Oriented object detection in aerial images based on area ratio of parallelogram

Xinyi Yu*, Mi Lin, Jiangping Lu, and Linlin Ou*
Zhejiang University of Technology, College of Information Engineering, Hangzhou, China

Abstract. Oriented object detection is a challenging task in aerial images since the objects in aerial images are displayed in arbitrary directions and are frequently densely packed. The mainstream detectors describe rotating objects using a five-parameter or eight-parameter representations, which have representation ambiguity for orientated object definition and border loss discontinuity in the regression process. We proposed an innovative representation method based on area ratio of parallelogram, called area ratio of parallelogram (ARP). Specifically, ARP regresses the minimum bounding rectangle of the oriented object and three area ratios. Three area ratios include the area ratio of a directed object to the smallest circumscribed rectangle and two parallelograms to the minimum circumscribed rectangle. It simplifies offset learning and eliminates the issue of angular periodicity or label point sequences for oriented objects. To further remedy the confusion issue of nearly horizontal objects, the area ratio between the object and its minimal circumscribed rectangle is employed to guide the selection of horizontal or oriented detection for each object. Moreover, the rotated efficient intersection over union loss with horizontal bounding box and three area ratios are designed to optimize the bounding box regression for rotating objects. Experimental results on remote sensing datasets, including HRSC2016, DOTA, and UCAS-AOD, show that the proposed method achieves superior detection performance than many state-of-the-art approaches. © 2022 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: 10.1117/1.JRS.16.034510]

Keywords: orientated object detection; aerial images; representation method; area ratio of parallelogram; rotated efficient intersection over union loss; remote sensing.

1 Introduction

Object detection is a vital computer vision technique that aims at classifying and locating the object in images. With the rapid development of satellite remote sensing images and the explosive growth of remote sensing data, object detection in aerial images has been a specific yet activated topic in computer vision. Among other applications, it is widely used in military surveillance, crop monitoring, traffic planning. Most traditional methods that rely on manual characteristics to identify objects are sensitive to environmental disruptions, such as clouds, sunlight, and rain. Object detection has advanced dramatically in recent years because of deep convolutional neural networks (CNN). Compared with traditional object detection methods, object detection algorithms based on deep learning have the advantages of high accuracy and strong robustness. They are frequently utilized for object detection in remote sensing images.

Object detection can generally be divided into horizontal detection and rotated detection based on the direction of detected boxes. Object detection in aerial images is challenging since the objects have various scales, aspect ratios, and orientations. Consequently, the horizontal object detector for aerial images may lead to misalignment between the detected bounding boxes and ground truth bounding boxes, as shown in Fig. 1(a). The horizontal object detector regresses the horizontal bounding box with four parameters: the abscissa and ordinates of the central point and the length and width of the bounding box. Compared with the horizontal object detector, the rotating object detector adds a regression angle with a total of five parameters.
parameters. As shown in Fig. 1(b), the rotating detector regresses the smallest circumscribed rectangle of objects, successfully resolving the problem of border misalignment and minimizing the overlapping area of dense objects. Therefore, the oriented object detector is preferred for capturing objects in aerial images.

Currently, oriented object detectors characterize oriented objects using five-parameter or eight-parameter regression techniques, which introduce representation ambiguity and border discontinuity. Five-parameter regression methods adopt \((x, y, w, h, \theta)\) to represent the rotating objects, which causes periodicity of angle and exchangeability of edges. The predicted angle may be beyond the defined range. Therefore, the model needs to regress the angle with the complicated form at the boundary, which increases the difficulty of predicting the angle and affects the performance of the model. This issue is especially pronounced for oriented objects with wide aspect ratios, such as harbors and bridges in aerial images, as well as long text lines in scene images. Eight-parameter regression methods typically utilize four vertices to represent the rotating objects, which creates ambiguity when defining the bounding box order of the four vertices. Theoretically, there are 24 different representations for a single bounding box, so they also face the problem of border discontinuity. As a result, whether employing five-parameter or eight-parameter regression, an object may be represented in various ways. They complicate the regression process and degrade detection performance further.

Several approaches have been proposed to redefine representation for oriented objects, such as P-RSDet, ProjBB, BBAVectors, and gliding vertex. These methods describe rotated boxes by altering the coordinate system, utilizing vector representations, and performing vertex offsets. However, they do not eliminate the issue of angular periodicity or vertex ordering, and the underlying boundary problem has not been totally solved. SCL and DCL are novel approaches to solve boundary issues. The prediction of the object angle is transformed into a classification problem to limit the forecast result to a specified range. Nevertheless, they lost some accuracy and increased the amount of calculation. GWD and KLD handle the discontinuity of boundary loss problem by modifying the loss function of the oriented bounding box, but there are angular offsets in the detection of square-shaped objects. In conclusion, the problem of border loss discontinuity in the detection of rotating targets has not yet been resolved and is currently the subject of ongoing exploration and investigation.

To solve the non-unique representation of rotating objects, this article suggests a straightforward yet effective representation called area ratio of parallelogram (ARP). It uses central point coordinates, width and height of the minimal bounding rectangle of the oriented object, and introduces three area ratios to represent a rotating object. Three area ratios are as follows: one is the area ratio of the rotating object to its smallest boundary rectangle; another two are the area ratios of two parallelograms to the minimum bounding rectangle of the rotating object (explained in Sec. 3.2). The area ratio can be used to determine the aspect ratio of the corresponding side at the intersection of the oriented object and the horizontal bounding box. It facilitates offset learning and is superior to forecast five-parameter or eight-parameter bounding boxes directly. Continuously, to further eliminate the issue of almost horizontal objects being confused, the

Fig. 1 Illustration of horizontal object detector and oriented object detector. (a) Horizontal object detector and (b) oriented object detector.
area ratio between the object and its minimum circumscribed rectangle is employed to determine whether to detect the object horizontally or oriented. Lastly, a rotation efficient loss intersection over union loss (R-EIoU) is suggested to handle the undifferentiable problem and enhance the accuracy of the rotating bounding box. In summary, the main contributions of this paper are:

1. An oriented object representation method (ARP) is proposed based on area ratio of parallelogram. It avoids discontinuous boundary loss in regression process by converting the area ratio into the aspect ratio of the corresponding side at the intersection of the oriented object and the horizontal bounding box. Compared to the baseline methods that learn five-parameter or eight-parameter, it achieves better performance. The area ratio between the object and its minimum circumscribed rectangle is used to determine whether to detect the object horizontally or oriented, resolving the problem that horizontal bounding box could not be represented.

2. As an alternative to the smooth $L_1$ loss, we propose a novel loss function called R-EIoU loss. It successfully solves the bias problem in near horizontal object detection and achieves higher accuracy in rotated detection when compared to the smooth $L_1$ loss.

3. Experiments show that our method can be easily embedded in different backbones and significantly improve the performance in oriented object detection. It achieves superior performance on four public datasets and three popular detectors.

## 2 Related Work

### 2.1 Horizontal Object Detection

In recent years, object detection has advanced significantly and can be roughly classified into two categories: two-stage and one-stage detectors. The two-stage detectors initially create several areas of interest [region of interests (RoIs)] and then utilize the characteristics of those areas to anticipate object categories and regress the bounding boxes. Numerous high-performance two-stage detectors have been developed, such as region-based convolutional neural network (R-CNN),\textsuperscript{6} Fast R-CNN,\textsuperscript{21} Faster R-CNN,\textsuperscript{7} and Mask R-CNN.\textsuperscript{22} Although two-stage approaches have shown promising results on some benchmarks, one-stage detectors simplify the detection process by treating it as a regression problem and achieving higher speed. YOLO\textsuperscript{8} and its variances (YOLO9000,\textsuperscript{23} YOLOv3,\textsuperscript{24} YOLOv4,\textsuperscript{25} and YOLOv5), SSD,\textsuperscript{26} and RetinaNet\textsuperscript{9} are representative single-stage approaches. Compared with anchor-based methods, various anchor-free techniques, such as Cornernet\textsuperscript{27} and CenterNet,\textsuperscript{28} have exploded in popularity in recent years, opening up a new direction for object detection.

