from experiments, and we focus on the detection effect of the network after adding some new models. While there is no specific index to measure the angle prediction accuracy, the angle prediction accuracy and the target detection accuracy are consistent, and the more accurate angle prediction implies a more accurate target detection.

$$t'_x=(x'-x_a)/w_a,t'_y=(y'-y_a)/h_a,$$
$$t'_w=\log(w'/w_a),t'_h=\log(h'/h_a) \quad (2)$$

$$t_x = (x - x_a)/w_a, t_y = (y - y_a)/h_a, \quad t_w = \log(w/w_a), t_h = \log(h/h_a) \tag{3}$$

$$L_{bcl}(\theta_{gt}, logits) = FL(Encode_{bcl}(\theta_{gt}), logits) \tag{4}$$

In Eqs. 2 and 3, *x, y, w, and h* are the central horizontal coordinate, central vertical coordinate, width and height of the box respectively. The variables *x, x_a, x'* are the central horizontal coordinates of the true value box, the proposed box and the predicted box, respectively, *y, y_a, y'* are the central vertical coordinates of the true value box, the proposed box and the predicted box, respectively, *w_a, w'* are the width of the proposed box and the predicted box, respectively, and *h_a, h'* are the height of the proposed box and the predicted box, respectively.

In Eq. 4, θ_{gt} is the angle of the truth frame; *logits* is the list of angular prediction probabilities of the prediction frame, as shown in Eq. 5, where *p* is each prediction probability in the list, *FL* is the focal loss function, and $Encode_{bcl}$ is the binary encoding function, as shown in Table 1.

$$logits = \log\frac{p}{1-p} \tag{5}$$

III.EXPERIMENTAL PARAMETERS AND EVALUATION INDEXES

A. EXPERIMENTAL DETAIL

The experimental environment for our work is shown in Table 3.

TABLE 3. Experimental environment.

Items	Setting
Operating System	Ubuntu16.04
CPU	Intel Xeon E5-2680, 3.3GHz
Memory	128GB
GPU	Nvidia TITAN V, 11GB
GPU corresponding driver	Nvidia Driver 435.21, CUDA 10.0
Programming Languages	Python3.5
Deep Learning Framework	Tensorflow1.13.1

We use the DOTA dataset in this paper. DOTA is one of the largest aerial image detection benchmarks with quadrangle annotations. DOTA contains 2806 aerial images from different sensors and platforms and the size of the image ranges from approximately 800 × 800 to 4000 × 4000 pixels. The fully annotated DOTA benchmark contains 188282 instances of 15 object categories: plane (PL), ship (SH), storage tank (ST), baseball diamond (BD), tennis court (TC), basketball court (BC), ground field track (GTF), harbor (HA), bridge (BR), small vehicle (SV), large vehicle (LV), roundabout (RA), swimming pool (SP), helicopter (HC), and soccer ball field (SBF).

Instead of the horizontal labeling method and the five-parameter labeling method, the DOTA dataset was chosen to use the quadrilateral labeling method to label the four vertices of the object that can be combined with Fig. 13 to understand the labeling method of this dataset.

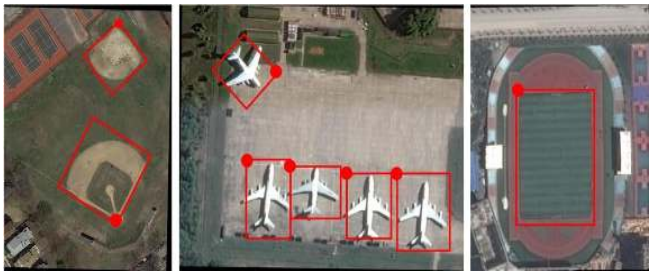


FIGURE 13. DOTA dataset annotation method.

Specifically, the starting point is marked first. Usually, the head of objects such as a baseball field, an airplane, or a vehicle is used as the starting point, but for objects such as a basketball court or a soccer field that do not have an obvious head, the top-left point is usually used as the starting point, and the remaining three vertices are then labeled clockwise.

The spatial resolution of the DOTA dataset is very high, giving rise to some difficulty in model training. Second, the size of the various types of objects in this dataset varies greatly, and most of the objects are small. For example, a car can be as small as 30 pixels and a bridge can be as large as 1200 pixels, which is 40 times the size of a car, as seen in Table 4. The DOTA dataset requires the model to be sufficiently flexible to handle both small and large objects. In addition, the objects in this dataset show a large variation in aspect ratio, further enhancing the difficulty of target detection in this dataset.

