Rotating Object Detection in Remote-Sensing Environment

Sixian Chan

Zhejiang University of Technology

Jingcheng Zheng

Zhejiang University of Technology

Lina Wang

Southeast Digital Economic Development Institute

Tingting Wang

Macau University of Science and Technology

Xiaolong Zhou (✉ xiaolong@ieee.org)

Quzhou College: Quzhou University <https://orcid.org/0000-0003-0732-5169>

Yingkun Xu

Zhejiang University of Technology

Kai Fang

Quzhou College: Quzhou University

Research Article

Keywords: Deep learning, Object detection, Remote sensing, Arbitrary orientation.

Posted Date: August 24th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-804822/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Soft Computing on April 15th, 2022. See the published version at <https://doi.org/10.1007/s00500-022-07058-z>.

Rotating Object Detection in Remote-Sensing Environment

Sixian Chan¹ · Jingcheng Zheng¹ · Lina Wang² · Tingting Wang³ ·
Xiaolong Zhou^{4*} · Yinkun Xu¹ · Kai Fang⁴

Received: date / Accepted: date

Abstract Deep learning models have become the mainstream algorithm for processing computer vision tasks. In object detection tasks, the detection box is usually set as a rectangular box aligned with the coordinate axis, so as to achieve the complete package of the object. However, when facing some objects with large aspect ratio and angle, the bounding box has to become large, which makes the bounding box contain a large amount of useless background information. In this study, a different approach is taken, using a method based on YOLOv5, the angle information dimension is increased in head part and angle regression added at the same time of the border regression, combining ciou and smoothl1 to calculate the bounding box loss, so that the resulting border box fits the actual object more closely.

Sixian Chan
E-mail: sxchan@zjut.edu.cn

Jingcheng Zheng
E-mail: 2112012071@zjut.edu.cn

Lina Wang
E-mail: 531123178@qq.com

Tingting Wang
E-mail: tingtingwang@ieee.org

Xiaolong Zhou
E-mail: xiaolong@ieee.org

Kai Fang
E-mail: kaifang@ieee.org

Yinkun Xu
E-mail: xyk@zjut.edu.cn

¹ College of computer science and technology, Zhejiang University of Technology

² Southeast Digital Economic Development Institute

³ Faculty of Information Technology, Macau University of Science and Technology, Macau, China

⁴ College of Electrical and Information Engineering, Quzhou College

*Corresponding author

At the same time, the original dataset tags are also pre-processed to calculate the angle information of interest. The purpose of these improvements is to realize object detection with angles in remote-sensing images, especially for objects with large aspect ratios, such as ships, airplanes, and automobiles. Compared with the traditional object detection model based on deep learning, experimental results show that the proposed method has a unique effect in detecting rotating objects.

Keywords Deep learning · Object detection · Remote sensing · Arbitrary orientation.

1 Introduction

Object detection is a basic problem in machine vision. It supports visual tasks, such as instance segmentation, object tracking, and action recognition, and has a wide range of applications in automatic driving, unmanned aerial vehicles, monitoring, and other fields [Haghofer et al.(2020)Haghofer, Dorl, and Oszwald, Janakiramaiah et al.(2021)Janakiramaiah, Prasad, and Krishna]. At present, a detector based on a convolutional neural network (CNN) has been developed that is unprecedented in terms of both accuracy and speed and exceeds the capabilities of traditional detectors [Mahalingam and M(2019),Mahalingam and Subramoniam(2020)]. In the process of object recognition and location, a deep learning-based detector always uses a bounding box parallel to the coordinate axis to represent the object, identifies the object with a rectangular shape, and then classifies and distinguishes the actual object or background in the aforementioned rectangular box [Sudha and Priyadarshini(2020),Ahmed et al.(2017)Ahmed, Dogra, Kar, and Roy]. Generally speaking, most instances of object detection, such as people, animals, buildings,

and other objects observed from the perspective of parallel ground, are parallel to the coordinate axis of the image with a small-aspect-ratio shape, so the rectangular box can better enwrap them and contain less background [Sun et al.(2019)Sun, Yan, Huang, and Sun, Sun et al.(2019)Sun, Yan, Huang, and Sun]. However, in some special cases, especially in the case of remote-sensing images [Araújo et al.(2020)Araújo, Fontinele, and Oliveira], when the camera observes some objects with large aspect ratio and disordered direction, such as ships, vehicles, and seaports, from an observation angle perpendicular to the ground, it is impossible to accurately surround the object only with a rectangular frame [Zhang et al.(2020)Zhang, Mu, Kou, and Zhao]. An overly long width or height will enlarge the bounding box to an unnecessary degree approaching exaggeration. When detecting high-density objects with angles, overlapping frames make it difficult to distinguish the actual position of a single object [Wu et al.(2020)Wu, Zhang, Wang, Wang, Wang, and Li]. At the same time, as important information, the direction of the object is often ignored in a series of applications, e.g., follow-up tracking.