Loss function plays a critical role in object detection. It has a significant impact on object detection performance and continues to improve with the advancement of object detectors. Since Smooth $L_1$ loss\textsuperscript{21} is robust and resistant to outliers, it is widely utilized in many object detectors, such as Fast R-CNN and Faster R-CNN. This approach assumes the four points of the bounding box are independent, whereas they are connected. It has been demonstrated that the use of IoU-induced loss\textsuperscript{29} such as GloU,\textsuperscript{30} DIoU,\textsuperscript{31} and Focal-EIoU,\textsuperscript{32} can assure the consistency of the final detection metric and loss in conventional horizontal detectors.

Prior to the appearance of datasets for remote sensing image object detection such as DOTA, remote sensing image object detection has mostly relied on horizontal object detection, such as ORSIm\textsuperscript{33} and FRIFB.\textsuperscript{34} However, these detectors generate bounding boxes solely in the horizontal direction, which frequently mismatch the objects in aerial images due to the arbitrarily rotated and densely distributed objects in remote sensing images. As a result, the oriented object detection has become a prominent direction in aerial images, and the identical is true for text detection in natural scenes.

### 2.2 Rotated Object Detection

Generally, rotated object detectors use oriented bounding boxes (OBBs) to describe the positions of objects, which eliminate the issue of misalignment and are excellent for capturing objects in aerial images and scene text. Numerous effective detectors for scene text detection have been
proposed, including R2CNN, RRPN, TextBoxes++, RRD, and EAST. These approaches are generally extended from horizontal object detectors by including an angle and some measures to adapt the detection of objects in scene text.

However, object detection is more challenging in aerial images since the objects are frequently small, numerous, and scattered with large-scale changes. It is hard to demonstrate promising performance in remote sensing images using the methods described already. Hence, many robust rotated object detectors have emerged in aerial images. For example, RoI transformer suggests RRoI leaner and RRoI warping. RRoI leaner transforms horizontal RoIs into rotated RoIs, and RRoI wrapping extracts the rotation-invariant feature from the oriented RoI for subsequent object classification and location regression. R3Det is a one-stage rotation detector based on RetinaNet that solves the feature misalignment problem by employing cascade regression and refining the bounding box. CFC-Net designs a polarized attention module, refines the preset horizontal anchor points, and adopts dynamic anchor learning strategy to achieve high performance in rotating object detection.

While the preceding solutions are promising, they overlook the issue of the discontinuous border. They lead to discontinuous regression loss and impede the convergence of network training. Thus, multiple rotational object detectors have been developed, each of which attempts to overcome the obstacles mentioned already in a unique way. For instance, gliding vertex glides the vertex of the horizontal bounding box to represent a rotatable object, alleviating the issue of border discontinuity and achieving more precise object detection. BBAVectors presents a box boundary-aware bounding box description. BBAVectors is distributed in the four quadrants of the Cartesian coordinate system, which is superior to directly forecasting the spatial characteristics of bounding boxes. P-RSDet achieves arbitrary-oriented object detection by predicting the center point and regressing one polar radius and two polar angles. ProjBB proposes a novel projection-based method that employs six-parameter to describe rotated bounding box: two points position on the projected line (one center point and one chosen vertex), a projection ratio and a quadrant label. Although the representation is unique, there are still issues with square-like detection. CSL and DCL regard angular prediction as a classification task to avoid the discontinuity of regression loss. SCRDet and RSDet propose IoU-smooth L1 loss and modulated loss to alleviate the sharp increase loss caused by the angular periodicity, allowing for improved handling of small, clutter, and rotated objects. Representation invariance loss (RIL) optimizes the bounding box regression for the rotating objects through RIL, which solves the inconsistency between the loss and localization quality caused by ambiguous representations. Gaussian Wasserstein distance (GWD) introduces the GWD loss to approximate the non-differentiable rotation IoU-induced loss, which can elegantly resolve border discontinuity and square-like problems.

In summary, all of the solutions already mentioned turn definitions of rotating objects or create more precise loss functions to handle the border discontinuity problem. Nonetheless, there is no fully unified answer to all of the issues previously noted. To increase the performance and speed in aerial images and scene text, the ARP method is proposed to accurately describe a rotating object and create an R-EIoU loss to approximate the non-differentiable rotation IoU.

3 Proposed Method

This section will elaborate on the suggested method in the sequence mentioned as follows. First, the general architecture of our technique is shown in Fig. 2. A single-stage rotation detector based on You Only Look Once v5 (YOLOv5) is utilized in this embodiment. Then, ARP is designed to describe rotating objects accurately, which has a unique representation. To be more precise, the directed object is regressed using its coordinates of the central point, the width and height of its smallest bounding rectangle, and three area ratios. It is easier to learn the orientation and scale information of the object, effectively solving the problem of boundary discontinuity. Additionally, the area ratio between the rotating object and its minimum circumscribed rectangle is also deployed as an obliquity factor for selecting the oriented or horizontal bounding boxes. At last, to more correctly express loss in oriented object identification, the R-EIoU loss is designed.
3.1 Network Architecture

YOLOv5 is one of the most advanced single-stage object detectors. We extend YOLOv5 to the oriented object detection to prove the effectiveness of ARP. Its network architecture is sketched in Fig. 2. Append three additional object variables \((\lambda_1, \lambda_2, \lambda_3)\) to the head of YOLOv5. The symbols \(H\) and \(W\) refer to the length and width of the grid, respectively, and \(B\) is the batch size of the sample. \(C\) and \(A\) represent the number of categories and predefined anchors, respectively, while \(8\) denotes the confidence and ARP representation method \((x, y, w, h, \lambda_1, \lambda_2, \lambda_3)\).

![Network architecture of proposed rotation detector (YOLOv5 as an embodiment). CSPDarknet53, FPN+PAN, and YOLO HEAD are the primary components of the proposed detector. We add three extra object variables \((\lambda_1, \lambda_2, \lambda_3)\) to the head of YOLOv5. The symbols \(H\) and \(W\) refer to the length and width of the grid, respectively, and \(B\) is the batch size of the sample. \(C\) and \(A\) represent the number of categories and predefined anchors, respectively, while \(8\) denotes the confidence and ARP representation method \((x, y, w, h, \lambda_1, \lambda_2, \lambda_3)\).](image)

3.2 General Representation Method and Existing Problems

The most common method of rotating bounding box representation in oriented object detection is the five-parameter representation, which uses \((x, y, w, h, \theta)\) to represent the rotating bounding box, where \((x, y)\) represents the center point coordinates of oriented bounding box, \((w, h)\) represents the width and height of the rotating bounding box, and \(\theta\) represents the angle of rotating bounding box. The five-parameter representation can be subdivided into the OpenCV representation and the long-edge representation based on the side length and angle information of the rotating bounding box. The OpenCV representation defines the angle as the angle between the edge closer to the X axis and the positive direction of the X axis. Since the angle is periodic, the angle is restricted to be in the range \([-90\ \text{deg}, 0\ \text{deg}]\) to avoid ambiguity in the definition of the bounding box, as shown in Fig. 3(a). The angle of the long-edge definition method is the angle between the long side of the rotating bounding box and the positive direction of the X axis, and its range is confined between \([-90\ \text{deg}, 0\ \text{deg}]\) as shown in Fig. 3(b).
As shown in Fig. 3(c), the OpenCV representation undergoes an abrupt change in loss when the oriented bounding box is close to the horizontal position. Assuming the blue box represents the preselected box \( (0, 0, 50, 25, -90 \text{ deg}) \), and the black box represents the ground-truth box \( (0, 25, 50, -10 \text{ deg}) \), the most direct and expedient regression method will rotate the blue box counterclockwise by a small angle to obtain the prediction box \( (0, 0, 50, 25, -100 \text{ deg}) \), as indicated by the solid arrow. However, the width and height of the anticipated box and the ground-truth box are reversed at this time, and there is a 90-deg variation in the angle, so the preselected box must be regressed in a more complicated way. First, the preselected box is rotated clockwise by a large angle to the box with dots, and then the length and width are simultaneously scaled to generate the final prediction box, as indicated by the dotted arrow, which renders the regression of the rotated box at the border point become unstable.

