Rotating object detection in remote-sensing environment

Sixian Chan · Jingcheng Zheng · Lina Wang · Tingting Wang · Xiaolong Zhou · Yinkun Xu · Kai Fang

Accepted: 28 November 2021 / Published online: 15 April 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract

Deep learning models have become the mainstream algorithm for processing computer vision tasks. In the tasks of object detection, the detection box is usually set as a rectangular box aligned with the coordinate axis, so as to achieve the complete packaging of the object. However, when facing some objects with large aspect ratio and angle, the bounding box must be enlarged, which makes the bounding box contain a large amount of useless background information. In this study, a different approach based on YOLOv5 is adopted. By this means, the angle information dimension is added at the head, and angle regression is also added at the same time of the boundary regression. Then the loss of the boundary box is calculated by combining ciou and smoothl1, so that the obtained boundary box is more closely suitable for the actual object. 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 and other state-of-the-art arbitrarily oriented 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, action recognition (Chen et al. 2021), and has a wide range of applications in automatic driving, unmanned aerial vehicles (Zhang et al. 2021), monitoring, and other fields (Haghofer et al. 2020; Janakiramaiah et al. 2021; Zhang et al. 2015a; Lin et al. 2019). 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; Zhang et al. 2015). In the process of object recognition and location, detectors based on deep learning always use bounding box parallel to the coordinate axes to represent the object, identify object with a rectangular shape, and then classify and distinguish actual object or background in the rectangular box. (Sudha and Priyadarshini 2020; Ahmed et al. 2018). 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). However, in some special cases, especially in the case of remote-sensing images (Araújo et al. 2020), when the camera observes some objects with large aspect ratio and disordered direction, from an observation angle perpendicular to the ground, such as ships, vehicles, and seaports, it is impossible to accurately surround the object with a rectangular frame only (Zhang et al. 2020b). Excessive width or height will enlarge the bounding box to an unnecessary extent. 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). 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 \((x_1,y_1, x_2,y_2, x_3,y_3, x_4,y_4)\) or the width and height of the object are not clearly defined. Therefore, in order to obtain the desired effect after training, the data are reprocessed before training to calculate the true width and height of the object and the deflection angle relative to the coordinate axis. To sum up, 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 aimed at solving the problem of object detection with rotation.

2) A method of reprocessing the existing data sets is proposed, which provides a real and effective bounding box for training.

3) The results show that the proposed method is better than the traditional horizontal region object detection method and arbitrarily oriented object detection 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 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 (Sea 2017), pays attention to the accuracy and reduces the amount of calculation to improve the detection speed. Feature pyramid network (FPN) method was proposed to deal with the scale change of the object in the image. Single Shot MultiBox Detector (SSD) (Liu et al. 2016), Yolo (Redmon et al. 2016), and RetinaNet (Lin et al. 2017) represent single-stage detection methods, and their single-stage structure endows them with faster detection speed. 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), 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 in aerial images, as shown in Fig. 2.
Arbitrarily oriented object detection. In recent years, object detection frameworks based on rotation quadrilateral or other polygonal bounding boxes have become very popular. For example, Rotation Region Proposal Network (RRPN) (Ma et al. 2018) 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) 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) improves R2CNN by adding feature fusion and spatial and channel attention mechanisms. Textboxes++ (Liao et al. 2018) implements the detection of a rotation box based on SSD. Refined Single-Stage Detector (R3det) (Yang et al. 2019a) 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.

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 \((x_1, y_1), (x_2, y_2)\) and the height of the rectangle to define the box. The common method is five-parameter regression, which adds the angle parameter \(\theta\) in addition to the basic parameters of \(xy\) and \(wh\) to realize boundary detection in any direction. The angle range is \(0 - 90^\circ\), as shown in Fig. 3a, and the angle \(\theta\) is the acute angle formed by the width (or height) and \(x\) axis. Another method is that the angle ranges from \(-90^\circ\) to \(+90^\circ\), where \(\theta\) is the angle between the longest side \((w)\) of the rectangle and the \(x\) axis, as shown in Fig. 3b.

3 Proposed method

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

Input the picture into the network shown in Fig. 4, the network is composed of backbone, neck and prediction. Some modules are omitted from the figure and only the overall structure is introduced. The backbone is used to extract image features. Plotting these three features through the neck section. In the neck part, a series of operations such as convolution, upsampling, and concatenating are performed on the three feature maps. In the predictions, 80, 40 and 20 represent three scales, \(cls\) represents the total category of the label, and 5 represents the information of the prediction box, including \(xywh\) and \(\theta\).

3.1 Angle parameter definition

In this study, the five-parameter \((x, y, w, h, \text{and } \theta)\) 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 object detection, the definition of the boundaries \(w, h\) 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 expressed. Therefore, to facilitate the implementation of boundary regression, the long 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.
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, the predictions are output at the head part for three scales.