The goal of the work described in this paper is to achieve a rotated bounding box detection of directional objects for remote-sensing images. Specifically, a separate information dimension about the object angle is introduced, along with the corresponding loss and regression functions. This method can well indicate the direction of the object to be detected. By determining the angle, the proposed bounding box shape has a more accurate width and height to fit the actual target, which is of great significance not only in intuitive visual experience, but also in the re-application of the result data. At the same time, most of the current remote-sensing datasets are labeled with four points $(x_1y_1, x_2y_2, x_3y_3, x_4y_4)$ or there is no clear definition of the width and height of the object. Therefore, to obtain the desired effect after training, the data are reprocessed before training to calculate the real width and height of the object and the deflection angle relative to the coordinate axis. To summarize, the main contributions of this study are as follows.

1) Angle is added as an additional channel in the YOLOv5 object detection model, and on the basis of the box loss CIOU, the related angle loss and regression function are introduced. These improvements are mainly needed to solve the problem of object detection with rotation.

2) A reprocessing method for the existing datasets is proposed so that it can provide a real and effective bounding box for training.

Fig. 1 The difference between horizontal detector and arbitrary-oriented bounding box. The green box is the horizontal detector, the red box is the rotated bounding box.

3) The results show that the proposed method is better than the traditional horizontal region object detection method for remote-sensing datasets.

The organization of paper is as follows: the introduction of issues with the background studies on object detection is discussed in Sect 2. Our proposed method about angle parameter definition and regression are illustrated in Sect 3. The results and analysis of ablation and comparative experiments are given in Sect 4.

2 Related work

Horizontal region object detection. Classical object detection detects a general object in an image with a horizontal boundary box. At present, many high-performance general object detection methods have been proposed. The two-stage object detection model, represented by fast Region-based Convolutional Network (Fast R-CNN) [Girshick(2015)] and faster R-CNN [Ren et al.(2017)Ren, He, Girshick, and Sun], pays attention to the accuracy and reduces the amount of calculation to improve the detection speed. Feature pyramid network (FNP) method was proposed to deal with the scale change of the object in the image. Single Shot MultiBox Detector (SSD) [Liu et al.(2016)Liu, Anguelov, Erhan, Y., and Berg], Yolo [Redmon et al.(2016)Redmon, Divvala, Girshick, and Farhadi], and RetinaNet [Lin et al.(2017)Lin, Goyal, Girshick, He, and Dollár] represent single-stage detection methods, and their single-stage structure endows them with faster detection speeds. Compared with anchor-based methods, many anchor-free methods have become very popular in recent years. CornerNet [Law and Deng(2020)], CenterNet [Zhou et al.(2019)Zhou, Wang, and Krähenbühl], and ExtremeNet try to predict some key points of an object, such as corners or extreme points, and then group these key points into a bounding box. However, the horizontal detector used in such applications cannot provide accurate direction and scale information, which brings difficulties to the practical application of object change detection and multi-

Fig. 2 Disadvantages of the traditional bounding box.

directional scene text sequence character recognition in aerial images, as shown in Fig 2.