The long-edge definition method also suffers from the boundary loss discontinuity problem, as shown in Fig. 3(d). Suppose that the blue box is the preset box \( (0, 0, 50, 25, -90 \text{ deg}) \) and the black box is the ground-truth box \( (0, 50, 25, 80 \text{ deg}) \). The simplest way of regression should be to get the predicted box \( (0, 0, 50, 25, -100 \text{ deg}) \) by rotating the blue box counterclockwise by a small angle, as shown by the solid arrow. However, at this time, there is a 180-deg difference between the prediction box and the ground-truth box in terms of angle, and the regression loss is large. Thus, the blue box will be regressed in another way, and it needs to rotate clockwise by a large angle to get the prediction box, as shown by the dashed arrow, which undoubtedly increases the complexity of the regression process of the rotated box.

As shown in Fig. 3(c), the OpenCV representation undergoes an abrupt change in loss when the oriented bounding box is close to the horizontal position. Assuming the blue box represents the preselected box \( (0, 0, 50, 25, -90 \text{ deg}) \), and the black box represents the ground-truth box \( (0, 25, 50, -10 \text{ deg}) \), the most direct and expedient regression method will rotate the blue box counterclockwise by a small angle to obtain the prediction box \( (0, 0, 50, 25, -100 \text{ deg}) \), as indicated by the solid arrow. However, the width and height of the anticipated box and the ground-truth box are reversed at this time, and there is a 90-deg variation in the angle, so the preselected box must be regressed in a more complicated way. First, the preselected box is rotated clockwise by a large angle to the box with dots, and then the length and width are simultaneously scaled to generate the final prediction box, as indicated by the dotted arrow, which renders the regression of the rotated box at the border point become unstable.

The long-edge definition method also suffers from the boundary loss discontinuity problem, as shown in Fig. 3(d). Suppose that the blue box is the preset box \( (0, 0, 50, 25, -90 \text{ deg}) \) and the black box is the ground-truth box \( (0, 0, 50, 25, 80 \text{ deg}) \). The simplest way of regression should be to get the predicted box \( (0, 0, 50, 25, -100 \text{ deg}) \) by rotating the blue box counterclockwise by a small angle, as shown by the solid arrow. However, at this time, there is a 180-deg difference between the prediction box and the ground-truth box in terms of angle, and the regression loss is large. Thus, the blue box will be regressed in another way, and it needs to rotate clockwise by a large angle to get the prediction box, as shown by the dashed arrow, which undoubtedly increases the complexity of the regression process of the rotated box.

The eight-parameter representation is another typical representation for OBBs. Since the eight-parameter representation frequently has multiple forms to represent a rotating bounding box, it is necessary to artificially set the ordering rules of the four vertices to make the detection box have only one representation. RSDet\textsuperscript{13} proposed a vertex ordering method based on vector fork product. First, determine the leftmost point of the rotation box by the vertex coordinates.
Then find the diagonal vertex, which is the third point, by fork multiplication, and finally use the vector formed by these two points and the fork multiplication method to find other two vertices according to counterclockwise or clockwise rotation. As shown in Fig. 4(a), the eight-parameter method typically uses four boundary points \((x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4)\) to represent the oriented object.

The eight-parameter representation method has an initial point shift at the horizontal border location, resulting in loss mutation and a substantial reduction in detection performance. As shown in Fig. 4(b), the black box is close to the horizontal position, and its vertex ordering is \(a\-b\-c\-d\), which is defined as \((x_a, y_a, x_b, y_b, x_c, y_c, x_d, y_d)\) using the eight-parameter definition approach. When the black box is turned counterclockwise with respect to the dashed box by a tiny angle, its vertex order changes to \(b\-c\-d\-a\). Theoretically, the loss of the two rotating boxes is minor due to the tiny rotation angle. But at this moment, the box representation is \((x_{a'}, y_{a'}, x_{b'}, y_{b'}, x_{c'}, y_{c'}, x_{d'}, y_{d'})\), which is vastly different in order of representation from that of the rotated box before rotation, causing a dramatic change in the regression loss.

In summary, the existing five-parameter or eight-parameter representations make the representation of the rotating bounding box correspond to the detection bounding box one by one by adding restrictions. However, at the boundary of the rotating bounding box, the ideal regression exceeds the restricted range, which creates the problem of boundary discontinuity and increases the difficulty of the regression of the oriented bounding box.

Remark 1: The existing five-parameter or eight-parameter representations suffer from border loss discontinuity in the oriented bounding box regression process because the ideal regression is beyond the range of the oriented bounding box definition. This problem can be solved by changing the representation of the oriented object or designing a superior loss function.

### 3.3 ARP Representation Method

#### 3.3.1 Construction of ARP representation method

Most current object detection systems use five- or eight-parameter regression techniques to describe rotating objects. However, as mentioned in Sec. 3.2, these representation methods suffer from the problem of discontinuous border loss. To address the aforementioned issues, this paper introduces a simple representation for oriented objects and a novel detection scheme that divides and conquers nearly horizontal and oriented object detection. To be more specific, we offer the ARP approach for precisely describing an oriented object parameterized by seven tuples \((x, y, w, h, \lambda_1, \lambda_2, \lambda_3)\), as shown in Fig. 5. For every rotating object \(P_o\) (see the blue box in Fig. 5), its left, upper, right, and bottom vertices are represented by \((a, b, c, d)\). The smallest circumscribed rectangle of a rotating object is \(P_h\) (view the black box in Fig. 5), its four vertices are denoted by \((A, B, C, D)\). \((x, y, w, h)\) denotes the central point, width, and height of the minimal bounding rectangle of the rotating object \(P_h\). \(\lambda_1\) is the area ratio between the oriented object \(P_o\) and minimum circumscribed rectangle \(P_h\), \(\lambda_2\) is the area ratio of the parallelogram \(P_a\) (see the green box in Fig. 5) to the smallest circumscribed rectangle \(P_h\). The parallelogram \(P_a\) is formed by the extension line of the right side of the top vertex of an oriented object and the extension line...
of the width of the smallest circumscribed rectangle. \( \lambda_3 \) is the area ratio of the parallelogram \( Pb \) (represented by the red box in Fig. 5) to the smallest circumscribed rectangle \( Ph \). The parallelogram \( Pb \) comprises the extension line of the right side of the top vertex on the oriented object and the extension line of the height of the smallest circumscribed rectangle. The parameter \( \lambda_1, \lambda_2, \) and \( \lambda_3 \) are defined as follows:

\[
\begin{align*}
\lambda_1 &= S_o/S_h, \\
\lambda_2 &= S_o/S_h, \\
\lambda_3 &= S_b/S_h,
\end{align*}
\]

(1)

where \( S_o, S_h, S_a, \) and \( S_b \) are the areas of the oriented object, minimum circumscribed rectangle, and two parallelograms, respectively.