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. 5a, the rotation angle is calculated according to the following formula:

$$\theta = \arcsin[(y_{x(min)} - y_{x(max)})/w]$$  \hspace{1cm} (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 definition of the angle, it is absolutely necessary to perform the conversion between the five-parameter method and the four-point annotation. 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 object 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. 5b, the conversion formula is obtained through a geometric analysis:

$$F_{xi} = \frac{(-1)^{L(Ox, Cx)}w \cos \theta - (-1)^{L(Oy, Cy)}h \sin \theta}{2} + Cx$$  \hspace{1cm} (2)

$$L(Ox, Cx) = \begin{cases} 0 & O_{x} > C_{x} \\ 1 & O_{x} < C_{x} \end{cases}$$  \hspace{1cm} (3)

(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($x_1y_1, x_2y_2, x_3y_3, x_4y_4$).
In which $F_x$ represents the transformed point $x_i$ and $L(x_0, C_x)$ is used to obtain the relative location of the original corner point $x$ and center point $x$, 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 required number of channels. For example, there are 80 detection categories in the proposed process, the target location 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 $L_i$ 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, firstly the label is reprocessed, then the upper left corner of each grid point is used as the origin to establish the coordinates, and finally $xy$ is converted to offset $tx$ and $ty$, relative to the upper left corner, so that the domain of $tw$ and $ty$ 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 coordinates, so that the domain of $tx$ and $ty$ becomes $[-0.5, 1.5]$, as 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.

$$b_x = \frac{2}{1 + e^{-tx}} - 0.5 + C_x \tag{4}$$

Eq. (4) represents the logistics calculation. $tx$ is the output value of the network, 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.

![Fig. 6 Build three positive samples with one groundtruth](image)

The result of Eq. (5) is expressed in radians and the angle $\theta$ 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. 7b, this function is likely to be discontinuous or non-differentiable, leading to non-convergence of the model. The second method is use complete-IoU (CIoU) to calculate the area intersection, and the angle loss is calculated by the SmoothL1 function alone. CIoU is an efficient loss function recently proposed.

![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](image)
and, as shown in Fig. 7a, it works with the width and height of 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:

\[
IoU = \frac{I(\tilde{B}, B)}{U(\tilde{B}, B)}
\]

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

\[
L_{ciou} = 1 - IoU + \frac{\rho^2(\tilde{b}, b)}{c^2} + \alpha v
\]  

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 \(B\), \(\rho\) calculates the Euclidean distance, and \(c\) denotes the minimum diagonal distance. The last item considers the aspect ratio:

\[
\alpha = \frac{v}{(1 - IoU) + v}
\]  

and \(\alpha\) is the parameter used to make a trade-off.

\[
v = \frac{4}{\pi^2} \left( \arctan \frac{\tilde{w}}{\tilde{h}} - \arctan \frac{w}{h} \right)^2
\]

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:

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

\(\theta\) 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. Backpropagation 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 2080ti 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 640640. According to the server conditions, 16 images were used as a batch to train on the GPU. All the comparative experiments were trained for 120 periods.

4.1 Datasets

DOTA v-1.0 (Xia et al. 2018): 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: plane (PL), ship (SH), storage tank (ST), baseball diamond (BD), tennis court (TC), basketball court (BC), ground track field (GTF), harbor (HA), bridge (BR), large vehicle (LV), small vehicle (SV), helicopter (HC), roundabout (RA), soccer ball field (SBF) and swimming pool (SP). 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.

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
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.

The proposed method focuses on improving the limitations of traditional horizontal region object detection in detecting high-density and high-aspect-ratio objects. Therefore, the proposed method is mainly compared with the traditional horizontal bounding box-based method on the aforementioned two 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. Different types of objects have their own characteristics in terms of angle and aspect ratio. Among so many messy, 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 Fig. 8, we selected 6 categories from the 15 categories of the DOTA dataset 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 and 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.

Regarding the comparison using the HRSC2016, the ship instance objects in this dataset are large and distributed as shown on the right of Fig. 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, so
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 | 800  | 66.5 | 52.7 | 13.1 | 27.0 | 45.7 | 68.8 | 41.0 | 88.5 | 40.5 | 45.1 | 22.6 | 28.2 | 21.1 | 51.5 | 41.6 | 25.0 | 42.2 |

Table 2. Comparison between different losses for oriented bounding box on HRSC2016 dataset

| Method       | Recall | mAP50 | mAP75 |
|--------------|--------|-------|-------|
| L\text{ciou}+L1 | 77.2   | 94.4  | 66.1  |
| L\text{ciou}+L2 | 77.3   | 94.6  | 67.3  |
| L\text{ciou}+SmoothL1 | 78.5   | 93.8  | 63.4  |
| L\text{ciou}+SmoothL1 | 80.5   | 96.5  | 69.34 |

Ciou is used to calculate the bounding box loss, and Smooth L1 is for the angle loss.