Arbitrarily oriented object detection. In recent years, object detection frameworks based on rotation quadrilateral or other polygon bounding boxes have become very popular. For example, Rotation Region Proposal Network (RRPN) [Ma et al.(2018)Ma, Shao, Ye, Wang, Wang, Zheng, and Xue] obtains the rotated region of interest (ROI) based on the rotated anchor, and extracts the depth feature from it; this is the first time that a rotation candidate box has been introduced to realize scene detection in any direction based on RPN architecture. Efficient and Accurate Scene Text Detector (EAST) [Zhou et al.(2017)Zhou, Yao, Wen, Wang, Zhou, He, and Liang] takes the distance between the feature points and four sides of the rotation frame as a new definition of the rotation object, and proposes the anchor-free rotation detection method mentioned above. Rotational Region CNN (R2CNN) [Jiang and Luo(2017)] is based on the framework of faster R-CNN, adding two kinds of pooling size to capture the text scene with width larger than height. SCRDet [Yang et al.(2019b)Yang, Yang, Yan, Zhang, Zhang, and Fu] improves R2CNN by adding feature fusion and spatial and channel attention mechanisms. Textboxes++ [Liao et al.(2018)Liao, Shi, and Bai] implements the detection of a rotation box based on SSD. Refined Single-Stage Detector (R3det) [Yang et al.(2019a)Yang, Liu, Yan, and Li] solves the problem that the receptive field of an anchor may not match the position and shape of the object. With the continuous updating of the object detection framework, most of these methods developed from traditional horizontal region detection do not achieve the best results in speed and accuracy. As a recently proposed object detection framework, YOLOv5 is single-stage, but its detection accuracy is higher than most of the object detectors while exhibiting high speed at the same time. Therefore, YOLOv5 is taken as the overall framework in the present work, and on that basis, angular target detection is implemented.

Fig. 3 Definitions of bounding boxes.

Classification for orientation information. To realize arbitrarily oriented object detection, each dataset and detector gives its own definition of rotation boundary. The DOTA dataset shows the coordinates of all four corners of the object. R2CNN uses the first two of the four clockwise corners (x_1y_1, x_2y_2) and the height of the rectangle to define the box. The common method is five-parameter regression, which adds the angle parameter θ in addition to the basic parameters of xy and wh information to realize boundary detection in any direction. The angle range is $0-90^\circ$, shown in Figure 3(a), and the angle θ is the acute angle formed by the width (or height) and x axis. Another method is that the angle ranges from -90° to $+90^\circ$, where θ is the angle between the longest side (w) of the rectangle and the x axis, as shown in Fig 3(b).

3 Proposed method

We first carry out label preprocessing, in which the processing here is mainly for the tag part, and the angle information is extracted from the spatial tag $xywh$ of the target. The data pre-processing can obtain the direction angle from -90° to $+90^\circ$ and re-define the classification of width and height.

Input the picture into the network shown in Figure 4, the network is composed of backbone, neck and prediction. Some modules are omitted in the figure and only the overall structure is introduced. The backbone is used to extract image features, from which three feature maps are extracted successively. Pass these three feature maps through the neck part. In the neck part, a series of operations such as convolution, upsampling, and splicing are performed on the three feature maps, and finally three scale predictions are output. In the predictions, **80**, **40** and **20** represent three scales, **cls** represents the total category of the label, and **5** repre-

Fig. 4 Architecture of the proposed rotation detector. Before training, perform label preprocessing, the pre-processed picture is input to the network, through the backbone and neck in the structure, and finally three-scale predictions are output.

sents the information of the prediction box, including $xywh$ and θ .

3.1 Angle parameter definition

In this study, the five-parameter (x , y , w , h , and θ) regression method is used to predict the rotated bounding box because the ship, car, aircraft, and other objects in remote-sensing images have a fixed aspect ratio, and people intuitively regard the direction parallel to the longer side as the movement direction of the object. In original horizontal area target detection, the definition of the boundary w , h is the side parallel to the coordinate axis x and y . In the rotation detection method, the boundary of the object is no longer simply parallel to the coordinate axis. The theoretical definition of w and h is meaningless, so which side of the object is w and which side is h cannot be clearly indicated. Therefore, to facilitate the implementation of boundary regression, the longer side is defined as w and the short side as h . In this way, the direction parallel to w is the moving direction of the object.

The definition of object direction is determined, and the method to calculate it is then sought. The angle between the long side w and the x axis is the direction of rotation. Considering that the required angle range is $[-90^\circ, 90^\circ]$, arcsin is chosen to calculate the angle. As shown in Fig 5(a), the rotation angle is calculated according to the following formula:

$$\theta = \arcsin [(y_{x(min)} - y_{x(max)})/w] \quad (1)$$

Two points on the longest side w , $y_{x(min)}$ represents the y value of the point with the smaller x , $y_{x(max)}$ is the opposite.