### 3.3.2 Side lengths and vertex coordinates solution

According to the ARP representation method in Fig. 5, the area of the minimum circumscribed and two parallelograms can be represented by the area of oriented object and triangles as follows:

\[
\begin{align*}
S_h &= S_o + 2(S_1 + S_2), \\
S_o &= S_o + 2(S_1 + S_3), \\
S_b &= S_o + 2(S_1 + S_4),
\end{align*}
\]

(2)

where \( S_1, S_2, S_3, \) and \( S_4 \) denote the areas of triangles \( \triangle Aba, \triangle bBc, \triangle EAa, \) and \( \triangle FbA \), respectively. Since the oriented object in aerial images is denoted by a rectangular box, and its smallest circumscribed rectangle is also a rectangle box, our represent method contains two pairs of similar triangles \( \triangle EAa \sim \triangle bBc \) and \( \triangle FbA \sim \triangle bBc \). Given their similarity ratios of \( k_1 \) and \( k_2 \), respectively. We can derive the following relationships between the four triangles:

\[
\begin{align*}
S_1/S_2 &= k_1^2, \\
S_4/S_2 &= k_2^2, \\
S_1/S_2 &= k_1k_2.
\end{align*}
\]

(3)
Combine Eqs. (2) and (3), the values of $k_1$ and $k_2$ can be obtained as follows:

$$
\begin{align*}
    k_1 &= \sqrt{\frac{(S_y-S_h)^2}{(S_y-S_h)(S_y-S_h)+(S_y-S_h)(S_y-S_h)}} , \\
    k_2 &= \sqrt{\frac{(S_y-S_h)^2}{(S_y-S_h)(S_y-S_h)+(S_y-S_h)(S_y-S_h)}} .
\end{align*}
$$

(4)

By substituting Eq. (4) into Eq. (1), $k_1$ and $k_2$ represented by the area ratios are as follows:

$$
\begin{align*}
    k_1 &= \sqrt{\frac{(\lambda_1-\lambda_2)^2}{(1-\lambda_1)(\lambda_2-\lambda_1)+(1-\lambda_2)(\lambda_3-\lambda_1)}} , \\
    k_2 &= \sqrt{\frac{(\lambda_1-\lambda_2)^2}{(1-\lambda_1)(\lambda_2-\lambda_1)+(1-\lambda_2)(\lambda_3-\lambda_1)}} .
\end{align*}
$$

(5)

Furthermore, the length and coordinates of the intersection point of a rotating object with its smallest circumscribed rectangle can also be acquired as follows:

$$
\begin{align*}
    h_1 &= |Aa| = \frac{k_1}{1+k_2} h , \\
    h_2 &= |aD| = \frac{1}{1+k_2} h , \\
    w_1 &= |Ab| = \frac{k_2}{1+k_2} w , \\
    w_2 &= |bB| = \frac{1}{1+k_2} w , \\
    a &: (x - w/2, y - h/2 + h_1) , \\
    b &: (x - w/2 + w_1, y - h/2) , \\
    c &: (x + w/2, y + h/2 - h_1) , \\
    d &: (x + w/2 - w_1, y + h/2) ,
\end{align*}
$$

where $h_1, h_2$ denote the length of line segment $Aa$ and $aD$, and $w_1, w_2$ indicate the length of the $Ab$ and $bB$ line segment. $(a, b, c, d)$ represent the coordinates of the four vertices of the oriented object. It is seen that the ARP representation method can accurately describe the position information of the rotating object in the image. We can convert the ARP to the more common eight-parameter representation method by applying the Eqs. (5), (6), and (7). Similarly, converting the eight-parameter representation method to our representation method is straightforward.

Thus, the ARP method accurately describes rotating objects and has a unique representation. It solves the discontinuous boundary problem effectively and significantly improves the detection accuracy of rotating objects.

### 3.3.3 Nearly horizontal bounding box

In practice, we observe that the detection fails when an object extremely approaches the horizontal bounding box, as shown in Fig. 6. When the oriented object is close to the horizontal bounding box, the parameter $\lambda_3$ and one of the parameters in $\lambda_2$ and $\lambda_1$ approach one infinitely. Thus, it will be impossible to obtain four vertex coordinates of the rotating object since the value of the parameter $k_1$ or $k_2$ is infinite [calculated by Eq. (5)]. This problem is regarded as an extreme case.

To address this case, we divide the rotating objects into OBBs and horizontal bounding boxes (HBBs) and process them independently in this work. The advantage of this classification strategy is that it allows us to convert difficult-to-handle corner cases into HBBs cases. A nearly horizontal object has a high area ratio $\lambda_1$ that is close to one (see Fig. 5), whereas a highly slender oriented object has a low area ratio $\lambda_3$ that is close to zero. Therefore, the area ratio $\lambda_4$ are adopted as an obliquity factor to guide the selection of OBB and HBB. Without introducing other parameters, the categories of OBB and HBB can be finished. Indeed, it is reasonable to represent nearly horizontal objects with horizontal bounding boxes.

**Remark 2**: The oriented bounding box representation based on the area ratio of the parallelogram represents the oriented bounding box with the minimum circumscribed rectangle and
three area ratios, and the area ratio of the rotating bounding box to the minimum circumscribed rectangle distinguishes the horizontal object and the oriented object. This representation method is unique and can successfully tackle the discontinuous boundary loss problem.

3.4 R-EIoU Loss

The proposed method is based on YOLOv5. Thus, the loss function makes some appropriate modifications based on YOLOv5. The multi-task loss is defined as follows:

$$L = \lambda_{box} \sum_{i=0}^{K \times K} \sum_{j=0}^{M} I_{ij}^{obj} L_{box} + \lambda_{obj} \sum_{i=0}^{K \times K} I_{ij}^{obj} L_{obj} + \lambda_{cls} \sum_{i=0}^{K \times K} \sum_{c \in cls} I_{ij}^{obj} L_{cls} + \lambda_{a} \sum_{i=0}^{K \times K} \sum_{j=0}^{M} I_{ij}^{obj} L_{a},$$

where $K \times K$ denotes the number of grids, $M$ indicates the number of candidate boxes produced by each grid. Each candidate box will get a corresponding bounding box through the network, so we finally obtain the $K \times K \times M$ bounding boxes. $I_{ij}^{obj}$ is a binary label that indicates whether the $j$ anchor box of the $i$ grid is responsible for detecting this object. If it is responsible, it is set to 1; otherwise, it is set to 0. The hyper-parameters $\lambda_{obj}$, $\lambda_{cls}$, $\lambda_{box}$, and $\lambda_{a}$ are used to balance the importance of each loss term. The confidence loss $L_{obj}$ and classification loss $L_{cls}$ are both Focal losses, which are the same as that in YOLOv5. $L_{box}$ signifies the loss for bounding box regression. $L_{a}$ denotes the loss for obliquity factor $\lambda_1$ regression, which is defined as the binary cross-entropy:

$$L_{cls} = - (\alpha \cdot \log(p(\alpha)) + (1 - \alpha) \log(1 - p(\alpha)))),$$

where $\alpha$ denotes a binary label of either 0 or 1, $p(\alpha)$ represents the probability that the output corresponds to the $\alpha$ label. The following is the definition of $\alpha$:

$$\alpha = \begin{cases} 1 & \text{(OBB)} \\ 0 & \text{(HBB)} \end{cases}, \quad \lambda_1 < \lambda_{thr},$$

where $\lambda_{thr}$ is the threshold used to differentiate between the oriented and horizontal bounding boxes, and its optimal value will be determined in Sec. 4.2.3. The loss function for the $L_{box}$ regression contains two terms for the horizontal bounding box $(x, y, w, h)$ and three area ratios $(\lambda_1, \lambda_2, \lambda_3)$. In terms of the horizontal bounding box, $L_{box}$ is the same as the CIoU loss in YOLOv5. Compared with the horizontal bounding, the oriented bounding box adds three parameters $\lambda_1, \lambda_2, \lambda_3$ in the head. Therefore, it needs to augment extra loss to calculate these three variables and facilitate their regression.