TABLE 4. Comparison of the proportion of targets of different sizes among the data sets.

Dataset	10-50 pixels	50-300 pixels	＞300 pixels
PASCAL VOC	0.14	0.61	0.25
MSCOCO	0.43	0.49	0.08
NWPU VHR-10	0.15	0.83	0.02
DOTA	0.57	0.41	0.02

B. EXPERIMENTAL PARAMETERS AND EVALUATION INDEXES

Table 5 lists the important experimental parameters of our method. The batch size is 1, corresponding to 1 image per training. We experimentally found that the best training results are achieved when the batch size is set to 1, so that we set the batch size to 1. The total number of training rounds (epochs) is 20; the momentum is 0.9; the initial learning rate is 0.0001; the Learning Rate Decay Rate is 10, indicating the decay rate of the learning rate, and the decay step of the

learning rate is 5, meaning that after every 5 rounds of training, the learning rate will decay by a factor of 10.

TABLE 5. Experimental parameters setting.

Parameter	Value
Batch Size	1
Epoch	20
Momentum	0.9
Learning Rate	0.0001
Learning Rate Decay Rate	10
Learning Rate Decay Step	5

The experiments still use the common evaluation metrics of object detection to evaluate the performance of our method, including the single-class average accuracy (AP), the multiclass average accuracy (mAP), precision, recall, and the F1-score.

C. EXPERIMENTAL PROCEDURE AND ANALYSIS OF RESULTS

Due to the large size of the images in the DOTA dataset, the images were cropped into smaller images of 600 pixel × 600 pixel for training prior to the training process. We generate new label information for the cropped images to facilitate the model training later. There are approximately 27,000 small images obtained after cropping. In the experiment, the training loss is reduced to less than 0.06 and we consider that the model has converged correctly. To verify the effectiveness of each module in our method, ablation experiments among the modules, including the NAS-FPN module, the binary coded label BCL module, and the *IoU* smooth L1 loss function, are performed first. Then, the method is compared with six existing high-performance rotating frame object detection methods to demonstrate the detection performance.

TABLE 6. *AP*, *mAP* results for ablation experiments on the DOTA dataset.

backbone	NAS-FPN	BCL	*IoU* smooth L1	PL	SH	ST	BD	TC	BC	GTF	HA	BR	SV	LV	RA	SP	HC	SBF	*mAP*
ResNet50	×	×	×	88.03	68.20	78.28	74.49	86.39	77.12	66.34	50.82	38.02	60.24	46.56	61.15	60.21	49.99	52.50	63.89
ResNet152	×	×	×	**88.92**	72.19	74.92	77.82	89.88	78.65	61.86	52.47	41.50	67.32	53.97	58.41	68.85	62.78	53.25	66.85
ResNet152	√	×	×	88.34	85.84	81.68	74.12	90.01	76.66	66.25	57.23	47.54	74.61	73.99	60.69	66.88	51.02	57.59	70.16
ResNet152	√	√	×	88.91	72.53	85.67	**83.18**	89.61	**85.95**	**69.32**	63.80	47.01	70.74	57.89	64.77	**72.66**	**66.38**	63.87	72.15
ResNet152	√	√	√	88.70	**86.41**	**86.31**	82.46	**90.02**	85.37	68.75	**67.80**	**52.81**	**78.51**	**81.45**	**65.20**	69.29	64.83	**65.10**	**75.53**

Note: Bolded font indicates the best results in each column, and the units of the values in the table are all %

In Table 6, when the base method uses ResNet50 as the feature extraction network, the average detection accuracy mAP is only 63.89%. When ResNet152 is used as the feature extraction network, the average detection accuracy (mAP) can reach 66.85%, which is an improvement of 2.96%. Therefore, ResNet152 is used as the feature extraction network in our method. After adding the NAS-FPN module, the mAP is improved by another 3.31%, when the binary coding module BCL is added again, the mAP is further improved by 1.99%, and finally, after adding the IoU smooth L1 loss function, the mAP reaches the highest value of 75.53%. Table 6 shows that the detection accuracy of most of the objects increases after the modules are added one by one. According to the above ablation experimental results, each module added to the basic method helps to improve the detection accuracy of the remote sensing image objects, illustrating the effectiveness of each module.