With a clear angle definition, one also must perform the conversion between the five-parameter method and the four-point annotation ($x_1y_1, x_2y_2, x_3y_3, x_4y_4$). The data-enhancement part of YOLOv5 will perform affine and color processing on the image. The data-enhancement part needs to recalculate the four corner points of the target. The final target detection result is to draw a rotating quadrilateral on the original image, and the specific coordinates of the four corner points must be known. As shown in Fig 5(b), the conversion formula is obtained through a simple geometric analysis:

$$F\mathbf{x}_i = \frac{(-1)^{L(Ox_i, Cx)}w \cos \theta - (-1)^{L(Oy_i, Cy)}h \sin \theta}{2} + Cx \quad (2)$$

$$L(O\mathbf{x}, C\mathbf{x}) = \begin{cases} 0 & Ox > Cx \\ 1 & Ox < Cx \end{cases} \quad (3)$$

In which $F\mathbf{x}_i$ represents the final point $\mathbf{x}_i$ after transformation and $L(O\mathbf{x}, C\mathbf{x})$ is used to obtain the relative location of the original corner point $\mathbf{x}$ and center point $\mathbf{x}$, other coordinates are calculated similarly.

3.2 Angle and box regression method

The angle dimension is added to solve the regression problem of the target rotation direction. From the perspective of YOLOv5 network structure, it is mainly divided into the backbone, neck, and head. The backbone extracts the deep features of the image, and the head outputs the prediction according to the number of channels needed. For example, there are 80 detection categories in the proposed process, the target location

(a) Use arcsin function to calculate the angle between the longest side and the x-axis.

(b) Calculate the coordinates of each corner point after rotation through geometric analysis, the original label(green box) is defined by 5 parameters($xywh\theta$), and the final label(red box) is defined by 4 points($x_1y_1, x_2y_2, x_3y_3, x_4y_4$).

Fig. 5 Angle parameter definition and conversion.

uses the four-parameter method (x, y, w, h), and then the final output matrix is $F \times F \times (80+4+1)$. F represents the size of the feature map that is output by the last layer of convolutional layer and '1' represents the probability that a certain pixel on the feature map is the center point of the target; the neck layer is between the above two and provides some useful modules, such as FPN. Therefore, in order to use the five-parameter method for positioning, an additional channel is added to the head layer to predict the angle value.

Before changing the number of channels in the head layer, it is first necessary to build the target, which is a major feature of YOLOv5 bounding regression. In the classification and positioning prediction matrix output by the head, if the object center point xy domain is the entire feature map, then the distance between ground truth and prediction may be too large, resulting in a large loss that is not conducive to convergence of the network. Thus, first the label is reprocessed, then the upper left-hand corner of each grid point is used as the origin to establish the coordinates, and finally xy is converted to offset t_x and t_y , relative to the upper left-hand corner, so that the domain of t_x and t_y is $[0, 1]$, which greatly reduces the loss. However, to increase the positive sample space, YOLOv5 uses one ground truth to establish three positive samples. These three positive samples are also calibrated with the original

Fig. 6 Build three positive samples with one groundtruth.

coordinates, so that the domain of t_x and t_y becomes $[-0.5, 1.5]$, as the shown in Fig 6.

The prediction output by the head cannot be directly calculated for the loss function. First, a logistics regression is required: the purpose of which is to limit the prediction within the range that is set.

$$b_x = \frac{2}{1 + e^{-t_x}} - 0.5 + C_x \quad (4)$$

Eq. (4) represents the logistics calculation. t_x is the network output value, after calculation plus grid point coordinates C_x , the actual position of the center point b_x is obtained, b_y can be gotten in the same way.

$$b_\theta = \frac{3}{1 + e^{-\theta}} - 1.5 \quad (5)$$

The result of Eq. (5) is expressed in radians and the angle θ is limited to the domain $[-1.5, 1.5]$ through logistics calculations.

Different loss functions are tried, with one combining an angle IOU calculation and the other separating an angle loss calculation. In the former method, the angle parameter is used to convert the horizontal rectangle into a rotated rectangle. An attempt was made to use the polygon IOU calculation formula that comes with the library function, but the internal calculation process was not revealed. Use of this method creates a huge amount of calculation. From the results in Fig 7(b), this function is likely to be discontinuous or non-differentiable, leading to non-convergence of the model. The second method is to calculate the area intersection by complete-IOU (CIOU), and the angle loss is calculated by the SmoothL1 function alone. CIOU is an efficient loss function recently proposed, and, as shown in Figure 7(a), it works with the width and height of

Fig. 7 (a)CIOU loss:the green box is groundtruth, the black is prediction, and the gray is the smallest outer rectangle. d denotes the distance between the two center point, and c represents the distance between the smallest outer rectangle.(b)Model convergence of the two loss functions.

the rectangle and the distance between the two center points.