Smooth $L1$ loss is a relatively excellent loss function in rotating object detection, which can promote the regression of the object without the problem of non-differentiable. Therefore, we also considered using smooth $L1$ loss to initially deal with the parameters of these three area ratios.
ratios. Given a predicted horizontal bounding box $B$ and an object horizontal bounding box $\hat{B}$, their losses are defined as follows:

$$
L_{box} = L_{\text{CIoU}} + \sum_{i=1}^{3}\text{smooth}L_1(\lambda_i - \hat{\lambda}_i),
$$

(11)

where $L_{\text{CIoU}}$ denotes the CIoU loss between the two horizontal boxes $B$ and $\hat{B}$. However, using CIoU to address horizontal bounding and smooth $L_1$ loss to address three area ratios generates boundary problems in the case of nearly horizontal bounding boxes. Therefore, we propose R-EIoU loss to better solve this problem. Given two predicted parallelogram boxes $P_a$ and $P_b$ and two object boxes $\hat{P}_a$ and $\hat{P}_b$ (see Fig. 5), the R-EIoU loss is defined as follows:

$$
L_{box} = 1 - \text{IoU} + \frac{\rho^2(b, \hat{b})}{c^2} + \frac{\rho^2(w_{pa}, \hat{w}_{pa})}{c^2_{wpa}} + \frac{\rho^2(h_{pb}, \hat{h}_{pb})}{c^2_{hp}}.
$$

(12)

where $\text{IoU}$ represent the Intersection over Union between the $B$ and $\hat{B}$. $b$ and $\hat{b}$ denote the central points of $B$ and $\hat{B}$, respectively. $\rho(\cdot) = \| \cdot \|$ indicates the Euclidean distance. The diagonal length of the smallest enclosing box that encompasses the two HBBs $B$ and $\hat{B}$ is denoted by $c$. $w_{pa}$ and $\hat{w}_{pa}$ signify the width of $P_a$ and $\hat{P}_a$, respectively. $c_{wpa}$ is the width of the smallest enclosing box covering the two parallelogram boxes $P_a$ and $\hat{P}_a$. $h_{pb}$ and $\hat{h}_{pb}$ denote the height of $P_b$ and $\hat{P}_b$, respectively. The height of the minimum enclosing box covering the two parallelogram boxes $P_b$ and $\hat{P}_b$ is denoted by $c_{hp}$. Namely, we divide the loss function into three parts: the IoU loss $1 - \text{IoU}$, the distance loss $\rho^2(b, \hat{b})/c^2$, and the area ratio loss $\rho^2(w_{pa}, \hat{w}_{pa})/c^2_{wpa} + \rho^2(h_{pb}, \hat{h}_{pb})/c^2_{hp}$. 

Remark 3: The R-EIoU loss associates the minimal circumscribed rectangle of the oriented bounding box with the three area ratios, hence, solving the problem of approximate horizontal object detection offset more effectively than the Smooth $L_1$ loss.

4 Experiments

Experiments are performed with Pytorch on a server with 8 GeForce RTX 2080 Ti and $11 \times 8$ $G$ memory. We first give a description of the datasets and then use these datasets to verify the advantages of the proposed method.

4.1 Datasets and Implementation Details

4.1.1 DOTA dataset

DOTA is a large-scale benchmark and challenge for object detection in aerial images, and it is currently the most comprehensive remote sensing image data set for object detection. There are three versions of this data set that has been released, including DOTA1.0, DOTA1.5, and DOTA2.0. Since the current rotation object detection algorithms are based on DOTA1.0, the experiment in this paper is mainly completed on DOTA1.0 to verify the performance of our proposed method. DOTA1.0 includes 2806 aerial images with 188,282 annotated instances, and the size of the image ranges from around 800 $\times$ 800 to 4,000 $\times$ 4,000 pixels. There are 15 categories in total, including plane (PL), baseball diamond (BD), bridge (BR), ground track field (GTF), small vehicle (SV), large vehicle (LV), ship (SH), tennis court (TC), basketball court (BC), storage tank (ST), soccer ball field (SBF), roundabout (RA), harbor (HA), swimming pool (SP), and helicopter (HC). It uses 1/2 of the images as the training set, 1/6 as the verification set, and 1/3 as the test set. The test set results are not public, and it needs to be uploaded to the server for the test. Since the aerial images in the DOTA dataset are of varying sizes, we divide the images
into 1024×1024 pixel sub-images with an overlap of 200 pixels, with is the same as RoI transformer.\textsuperscript{10}

### 4.1.2 HRSC2016 dataset

High-resolution ship collections 2016 (HRSC2016)\textsuperscript{2} is a challenging dataset for ship detection in aerial images. All images come from six famous ports, including two scenes of marine vessels and offshore vessels. It contains 1061 images and more than 20 categories of ships in various appearances. The image size ranges from 300×300 to 1500×900. The training, validation and test sets contain 436 images (including 1207 instances), 181 images (including 541 instances) and 444 images (including 1228 instances).

### 4.1.3 UCAS-AOD dataset

USCA-AOD\textsuperscript{44} is a remote sensing data set for the detection of vehicles and planes issued by the University of Chinese Academy of Sciences. It contains 1510 aerial images of ∼659×1,280 pixels, with two categories of 14,596 instances in total. In line with ICN\textsuperscript{45} and DOTA,\textsuperscript{1} we randomly select 1110 images for training and 400 images for testing.

### 4.1.4 ICDAR2015 dataset

ICDAR2015\textsuperscript{46} is a data set for scene text detection and location proposed by ICDAR 2015 Robust Reading Competition. It contains a total of 1500 images, of which 1000 are utilized for training and the remainder for testing. Text regions are annotated by the four vertices of the quadrangle.

### 4.1.5 Baseline and training details

To demonstrate the generality of our system, we conduct experiments on three aerial benchmarks and one scene text benchmark. RetinaNet,\textsuperscript{9} R\textsuperscript{3}Det,\textsuperscript{11} and YOLOv5 are adopted as the baseline in all ablation studies and adopt YOLOv5 as final experiments. The initial learning rates for RetinaNet, R\textsuperscript{3}Det, and YOLOv5 are 0.0005, 0.001, and 0.01, respectively. The momentum and weight decay are 0.9 and 0.0001, respectively.

The baseline of our experiment is YOLOv5 unless otherwise specified. The training epoch for DOTAv1.0, HRSC2016, USAS-AOD, and ICDIR2015 are 300, 200, 150, and 150, respectively. Additionally, the number of iterations of each epoch is proportional to the sample size in the dataset. The warm-up approach determines an appropriate learning rate during the first quarter of the training epochs. The final detection results are post-processed using rotating non-maximum suppression during the inference. The hyperparameters $\lambda_{\text{box}}, \lambda_{\text{cls}}, \lambda_{\text{obj}}, \text{ and } \lambda_{\alpha}$ in Eq. (8) are set to 0.05, 0.3, 0.7, and 0.8, respectively. Without the extra declaration, other hyper-parameter settings related to the program are consistent with YOLOv5.

### 4.2 Ablation Study

In this section, we conduct a series of ablation experiments on DOTAv1.0, HRSC2016, UCAS-AOD, and ICDAR2015 datasets to evaluate the effectiveness of our proposed method. For fair compassion, all baseline methods are implemented using similar settings to the proposed method. We can take note that sophisticated data augmentation techniques are not employed in this section.