TABLE 7. Precision, recall, and F1-score before and after the method improvement.

Method	Precision	Recall	F1-score
Basic Method	28.93%	87.96%	40.31%
Our Method	42.30%	83.07%	52.56%

According to Table 7, compared with the basic method, the precision and F1-score have improved significantly, although the recall has decreased slightly, indicating that our method can detect the objects more accurately and comprehensively, and the detection performance is better than that of the basic method.

Fig. 14 shows the detection results of the basic method and our method on some images in the DOTA dataset, and only some of the original images are captured here to make the detection results more obvious. From the figure, we can see that the basic method is prone to mis-detection, such as the roundabout intersection and port in the figure; and the angle prediction of the basic method is not accurate enough, such as for the soccer field and tennis court in the figure. By contrast, our method can detect most of the objects with higher object detection accuracy, and more importantly, it can mark the location of the object using a rotating detection frame with a more accurate angle.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107358, IEEE Access

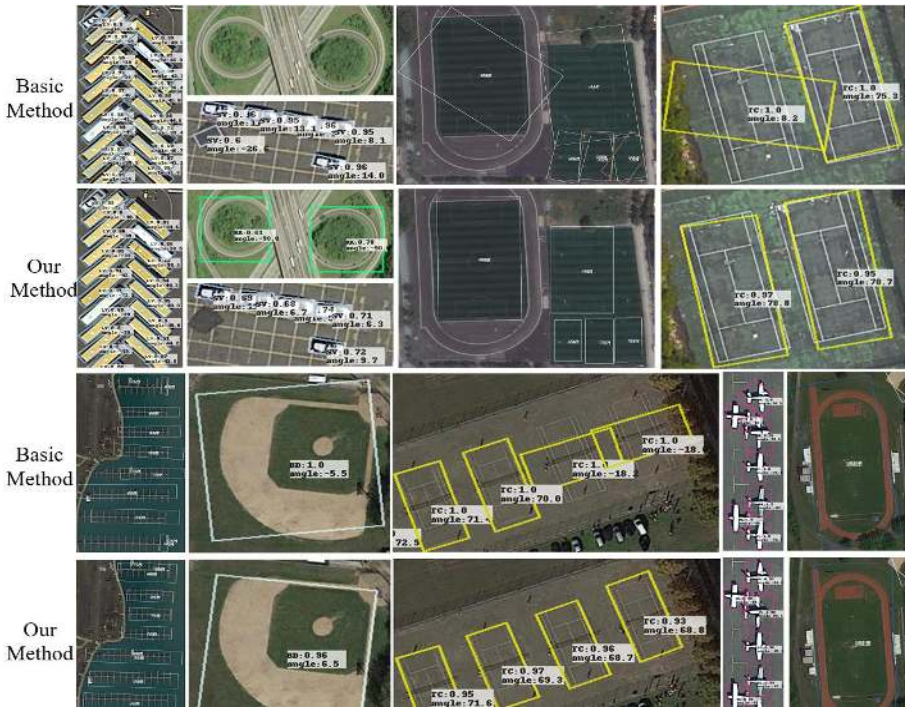


FIGURE 14. Comparison of the detection results of the basic method and our method.

We compared our method with eight existing high-performance methods, including R2CNN (Rotational Region CNN) [27], RRPN (Rotation Region Proposal Networks) [28], RetinaNet [29], ICN [30], RoI Transformer [17], CADNet [31], MFIAR-Net [32] and DRN [33]. The results of the comparison experiments are shown in Table 8, which includes the average detection accuracy *AP* and the multiclass average detection accuracy *mAP* for each type of target in the DOTA dataset. Since the publisher of the DOTA dataset does not publish the truth labels of the test set, the AP and *mAP* values discussed here are obtained by submitting the prediction files to the official DOTA evaluation server for evaluation.

TABLE 8. *AP*, *mAP* results of each method on the DOTA dataset.