It includes a simple IOU calculation in which one first calculates the intersection and union of the two bounding boxes:

$$IOU = \frac{I(\tilde{B}, B)}{U(\tilde{B}, B)} \quad (6)$$

$\tilde{B}$ denoting ground truth, B denoting prediction box, and then performs the logarithmic operation. The complete loss function is described as follows:

$$L_{cious} = 1 - IOU + \frac{\rho^2(\tilde{b}, b)}{c^2} + \alpha v \quad (7)$$

The last two items are penalties that penalize the center-point distance and aspect ratio, where $\tilde{b}$, b denotes the center point of $\tilde{B}$ and B , ρ calculates the euclidean distance, and c denotes the minimum diagonal distance. The last item considers the aspect ratio:

$$\alpha = \frac{v}{(1 - IOU) + v} \quad (8)$$

and α is the parameter used to make a trade-off.

$$v = \frac{4}{\pi^2} (\arctan \frac{\tilde{w}}{\tilde{h}} - \arctan \frac{w}{h})^2 \quad (9)$$

Here v is used to measure the consistency of the aspect ratio. The above CIOU loss is used to calculate the area intersection ratio of ground truth and prediction, the angle is calculated by separate SmoothL1 loss:

$$L_{SmoothL1} = \begin{cases} 0.5(\tilde{\theta} - \theta)^2 & |\tilde{\theta} - \theta| < 1 \\ |\tilde{\theta} - \theta| - 0.5 & otherwise \end{cases} \quad (10)$$

θ is the prediction and $\tilde{\theta}$ is the ground-truth angle.

The above is the regression and loss function of the two branches of the bounding box and the angle. Back-propagation will gradually reduce the loss to achieve

the expected detection result. The specific implementation effect will be described in detail in the following experimental analysis.

4 Experiments

To show the difference between the proposed method and the previous object detection models with or without anchors, comparative experiments were conducted on different benchmark datasets (DOTA, HRSC2016). The training and testing tasks were carried out on a computer equipped with an Intel (R) core (TM) i-9900-KFCPU@3.60GHz processor, with 32 GB of RAM, a GeForce RTX 1080ti graphical processing unit (11 GB global memory), and running the Ubuntu 18.04 LTS operating system.

As the newly released YOLOv5 shows good results in detection accuracy and speed, it was decided to inherit most of the basic configuration of the model, and set the learning rate and momentum to 0.01 and 0.937, respectively. As a classic deep neural network, Darknet-53 combines the characteristics of ResNet to ensure the super-strong expression of features and avoid the gradient problem caused by an overly deep network, and was used as the backbone network of the experiments described herein. Before training, the image was enlarged and enhanced, which is more conducive to feature extraction and improves the robustness of the training results. In the training phase, the size of the input image was adjusted to 640×640. According to the server conditions, 16 images were used as a batch to train on the GPU. All the experiments involved in the comparison trained 120 epochs.

4.1 Datasets

DOTA v-1.0 [Xia et al.(2018)Xia, Bai, Ding, Zhu, Belongie, and Zhang]: Currently the largest optical remote-sensing-image dataset (image sources are Google Earth and two Chinese satellites, i.e., gf-2 and JL-1). The dataset contains 2,806 remote-sensing images (image sizes range from 800×800 to 4000×4000), with 188,282 instances that are divided into 15 categories: aircraft, boats, piggy bank, baseball infield, tennis court, basketball court, track and field, harbor, bridge, large vehicle, small vehicle, helicopter, soccer field, circular route, and swimming pool. Each instance is labeled as a quadrilateral with arbitrary shape and direction determined by four clockwise points $(x_1y_1, x_2y_2, x_3y_3, x_4y_4)$. Officially, half of the images were used as the training set, one-sixth as the verification set, and one-third as the test set.

Table 1 Four traditional horizontal region target detection and rotated bounding box detection results on DOTA, including each type of detection accuracy (AP) and overall performance (mAP).