#### 4.2.1 Evaluation of ARP representation method

The RetinaNet, R\textsuperscript{3}Det, and YOLOv5 are employed as our baseline to evaluate the effectiveness of the ARP representation method. Notably, except for the output box representations, the baseline method shares the same architecture and loss (smooth $L_1$ loss) as the proposed method.
We compare our ARP method to others that employ different representations, as given in Table 1. The suggested method outperforms OpenCV representation (DOC) on RetinaNet, R³Det, and YOLOv5 by 3.03%, 0.96%, and 1.71%, respectively, on the DOTAv1.0 dataset. Additionally, comparative tests are conducted on the HRSC2016 dataset to confirm the boosting effect of ARP. Results show that our ARP method also achieves predominant performance. The result suggests that the ARP representation approach is superior to learning the five-parameter (DOC and DLE) and angle classification methods (CSL and DCL) for detecting orientated objects. Figure 7 shows a partial comparison of the detection effects of the proposed ARP oriented bounding box representation and the common oriented bounding box representation. It reveals that the ARP bounding box representation can locate the location of the oriented object more precisely than other rotating bounding box representations, thereby, reducing the offset of the detection results and achieving a higher detection accuracy. This demonstrates conclusively that the ARP representation can successfully overcome the problem of boundary loss discontinuity, hence, enhancing the accuracy of oriented object detection.

### 4.2.2 Evaluation of R-EIoU loss

The R-EIoU loss for our ARP approach is constructed in the manner described in Sec. 3.4. As given in Table 2, four distinct datasets are utilized to test the effectiveness of the R-EIoU loss. As mentioned previously, the basic model is YOLOv5. The regression head produces the orientated bounding box, which is represented by four bounding box parameters and three area ratios. R-EIoU loss achieves mean average precision (mAP) values of 73.33%, 89.46%, 96.58%, and 84.43% on the DOTAv1.0, HRSC2016, USAS-AOD, and ICDIR2015, respectively, which is an

| Method  | Backbone     | Datasets         | BOX DEF | mAP   |
|---------|--------------|------------------|---------|-------|
| RetinaNet | Resnet50     | DOTAv1.0         | DLE     | 64.17 |
|         |              |                  | DOC     | 65.73 |
|         |              |                  | CSL     | 67.38 |
|         |              |                  | DCL     | 67.39 |
|         |              |                  | ARP (ours) | 68.76 |
|         |              |                  | HRSC2016 | 84.28 |
|         |              |                  | ARP (ours) | 86.43 |
| R³Det   | Resnet50     | DOTAv1.0         | DOC     | 70.66 |
|         |              |                  | DCL     | 71.21 |
|         |              |                  | ARP (ours) | 71.62 |
|         |              |                  | HRSC2016 | 88.52 |
|         |              |                  | ARP (ours) | 88.93 |
| YOLOv5  | CSPDarknet53 (yolov5x) | DOTAv1.0 | DOC     | 71.12 |
|         |              |                  | ARP (ours) | 72.83 |
|         |              |                  | HRSC2016 | 88.65 |
|         |              |                  | ARP (ours) | 89.07 |

Note: The bolded method is intended to distinguish our method from other algorithms.
improvement of 0.50%, 0.39%, 0.42%, and 0.27% over the baseline model. Figure 8 shows the detection results of Smooth $L_1$ loss and the R-EIoU loss proposed in this paper for approximate horizontal objects, demonstrating that the R-EIoU loss can effectively solve the problem of detection result bias that exists in the detection of smooth $L_1$ loss for approximate horizontal objects. R-EIoU loss of the orientated bounding box combines horizontal bounding box and area ratio, resolving boundary difficulties and promoting network convergence. Thus, the R-EIoU loss design is effective for our ARP approach.

### Table 2: Ablation study for R-EIoU loss on three datasets. Where R-EIoU loss denotes the loss function proposed in this paper.

| Method  | Backbone                  | Datasets       | Loss            | mAP   |
|---------|---------------------------|----------------|-----------------|-------|
| YOLOv5  | CSPDarknet53 (yolov5x)    | DOTA v1.0      | Smooth $L_1$ loss | 72.83 |
|         |                           |                | R-EIoU loss     | 73.33 |
|         |                           | HRSC2016       | Smooth $L_1$ loss | 89.07 |
|         |                           |                | R-EIoU loss     | 89.46 |
|         |                           | UCAS-AOD       | Smooth $L_1$ loss | 96.16 |
|         |                           |                | R-EIoU loss     | 96.58 |
|         |                           | ICDAR2015      | Smooth $L_1$ loss | 84.16 |
|         |                           |                | R-EIoU loss     | 84.43 |

**Fig. 8** Comparison of detection results by two losses. (a) Smooth $L_1$ loss and (b) R-EIoU loss.
4.2.3 Hyperparameters of ARP

The parameter sensitivity experiments are carried out to establish the appropriate hyperparameter settings. As discussed in Sec. 3.2.3, ARP regression is accurate for oriented objects but less accurate for objects close to the level due to the possibility of confusion. YOLOv5 is used as a baseline to examine the influence of various obliquity factor thresholds $\lambda_{thr}$ on four datasets. As shown in Fig. 9, the best thresholds in DOTAv1.0, HRSC2016, USAS-AOD, and ICDIR2015 are 0.94 (0.95), 0.92, 0.96, and 0.91 (0.92), respectively. Indeed, when a low threshold is applied, HBBs represent some oriented objects, resulting in lower oriented object identification performance. However, when a wide threshold is employed, almost horizontal objects may have difficulty discriminating between horizontal and orientated bounding boxes. Therefore, it is vital for us to choose appropriate threshold values across distinct datasets.

4.3 Comparison with State-of-Art Methods

We compare our proposed method to other state-of-the-art methods in three aerial image detection datasets, DOTA, HRSC2016, and UCAS-AOD, as well as one scene text detection dataset, ICDAR2015.

4.3.1 Results on DOTA

Table 3 compares the proposed approach to various outstanding detectors on DOTA dataset. Because ground-truth labels of the test set are not publicly available, the detection results should be submitted to the official DOTA assessment site. Our approach achieves 75.31% and 76.05% mAP, respectively, using YOLOv5x and YOLOv5x6 as the backbone. Moreover, As can be observed, introducing R-EIoU improves performance by 1.69% and results in 77.74% mAP. In addition, it achieves 79.93% mAP with YOLOv5x6 in the multi-scale (0.5, 1.0, 1.5, 2) during training and testing. The experimental evidence indicates that our method outperforms previous methods, particularly for objects with dense packing and many directions (see SV, LV, SH, and TC in Table 3). Figure 10 shows the detection impact of the proposed technique on the DOTA dataset. The image demonstrates that the proposed technique can achieve accurate detection of oriented objects in remote sensing images and effectively resolving the problem of detection result bias in oriented object detection. These detection results reveal conclusively that our method can successfully resolve the problem of loss discontinuity in the detection of spinning targets, hence enhancing the detection precision of rotating objects.
Table 3  Performance evaluation of OBB task on DOTAv1.0 dataset. R-101 indicates ResNet-101 (likewise for R-50), and H-104 stands for Hourglass-104. YOLOv5x and YOLOv5x6 denote different models in YOLOv5. * denotes multi-scale training and testing.