Method	PL	SH	ST	BD	TC	BC	GTF	HA	BR	SV	LV	RA	SP	HC	SBF	*mAP*
R2CNN	80.94	55.81	72.39	65.67	90.67	66.92	67.44	55.14	35.34	59.92	50.91	52.23	53.35	48.22	55.06	60.67
RRPN	88.52	57.25	67.38	71.20	90.81	72.84	59.30	53.08	31.66	51.85	56.19	52.84	51.94	53.58	56.69	61.01
RetinaNet	88.92	75.24	75.07	67.67	90.87	73.95	56.83	51.05	33.55	66.11	73.28	56.72	55.86	21.46	43.77	62.02
ICN	81.40	70.00	78.20	74.30	90.80	79.10	70.30	67.00	47.70	64.90	67.80	62.90	64.20	50.20	53.60	68.20
RoI Transformer	88.64	83.59	81.46	78.52	90.74	77.27	**75.92**	62.83	43.44	68.81	73.68	53.54	58.93	47.67	58.39	69.56
CADNet	87.80	76.60	73.30	82.40	**90.90**	79.20	73.50	62.00	49.40	71.10	63.50	60.90	67.00	62.20	48.40	69.90
MFIAR-Net	89.62	77.81	**86.86**	**84.03**	90.85	85.40	70.30	66.25	52.41	70.13	67.64	**66.68**	**70.21**	62.11	63.21	73.49
DRN	**89.71**	85.84	84.89	82.34	90.57	**86.18**	64.10	69.30	47.22	76.22	74.43	61.93	69.63	58.48	57.65	73.23
Our Method	88.70	**86.41**	86.31	82.46	90.02	85.37	68.75	**67.80**	**52.81**	**78.51**	**81.45**	65.20	69.29	**64.83**	**65.10**	**75.53**

Note: Bolded font indicates the optimal results in each column, and the units of the values in the table are all in %

As illustrated in Table 8, our method not only has the optimal multiclass average accuracy compared with the eight object detection methods, but also has improved the single-class average accuracy for most of the objects. This reflects the greater ability of our method to use more accurate rotated detection frames to indicate the location and class of objects in remote sensing images.

As shown in Fig. 15, due to the large size of the measured images, small images containing typical scenes are selectively shown here. It is observed from the figure, that our method can accurately detect the position of the objects by using a rotating frame with an angle, and can also give the approximate angle value of the rotating frame.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107358, IEEE Access

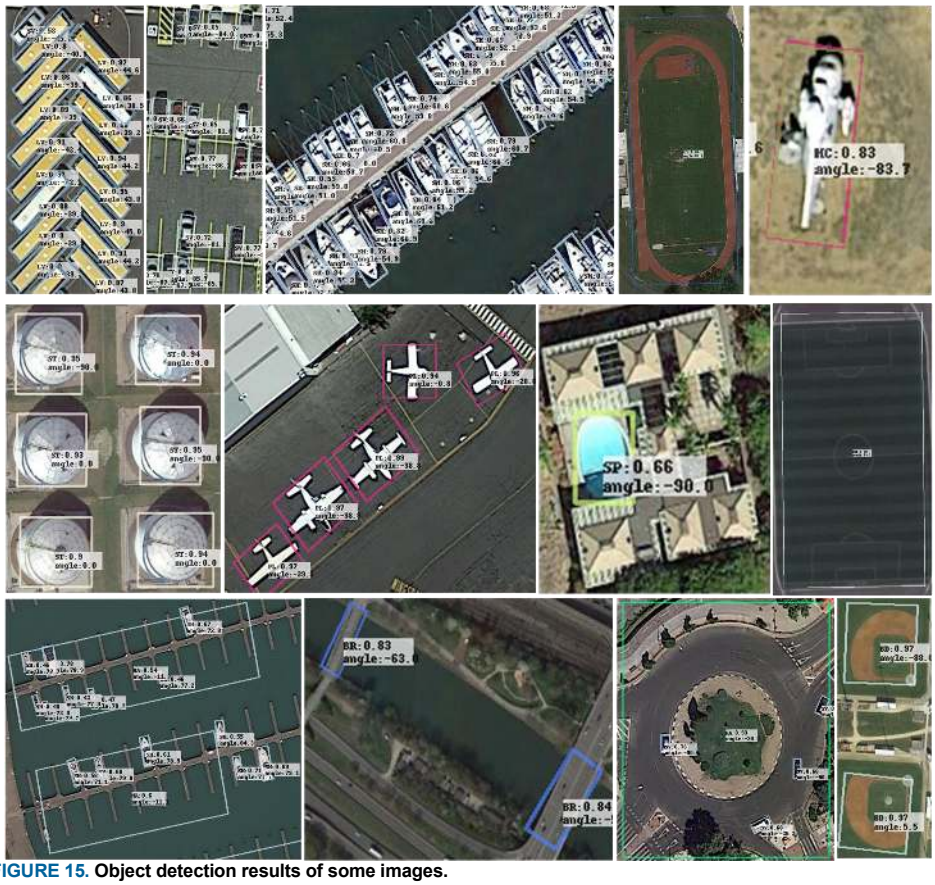


FIGURE 15. Object detection results of some images.