Method	Backbone	Size	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
SSD	VGG16	512	39.8	9.1	0.6	13.2	0.3	0.4	1.1	16.2	27.6	9.2	27.2	9.1	3.0	1.1	1.0	10.6
YOLOV2	DarkNet19	416	39.6	20.3	36.6	23.4	8.9	2.1	4.8	44.3	38.4	34.7	16.0	37.6	47.3	25.5	7.5	21.4
RFCN	ResNet101	800	37.8	9.1	0.6	13.2	0.3	0.4	1.1	16.2	27.6	9.2	27.2	9.1	10.5	25.2	18.0	26.8
FR-H	ResNet101	800	47.2	61.0	9.8	51.7	14.9	12.8	6.9	56.3	60.0	57.3	47.8	48.7	8.2	37.3	23.1	32.3
ROD(our method)	DarkNet53	640	67.4	52.5	14.6	27.4	47.2	69.3	41.4	88.2	41.3	23.0	29.3	19.3	52.0	42.0	26.9	42.8

Fig. 8 Rotating object training and detection process in remote sensing environment. The left part is the DOTA dataset, and the right part is the HRSC2016 dataset. In each part, the first column is the input, the second column is the image after data enhancement, the third column is the feature map output by the angle channel of the prediction layer, and the last column is the detection result.

HRSC2016: A public remote-sensing ship dataset with resolutions ranging from 2 to 0.4 m and image sizes ranging from 300×300 to 1500×900 contains three levels of tasks (single class, four categories, and 19 types) and 1,070 images (Google Earth) and 2,976 instances, using the rotating box annotation.

4.2 Ablation experiments

In this part, the impact of the proposed method on the control experiment based on the HRSC2016 datasets is discussed.

Different loss functions, including the proposed SmoothL1 loss with L1 and L2, and CIOU loss with the original IOU loss, were compared against the HRSC2016 datasets. Table 2 lists the comparison results in detail,

including the recall rate and accuracy results. It can be clearly seen that the recommended IOU loss is nearly 3% higher than that of the original IOU under an accuracy of mean average precision (mAP50), and it is nearly 6% higher than that of the original CIOU under the accuracy of mAP75. Regarding the other two angle losses, it can be seen from the comparison results presented in the table that SmoothL1 loss exhibits good improvement compared with both L1 and L2 loss, which proves that the proposed method has obvious advantages.

4.3 Benchmark results

The proposed method focuses on improving the limitations of traditional horizontal region object detection in detecting high-density and high-aspect-ratio ob-

Table 2 Comparison between different losses for oriented bounding box on HRSC2016 dataset. Ciou is used to calculate the bounding box loss, and Smooth L1 is for the angle loss.

Method	Recall	mAP ₅₀	mAP ₇₅
$L_{ciou}+L1$	77.2	94.4	66.1
$L_{ciou}+L2$	77.3	94.6	67.3
$L_{iou}+SmoothL1$	78.5	93.8	63.4
$L_{ciou}+SmoothL1$	80.5	96.5	69.34

jects. Therefore, the proposed method is mainly compared with the traditional horizontal bounding box-based method on the aforementioned three benchmark sets; it is also compared with the previously developed arbitrarily oriented object detection method.

DOTA, as the largest remote-sensing dataset, contains a wealth of target instance types and complex ground conditions. The angles and aspect ratios of different types of objects have their own characteristics. In such a large number of cluttered, multi-directional, and multi-dimensional objects, DOTA is taken as the main comparison dataset. To better understand training, we visualize the training process of the DOTA dataset as shown on the left in Figure 8, we selected 6 categories from the 15 categories of the DOTA data set for visualization, namely large-vehicle, small-vehicle, plane, tennis-court, harbor and ship. As we can see from the figure, after data enhancement, the image is rotated and divided into small pieces into the model, from the third column we can see that the angle semantic information of the object can be better expressed in the angle channel feature map, and finally combined with the angle regression, we can realize the rotating object detection in the remote sensing environment. The specific benchmark experiment results are shown in Table 1. Compared with other horizontal detectors, our mAP is at least about 10% higher than others, especially for plane, large-vehicles tennis-court, with accuracies of 67.4%, 69.3% and 88.2% respectively. We think this is a good improvement.

Table 3 Test results on HRSC2016. The first six kinds of detectors are also based on the rotated bounding box.