| Methods          | Backbone | PL  | BD  | BR  | GTF | SV  | LV  | SH  | TC  | BC  | ST  | SBF | RA  | HA  | SP  | HC  | mAP |
|------------------|----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Two-stage:       |          |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| RRPN66           | R-101    | 88.52 | 71.20 | 31.66 | 59.30 | 51.85 | 57.25 | 90.81 | 72.84 | 67.38 | 56.69 | 52.84 | 53.08 | 51.94 | 53.58 | 61.01 |
| ICN65            | R-101    | 81.36 | 74.30 | 47.70 | 70.32 | 64.89 | 67.82 | 69.98 | 90.76 | 79.06 | 78.20 | 53.64 | 62.90 | 67.02 | 64.17 | 50.23 | 68.16 |
| RoI Trans.61     | R-101    | 88.64 | 78.52 | 43.44 | 75.92 | 68.81 | 73.68 | 83.59 | 90.74 | 77.27 | 81.46 | 58.39 | 53.54 | 62.83 | 58.93 | 47.67 | 69.56 |
| CAD-Net67        | R-101    | 87.80 | 82.40 | 49.40 | 73.50 | 71.10 | 63.50 | 76.70 | 90.90 | 79.20 | 73.30 | 58.93 | 62.20 | 69.00 | 62.20 | 69.90 |     |
| SCRDet61         | R-101    | 89.98 | 80.65 | 52.09 | 68.36 | 68.36 | 60.32 | 72.41 | 90.85 | 87.94 | 86.86 | 65.02 | 66.68 | 66.25 | 68.24 | 65.21 | 72.61 |
| Gliding Vet.66   | R-101    | 89.64 | 85.00 | 52.26 | 77.34 | 73.01 | 73.14 | 86.82 | 90.74 | 79.02 | 86.81 | 59.55 | 70.91 | 72.94 | 70.86 | 57.32 | 75.02 |
| MASKOBB48        | RX-101   | 89.56 | 85.95 | 54.21 | 72.90 | 76.52 | 74.16 | 85.63 | 89.85 | 83.81 | 86.48 | 54.89 | 69.64 | 73.94 | 69.09 | 63.32 | 75.33 |
| CSL67            | R-152    | 90.25 | 85.53 | 54.64 | 75.31 | 70.44 | 73.51 | 77.62 | 90.84 | 86.15 | 86.69 | 69.60 | 68.04 | 73.83 | 71.10 | 68.93 | 76.17 |
| SCRDet++49       | R-101    | 90.05 | 84.39 | 55.44 | 73.99 | 77.54 | 71.11 | 86.05 | 90.67 | 87.32 | 87.08 | 69.62 | 68.90 | 73.74 | 71.29 | 65.08 | 76.81 |
| Methods   | Backbone | PL  | BD  | BR  | GTF | SV  | LV  | SH  | TC  | BC  | ST  | SBF | RA  | HA  | SP  | HC  | mAP |
|-----------|----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Single-stage:                  |          |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| DRN                               | H-104    | 88.91 | 80.22 | 43.52 | 63.35 | 73.48 | 70.69 | 84.94 | 90.14 | 83.85 | 84.11 | 50.12 | 58.41 | 67.62 | 68.60 | 52.50 | 70.70 |
| DAL                               | R-101    | 88.61 | 79.69 | 46.27 | 70.37 | 65.89 | 76.10 | 78.53 | 90.84 | 79.98 | 78.41 | 58.71 | 62.02 | 69.23 | 71.32 | 60.65 | 71.78 |
| RSDet                             | R-101    | 89.99 | 82.90 | 48.60 | 65.20 | 69.50 | 70.10 | 70.20 | 90.50 | 85.60 | 83.40 | 62.50 | 63.90 | 65.60 | 67.20 | 68.00 | 72.20 |
| P-RSDet                           | R-101    | 88.58 | 77.83 | 50.44 | 69.29 | 71.10 | 75.79 | 78.66 | 90.88 | 80.10 | 81.71 | 57.92 | 63.03 | 66.30 | 69.77 | 63.13 | 72.30 |
| RIDet-Q                           | R-101    | 87.38 | 75.64 | 44.75 | 70.32 | 77.87 | 79.43 | 87.43 | 90.72 | 81.16 | 82.52 | 59.36 | 63.63 | 68.11 | 71.94 | 51.42 | 72.78 |
| CFC-Net                           | R-50     | 89.08 | 80.41 | 52.41 | 70.02 | 76.28 | 78.11 | 87.21 | 90.89 | 84.47 | 85.64 | 60.51 | 61.52 | 67.82 | 68.02 | 50.09 | 73.50 |
| RIDet-O                           | R-101    | 88.94 | 78.45 | 46.87 | 72.63 | 77.63 | 80.68 | 88.18 | 90.55 | 81.33 | 83.61 | 64.85 | 63.72 | 73.09 | 73.13 | 56.87 | 74.70 |
| RBAVector                         | R-101    | 88.63 | 84.06 | 52.13 | 69.56 | 78.26 | 80.40 | 88.06 | 90.87 | 87.23 | 86.39 | 56.11 | 65.62 | 67.10 | 72.08 | 63.96 | 75.36 |
| R3Det                              | R-152    | 89.80 | 83.77 | 48.11 | 66.77 | 78.76 | 83.27 | 87.84 | 90.82 | 85.38 | 85.51 | 65.67 | 62.68 | 67.53 | 78.56 | 72.62 | 76.47 |
| PolarDet                          | R-101    | 89.65 | 87.07 | 48.14 | 70.97 | 78.53 | 80.34 | 87.45 | 90.76 | 85.63 | 86.67 | 61.64 | 70.32 | 71.92 | 73.09 | 67.15 | 76.64 |
| R3Det-DCL                         | R-152    | 89.26 | 83.60 | 53.54 | 72.76 | 79.04 | 82.56 | 87.31 | 90.67 | 86.59 | 86.98 | 67.49 | 66.88 | 73.29 | 70.56 | 69.99 | 77.37 |
| S2A-Net                           | R-50     | 89.89 | 83.60 | 57.74 | 81.95 | 79.94 | 83.19 | 89.11 | 90.78 | 84.87 | 87.81 | 70.30 | 68.25 | 78.30 | 77.01 | 69.58 | 79.42 |
| ROSD                              | R-101    | 88.28 | 85.74 | 59.79 | 77.46 | 79.48 | 84.02 | 88.32 | 90.82 | 87.45 | 85.65 | 66.80 | 66.68 | 78.78 | 82.52 | 74.70 | 79.76 |
| ARP                               | YOLOv5x  | 85.19 | 83.47 | 51.36 | 75.32 | 73.12 | 78.49 | 85.16 | 90.45 | 80.19 | 86.39 | 59.45 | 64.83 | 76.92 | 77.35 | 61.94 | 75.31 |
| ARP                               | YOLOv5x6 | 85.97 | 83.04 | 52.73 | 76.11 | 75.92 | 79.50 | 88.91 | 90.68 | 79.15 | 88.33 | 61.20 | 63.28 | 77.29 | 75.83 | 62.94 | 76.05 |
| ARP+R-EIoU                        | YOLOv5x6 | 88.53 | 83.71 | 53.76 | 80.07 | 77.42 | 82.19 | 84.70 | 90.79 | 86.73 | 87.44 | 64.04 | 63.39 | 76.77 | 80.67 | 66.03 | 77.74 |
| ARP+R-EIoU                        | YOLOv5x6 | 89.32 | 85.28 | 58.12 | 81.12 | 78.54 | 84.11 | 89.43 | 90.91 | 87.42 | 87.88 | 68.76 | 67.01 | 77.87 | 81.87 | 71.32 | 79.93 |

Note: The bolded values are intended to more intuitively reflect the optimal value of each category and the whole, whereas the bolded method is intended to distinguish our method from other algorithms.
4.3.2 Results on HRSC2016

The experimental results on HRSC2016 are evaluated under PASCAL VOC2007 and VOC2012 metrics. HRSC2016 dataset contains lots of large aspect ratios and multiple directions ship, which poses a considerable challenge for oriented object detection. As shown in Table 4, our method achieves 89.86% and 96.27% mAP under VOC2007 and VOC2012 metrics, respectively, outperforming most methods. Additionally, our model with YOLOv5× reaches 30 FPS on RTX 2080 Ti GPU, exceeding major of detectors. More crucially, our model with YOLOv5×6 delivers state-of-the-art performance, scoring around 90.52% and 96.63% on 2007 and 2012 evaluation metrics, respectively. Figure 11 shows the detection effect graph of the proposed method on the HRSC2016 dataset, from which it is evident that the proposed method is still advantageous in the detection of slender objects, such as ships, and can accurately detect the position of ships.