Our approach is an improvement for on RetinaNet. Since there is no precedent for combining these techniques, it is unclear whether their combination will produce better detection results. We performed ablation experiments and compared the accuracy with some other methods. According to the above experimental results, our proposed method (ResNet152+ NAS-FPN+ BCL+ IoU smooth L1) is superior to the other methods.

IV.CONCLUSION

In this paper, we proposed a rotation detector based on angle classification. The embodiment is improved on the basis of RetinaNet. First, we use the residual network to extract the features in images, and the feature pyramid network based on neural structure search is used to fuse the extracted feature maps to obtain feature maps of different scales. Then angle classification is used to avoid the problem of periodicity of angle, while the five-parameter definition method with 180º range is adopted to solve the problem of exchangeability of edges. Finally, the *IoU* smooth L1 function is added to the loss function to further eliminate the boundary discontinuity problem. The effectiveness of our method is verified by ablation experiments and comparison experiments, and the dataset used is the DOTA dataset. The results of the ablation experiments show that each proposed module contributes to improving the object detection accuracy. The results of the comparison experiment further demonstrate that the proposed method has higher accuracy in remote sensing image object detection compared with the comparison methods, and also can locate the objects with more accurate rotation detection frames.

In the field of machine learning, learning tasks can be broadly classified into the two categories of supervised learning and unsupervised learning. Usually, both need to learn predictive models from training datasets containing a large number of training samples. Although current supervised learning techniques have achieved great success, it is important to note that it is difficult to obtain strongly supervised information such as full truth labels for many tasks due to the high cost of the data labeling process. Unsupervised learning is quite difficult due to the slow development of the learning process. Therefore, weakly supervised learning has been gradually attracting attention. Several studies using weakly supervised learning in combination with perspective regression algorithms have already been reported. From the point of view of training, the difference between classification models and regression models is their loss functions. Some work on detection using weakly supervised learning with angle regression algorithms has been carried out [34], and we believe that detection methods based on angle classification can also be implemented using a deep learning approach of supervised learning; we expect to conduct research in this area in future work.

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. (CVPR),* 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,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[4] 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

[5] J. Redmon and A. Farhadi, ''YOLOv3: An incremental improvement,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Apr. 2018, pp. 1–6.

[6] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, ''Accurate object localization in remote sensing images based on convolutional neural networks,'' *IEEE Trans. Geoscience. Remote Sens*., vol. 55, no. 5, pp. 2486–2498, May 2017.

[7] G. Cheng, P. Zhou, and J. Han, ''Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images,'' *IEEE Trans. Geosci. Remote Sens*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[8] J. Yan, H. Wang, M. Yan, D. Wenhui, X. Sun, and H. Li, ''IoU-adaptive deformable R-CNN: Make full use of IoU for multi-class object detection in remote sensing imagery,'' *Remote Sens*, vol. 11, no. 3, p. 286, Feb. 2019.

[9] S. Chen, R. Zhan, and J. Zhang, ''Geospatial object detection in remote sensing imagery based on multiscale single-shot detector with activated semantics,'' *Remote Sens*, vol. 10, no. 6, p. 820, 2018.

[10] W. Liu, D. Anguelov , D. Erhan, C. Szegedy , S. Reed, C.-Y . Fu, and A. C. Berg, "SSD: Single shot multi box detector," in *Proc Eur. Conf. Comput. Vis.* Springer, 2016, pp.21–37.

[11] 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.

[12] Liu, G., Sun, X., Zhang, Y., and Zheng, "A new method on inshore ship detection in high-resolution satellite images using shape and context information," *IEEE geoscience and remote sensing letters*, vol. 11, no. 3, pp. 617–621, 2014.

[13] H. He, Y. Lin, F. Chen, H. M. Tai, and Z. Yin, "Inshore ship detection in remote sensing images via weighted pose voting," *IEEE Transactions on Geoscience Remote Sensing*, vol. 55, no. 6, pp. 3091–3107, 2017.