Method	Backbone	Size	mAP ₅₀
R ² CNN	ResNet101	800×800	73.03
RC1RC2	VGG16	—	75.5
RRPN	ResNet101	800×800	79.08
R ² PNN	VGG16	—	79.6
RetinaNet-R	ResNet101	800×800	89.18
ROD(our method)	DarkNet53	640×640	96.5

Regarding the comparison using the HRSC2016, the ship instance objects in this dataset are large and distributed as shown on the right of Figure 8. Because

only the ship category is available in this dataset, and the distribution is relatively scattered, the difficulty of training is easier than that of DOTA, the angular channel output can also be used to produce clearer feature maps. We compare our method with other state-of-the-art rotation detectors, it can be seen that the proposed method exhibits excellent performance in scene detection facing any direction, as shown in Table 3. Compare to other rotation detectors, our model achieves an excellent performance of approximately 96.5%.

Fig. 9 Samples results using YOLOV5 (green boxes) and rotated bounding detector (red boxes) on DOTA (first row), and joining RetinaNet-R (yellow boxes) to compare on HRSC2016 (second row).

The visual results using the DOTA and HRSC2016 dataset are shown in Figure 9. It can be observed that, compared with the traditional object detection based on a horizontal bounding box, rotated bounding box detection has excellent performance in the remote-sensing environment, whether for small objects like vehicles or large objects like ships and seaports, which have complex and confusing directions. The proposed method has been greatly improved on the DOTA datasets. From the first row of DOTA detection results in the figure, the rotational detection (red boxes) can more clearly distinguish the position and direction of each objects in the face of the gathered vehicles. For objects with large aspect ratios such as tennis-court and harbor, our method also eliminates the background and wraps the target more closely. In HRSC2016 (second row), our method (red boxes) is not only compared with traditional horizontal detectors (green boxes), but also has good results compared with other rotating detectors (yellow boxes).

5 Conclusion

In this study, a model of rotated bounding box detection based on YOLOv5 was proposed. This model

was introduced to solve the problem of traditional horizontal bounding box detectors that detect targets with high density, high aspect ratio, and overlapping detection frames with difficulty. On the basis of the original model of YOLOv5, the rotation angle channel and corresponding angle loss function were added. To achieve the training effect, the pre-processing of data labels was set up to define and calculate the width, height, and angle of the objects. To prove the effectiveness of the model, two public remote-sensing datasets were selected to conduct comparative experiments and evaluate the performance of the model against both a traditional horizontal bounding box detector and a nearest rotated bounding box detector. Experimental data and visual analysis showed that the model based on YOLOv5 is an effective choice for multi-directional remote-sensing image detection.

Acknowledgements This study was supported by the National Nature Science Foundation of China (Grant No.61906168,61876168); Quzhou Science and Technology Projects (Grant No.2020K19); Zhejiang Provincial Natural Science Foundation of China (Grant No.LY20F020029).

Author Contributions Sixian Chan:concept propose, funding support, draft review and algorithm design. Jingcheng Zheng:coding, experimental evaluation and manuscript draft writing. Lina Wan and Tingting Wang: experiments and analysis. Xiaolong Zhou:funding support and draft review. Yinkun Xu:funding support. Kai Fang:literature review.

Declarations

Conflict of interest All authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