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

| Method     | Backbone | Size    | mAP50 |
|------------|----------|---------|-------|
| \text{R}^2\text{CNN} | ResNet101 | 800 × 800 | 73.03 |
| \text{RC1&RC2} | VGG16    | –       | 75.5  |
| \text{RRPN}  | ResNet101 | 800 × 800 | 79.08 |
| \text{R}^2\text{PNN} | VGG16    | –       | 79.6  |
| \text{RetinaNet-R} | ResNet101 | 800 × 800 | 89.18 |
| \text{RoI-Transformer} | ResNet101 | 512 × 800 | 86.20 |
| \text{R}^2\text{Det} | ResNet101 | 800 × 800 | 89.26 |
| \text{RSDet}   | ResNet101 | 800 × 800 | 86.5  |
| ROD (our method) | DarkNet53 | 800 × 800 | 96.2  |

The visual results using the DOTA and HRSC2016 dataset are shown in Fig. 9. It can be observed that: the rotated bounding box detection has excellent performance in remote-sensing compared with the traditional target detection based on horizontal bounding box. 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.6190 6168, 61876168); Quzhou Science and Technology Projects (Grant No.2020K19); Zhejiang Provincial Natural Science Foundation of China (Grant No. LY20F020029).

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

Funding The authors have not disclosed any funding.

Data availability Enquiries about data availability should be directed to the authors.

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 SA, Dogra DP, Kar S, Roy PP (2018) Unsupervised classification of erroneous video object trajectories. Soft Comput 22(14):4703–4721

Araijo P, Fontainele J, Oliveira L (2020) Multi-perspective object detection for remote criminal analysis using drones. IEEE Geosci Remote Sens Lett 17(7):1283–1286

Chen ZM, Jin X, Zhao BR, Zhang X, Guo Y (2021) Hcc: Hierarchical context embedding for region-based object detection. IEEE Trans Image Process 30:6917–6929

Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp 1440–1448

Haghofer A, Döfl S, Oszwald SM (2020) Evolutionary optimization of image processing for cell detection in microscopy images. Soft Comput 24(23):17847–17862

Janakiramaiah A, Prasad LN, Krishna M (2021) Military object detection in defense using multi-level capsule networks. Soft Computing pp 1–15

Jiang Y, Luo Z (2017) R2CNN: rotational region CNN for orientation robust scene text detection. CoRR abs/1706.09579, arXiv:1706.09579

Law H, Deng J (2020) CornerNet: detecting objects as paired keypoints. Int J Comput Vision 128(3):642–656

Liao M, Shi B, Bai X (2018) Textboxes++: a single-shot oriented scene text detector. IEEE Trans Image Process 27(8):3676–3690

Lin H, Zhang T, Chen Z, Song H, Yang C (2019) Adaptive fuzzy gaussian mixture models for shape approximation in robot grasping. Int J Fuzzy Syst 21(4):1026–1037

Lin TY, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. IEEE Tran Pattern Anal Mach Intell 99:2999–3007

Liu W, Anguelov D, Erhan D, YSC, Berg AC (2016) Ssd: Single shot multibox detector. European Conference on Computer Vision pp 21–37

Ma J, Shao W, Ye H, Wang L, Wang H, Zheng Y, Xue X (2018) Arbitrary-oriented scene text detection via rotation proposals. IEEE Trans Multimedia 20(11):3111–3122. https://doi.org/10.1109/TMM.2018.2818020

Mahalingam TM (2019) Cbld: a refined w++ cluster-based frame difference approach for efficient moving object detection. Soft Comput 23(21):10661–10679

Mahalingam T, Subramoniam M (2020) Optimal object detection and tracking in occluded video using dnn and gravitational search algorithm. Soft Comput 24(24):18301–18320

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 S (2017) Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149

Sudha D, Priyadarshini J (2020) An intelligent multiple vehicle detection and tracking using modified vibe algorithm and deep learning algorithm. Soft Comput 24(21):1–13

Sun W, Yan D, Huang J, Sun C (2019) Small-scale moving target detection in aerial image by deep inverse reinforcement learning. Soft Comput 24(11):1–12

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 Geosci Remote Sensing Lett. https://doi.org/10.1109/LGRS.2020.3042465

Xia GS, Bai X, Ding J, Zhu Z, Belongie M, Zhang L (2018) Dopa 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 X, Liu Q, Yan J, Li A (2019a) R3det Refined single-stage detector with feature refinement for rotating object. CoRR abs/1908.05612, arxiv:1908.05612

Yang X, Yang J, Yan J, Zhang Y, Zhang X, Fu K (2019b) Scrdet 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 J, Li M, Feng Y, Yang C (2020) Robotic grasp detection based on image processing and random forest. Multimed Tools Appl 79(3):2427

Zhang S, Mu X, Kou G, Zhao J (2020) Object detection based on efficient multiscale auto-inference in remote sensing images. IEEE Geosci Remote Sensing Lett. https://doi.org/10.1109/LGRS.2020.300461

Zhang X, Hu W, Xie N, Bao H, Maybank S (2015) A robust tracking system for low frame rate video. Int J Comput Vision 115(3):279–304

Zhang X, Fan M, Wang D, Zhou P, Tao D (2021) Top-k feature selection framework using robust 0–1 integer programming. IEEE Trans Neural Netw Learn Syst 32(7):3005–3019

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 X, Wang D, Krähenbühl P (2019) Objects as points. CoRR arXiv:1904.07850

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.