[14] L. Liu, Z. Pan, and B. Lei, "Learning a rotation invariant detector with rotatable bounding box," 2017, *arXiv: 1711.09405*. [Online]. Available: http://arxiv.org/abs/1711.09405.

[15] Z. Zhang, W. Guo, S. Zhu, and W. Yu, "Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks," *IEEE Geoscience and Remote Sensing Letters*, vol.15, no. 11, pp. 1745–1749, 2018.

[16] X. Yang, J. Yang, J. Yan, Y. Zhang, T. Zhang, Z. Guo, X. Sun, and K. Fu, "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV),* Oct. 2019, pp.8232–8241.

[17] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc IEEE/CVF Conf. Comput. Vis. Pattern Recognit (CVPR),* Jun. 2019, pp.2849-2858.

[18] W. Qian, X. Yang, S. Peng, Y. Guo, and J. Yan, "Learning modulated loss for rotated object detection," 2019, *arXiv: 1911.08299*. [Online]. Available: http://arxiv.org/abs/1911.08299.

[19] X. Yang and J. Yan, "Arbitrary-Oriented Object Detection with Circular Smooth Label," 2020, *arXiv: 2003.05597v2*. [Online]. Available: http://arxiv.org/abs/2003.05597v2.

[20] G. Ghiasi, T. Y. Lin, and Q. V. Le, "Nas-fpn: Learning scalable feature pyramid architecture for object detection," in *2019 IEEE/CVF Conf. Comput. Vis. Pattern Recognit (CVPR)*,2019, pp.7036-7045

[21] Zoph B, Le Q V. "Neural Architecture Search with Reinforcement Learning,"2016,*arXiv:1611.01578.*[Online]Available:http://arxiv.org/abs/1611.01578.

[22] X. Yang, S. Hao, K. Fu, J. Yang, S. Xian, M. Yan, and G. Zhi, "Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks," *Remote Sensing*, vol. 10, no. 1, p. 132, 2018.

[23] M. Liao, B. Shi, and X. Bai, "Textboxes++: A single-shot oriented scene text detector," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3676–3690, 2018.

[24] Y. Liu, S. Zhang, L. Jin, L. Xie, Y. Wu, and Z. Wang, "Omnidirectional scene text detection with sequential-free box discretization," in *Twenty-Eighth International Joint Conference on Artificial Intelligence IJCAI-19*, 2019.

[25] Y. Xu, M. Fu, Q. Wang, Y. Wang, and X. Bai, "Gliding vertex on the horizontal bounding box for multi-oriented object detection," 2019, *arXiv:1911.09358*.[Online].Available:http://arxiv.org/abs/1911.09358.

[26] X. Yang, L. Hou, Y. Zhou, W. Wang, and J. Yan, "Dense label encoding for boundary discontinuity free rotation detection," 2020, *arXiv:2011.09670*.[Online].Available:http://arxiv.org/abs/2011.09670.

[27] Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, P. Fu, and Z. Luo, "R2cnn: Rotational region cnn for orientation robust scene text detection,"2017,*arXiv:1706.09579*.[Online].Available:http://arxiv.org/abs/1706.09579.

[28] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, " Arbitrary-oriented scene text detection via rotation proposals," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 3111-3122,2018.

[29] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollr, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. PP, no. 99, pp. 2999–3007, 2017.

[30] S. M. Azimi, E. Vig, R. Bahmanyar, M. Körner, and P. Reinartz, "Towards multi-class object detection in unconstrained remote sensing imagery," in *Asian Conference on Computer Vision. Springer*, 2018, pp. 150–165.

[31] G. Zhang, S. Lu, and W. Zhang, "Cad-net: A context-aware detection network for objects in remote sensing imagery," IEEE Transaction son Geoscience and Remote Sensing, vol. PP, no. 99, pp. 1–10, 2019.

[32] F Yang, W Li, H Hu, and P Wang. Multi-Scale Feature "Integrated Attention-Based Rotation Network for Object Detection in VHR Aerial Images," *Sensors (Basel, Switzerland)*, vol 20, no 6, 2020.

[33] X Pan, Y Ren, K Sheng, W Dong and C Xu. "Dynamic Refinement Network for Oriented and Densely Packed Object Detection," in *Proc IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR),*2020.

[34] C Shi and Y Huang. "Cap-Count Guided Weakly Supervised Insulator Cap Missing Detection in Aerial Images," *IEEE Sensors Journal,* vol. 21, no. 21, pp. 1-1, 2020.