- [Ahmed et al.(2017)Ahmed, Dogra, Kar, and Roy] Ahmed SA, Dogra DP, Kar S, Roy PP (2017) Unsupervised classification of erroneous video object trajectories. *Soft Computing*
- [Araújo et al.(2020)Araújo, Fontinele, and Oliveira] Araújo P, Fontinele J, Oliveira L (2020) Multi-perspective object detection for remote criminal analysis using drones. *IEEE Geoscience and Remote Sensing Letters* 17(7):1283–1286, DOI 10.1109/LGRS.2019.2940546
- [Girshick(2015)] Girshick R (2015) Fast r-cnn. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp 1440–1448
- [Haghofer et al.(2020)Haghofer, Dorl, and Oszwald] Haghofer A, Dorl S, Oszwald SM (2020) Evolutionary optimization of image processing for cell detection in microscopy images. *Soft Computing* 24(23):17847–17862
- [Janakiramaiah et al.(2021)Janakiramaiah, Prasad, and Krishna] Janakiramaiah A, Prasad LN, Krishna M (2021) Military object detection in defense using multi-level capsule networks. *Soft Computing* pp 1–15
- [Jiang and Luo(2017)] Jiang Y, Luo Z (2017) R2CNN: rotational region CNN for orientation robust scene text detection. *CoRR* abs/1706.09579, URL <http://arxiv.org/abs/1706.09579>, 1706.09579
- [Law and Deng(2020)] Law H, Deng J (2020) Cornernet: Detecting objects as paired keypoints. *International Journal of Computer Vision* 128(3):642–656
- [Liao et al.(2018)Liao, Shi, and Bai] Liao M, Shi B, Bai X (2018) Textboxes++: A single-shot oriented scene text detector. *IEEE Transactions on Image Processing* 27(8):3676–3690
- [Lin et al.(2017)Lin, Goyal, Girshick, He, and Dollár] Lin TY, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis & Machine Intelligence* PP(99):2999–3007
- [Liu et al.(2016)Liu, Anguelov, Erhan, Y., and Berg] Liu W, Anguelov D, Erhan D, Y SC, Berg AC (2016) Ssd: Single shot multibox detector. *European Conference on Computer Vision* pp 21–37
- [Ma et al.(2018)Ma, Shao, Ye, Wang, Wang, Zheng, and Xue] Ma J, Shao W, Ye H, Wang L, Wang H, Zheng Y, Xue X (2018) Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia* 20(11):3111–3122, DOI 10.1109/TMM.2018.2818020
- [Mahalingam and M(2019)] Mahalingam T, M (2019) Cbdf: a refined w4+ cluster-based frame difference approach for efficient moving object detection. *Soft Computing* 23(21):10661–10679
- [Mahalingam and Subramoniam(2020)] Mahalingam T, Subramoniam M (2020) Optimal object detection and tracking in occluded video using dnn and gravitational search algorithm. *Soft Computing* 24(24):18301–18320
- [Redmon et al.(2016)Redmon, Divvala, Girshick, and Farhadi] Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 779–788
- [Ren et al.(2017)Ren, He, Girshick, and Sun] Ren S, He K, Girshick R, Sun J (2017) Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 39(6):1137–1149
- [Sudha and Priyadarshini(2020)] Sudha D, Priyadarshini J (2020) An intelligent multiple vehicle detection and tracking using modified vibe algorithm and deep learning algorithm. *Soft Computing* 24(21):1–13
- [Sun et al.(2019)Sun, Yan, Huang, and Sun] Sun W, Yan D, Huang J, Sun C (2019) Small-scale moving target detection in aerial image by deep inverse reinforcement learning. *Soft Computing* 24(11):1–12
- [Wu et al.(2020)Wu, Zhang, Wang, Wang, Wang, and Li] Wu Y, Zhang K, Wang J, Wang Y, Wang Q, Li Q (2020) Cdd-net: A context-driven detection network for multiclass object detection. *IEEE Geoscience and Remote Sensing Letters* pp 1–5, DOI 10.1109/LGRS.2020.3042465
- [Xia et al.(2018)Xia, Bai, Ding, Zhu, Belongie, and Zhang] Xia GS, Bai X, Ding J, Zhu Z, Belongie M, Zhang L

- (2018) Dots: A large-scale dataset for object detection in aerial images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 3974–3983
- [Yang et al.(2019a)Yang, Liu, Yan, and Li] Yang X, Liu Q, Yan J, Li A (2019a) R3det: Refined single-stage detector with feature refinement for rotating object. CoRR abs/1908.05612, URL <http://arxiv.org/abs/1908.05612>, 1908.05612
- [Yang et al.(2019b)Yang, Yang, Yan, Zhang, Zhang, and Fu] Yang X, Yang J, Yan J, Zhang Y, Zhang X, Fu K (2019b) Srdet: Towards more robust detection for small, cluttered and rotated objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 8232–8241
- [Zhang et al.(2020)Zhang, Mu, Kou, and Zhao] Zhang S, Mu X, Kou G, Zhao J (2020) Object detection based on efficient multiscale auto-inference in remote sensing images. IEEE Geoscience and Remote Sensing Letters pp 1–5, DOI 10.1109/LGRS.2020.3004061
- [Zhou et al.(2017)Zhou, Yao, Wen, Wang, Zhou, He, and Liang] Zhou X, Yao C, Wen H, Wang Y, Zhou S, He W, Liang J (2017) East: An efficient and accurate scene text detector. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp 5551–5560
- [Zhou et al.(2019)Zhou, Wang, and Krähenbühl] Zhou X, Wang D, Krähenbühl P (2019) Objects as points. CoRR abs/1904.07850, URL <http://arxiv.org/abs/1904.07850>, 1904.07850