4.3.3 Results on UCAS-AOD

VOC2012 measures are used to evaluate the experimental results on USAC-AOD dataset. As illustrated in Table 5, Our approach yields 97.04% mAP in testing, outperforming existing advanced rotation detectors such as R3Det and PolarDet. Furthermore, our technique has the best performance in detecting small cars, indicating that it is robust to small and densely scattered objects. The visualization results are shown in Fig. 12.

4.3.4 Results on ICDAR2015

The generalizability of our method across different scenes was also tested through experiments on the scene text detection dataset. As given in Table 6, our algorithm yields an 84.83% F-measure, outperforming the majority of aerial detectors and text detectors, such as CSL and RRD. However, our strategy does not account for the case of a significant number of lengthy
Table 4 Results and speed comparison on the HRSC2016 dataset.

| Method               | Backbone  | mAP (07) | mAP (12) | FPS |
|----------------------|-----------|----------|----------|-----|
| R2CNN25              | ResNet101 | 73.07    | 79.73    | 2   |
| RRPN26               | ResNet101 | 79.08    | 85.64    | 3.5 |
| ROI-Transformer10    | ResNet101 | 86.20    | —        | 6   |
| RSDet13              | ResNet50  | 86.50    | —        | —   |
| Gliding vertex12     | ResNet101 | 88.20    | —        | 10  |
| BBAVectors16         | ResNet101 | 88.60    | —        | 11.7|
| DAL51                | ResNet101 | 88.95    | —        | 34  |
| R3Det11              | ResNet101 | 89.26    | 96.01    | 12  |
| R3Det-DCL18          | ResNet101 | 89.46    | 96.41    | —   |
| R4Det65              | ResNet101 | 89.56    | —        | 6.5 |
| CSL17                | ResNet50  | 89.62    | 96.10    | —   |
| R2Det-O22            | ResNet101 | 89.63    | —        | —   |
| CFC-Net40            | ResNet101 | 89.70    | —        | 28  |
| S2A-Net53            | ResNet101 | 90.17    | 95.01    | 12.7|
| PolarDet52           | ResNet50  | 90.46    | —        | 25  |
| ARP+R-EIoU (Ours)    | YOLOv5x   | 89.86    | 96.27    | 30  |
| ARP+R-EIoU (Ours)    | YOLOv5x6  | 90.52    | 96.63    | 14  |

Note: The bolded values are intended to more intuitively reflect the optimal value of each category and the whole, whereas the bolded method is intended to distinguish our method from other algorithms.

Fig. 11 Detection results on HRSC2016 dataset with our method.
texts that are incorrectly identified as multiple short paragraphs. As a result, our system still falls short of state-of-the-art scene text identification.

5 Conclusions
In this paper, the ARP representation method is suggested for the definition of OBBs. It represents an oriented object by utilizing the central point coordinates, width and height of the smallest bounding rectangle of the oriented object and introducing three area ratios. Based on this, the area ratio of the oriented object and its smallest circumscribed rectangle is adopted to distinguish nearly horizontal objects. ARP method has a unique representation and facilitates offset learning. Thus, it is superior to the method of directly anticipating five-parameter or eight-parameter detection results on USAC-AOD dataset.

| Method                | Plane | Car | mAP  |
|-----------------------|-------|-----|------|
| YOLOv2                | 96.60 | 79.20 | 87.90 |
| R-DFPN                | 95.60 | 82.50 | 89.20 |
| DRBox                 | 94.90 | 85.00 | 89.95 |
| S²ARN                 | 97.60 | 92.20 | 94.90 |
| RetinaNet-H           | 97.34 | 93.60 | 95.47 |
| FADet                 | 98.69 | 92.72 | 95.71 |
| R³Det                 | 98.20 | 94.14 | 96.17 |
| RSDet                 | 98.04 | 94.97 | 96.50 |
| SCRDet++              | 98.93 | 94.97 | 96.95 |
| PolarDet              | **99.08** | 94.96 | 97.02 |
| ARP+R-EIoU (Ours)     | 98.91 | **95.17** | **97.04** |

Note: The bolded values are intended to more intuitively reflect the optimal value of each category and the whole, whereas the bolded method is intended to distinguish our method from other algorithms.

Fig. 12 Visualization of detection results on UCAS-AOD dataset with our method.
parameter bounding boxes. Furthermore, R-EIoU loss is designed to connect the minimum circumscribed rectangle of the rotating object and three area ratios proposed by the ARP method. It accelerates network convergence and achieves higher performance in rotated detection when compared to the smooth $L_1$ loss. Extensive detailed ablation experiments and comparative experiments on multiple rotation detection datasets, including DOTA, HRSC2016, UCAS-AOD, and ICDAR2015, proved the superiority of our method. Nevertheless, the proposed technique has a large number of preset boxes and uneven positive and negative samples. It is still worth considering how to create a label assignment method based on the oriented bounding box that can achieve relatively decent detection accuracy while lowering the amount of calculation.

Acknowledgments

We gratefully acknowledge the support of National Key R&D Program of China (Grant No. 2018YFB1308400) and Natural Science Foundation of Zhejiang Province (Grant No. NO.LY21F030018).

References

1. G. S. Xia et al., “DOTA: a large-scale dataset for object detection in aerial images,” in Proc. IEEE Conf. Comput. Vision and Pattern Recognit., pp. 3974–3983 (2018).
2. Z. K. Liu et al., “A high-resolution optical satellite image dataset for ship recognition and some new baselines,” in Int. Conf. Pattern Recognit. Appl. and Methods, pp. 2324–2331 (2017).
3. Y. Li et al., “Automatic target detection in high-resolution remote sensing images using a contour-based spatial model,” IEEE Geosci. Remote Sens. Lett. 9(5), 886–890 (2012).
4. J. W. Han et al., “Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding,” ISPRS J. Photogramm. Remote Sens. 89, 37–48 (2014).
5. D. Hong et al., “An augmented linear mixing model to address spectral variability for hyperspectral unmixing,” IEEE Trans. Image Process. 28, 1923–1938 (2018).
6. R. Girshick et al., “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 580–587 (2014).
7. S. Q. Ren et al., “Faster R-CNN: towards real-time object detection with region proposal networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6), 1137–1149 (2016).
8. J. Redmon et al., “You only look once: unified, real-time object detection,” in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 779–788 (2016).
9. T. Y. Lin et al., “Focal loss for dense object detection,” in *IEEE Int. Conf. Comput. Vision*, pp. 2980–2988 (2017).
10. J. Ding et al., “Learning ROI transformer for oriented object detection in aerial images,” in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 2849–2858 (2019).
11. X. Yang et al., “R3Det: refined single-stage detector with feature refinement for rotating object,” in *AAAI Conf. Artif. Intell.*, pp. 3163–3171 (2021).
12. Y. C. Xu et al., “Gliding vertex on the horizontal bounding box for multi-oriented object detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(4), 1452–1459 (2020).
13. W. Qian et al., “Learning modulated loss for rotated object detection,” in *AAAI Conf. Artif. Intell.*, Vol. 35, No. 3, pp. 2458–2466 (2021).
14. L. Zhou et al., “Arbitrary-oriented object detection in remote sensing images based on polar coordinates,” *IEEE Access*, 8, 223373–223384 (2020).
15. Q. Wu et al., “Bounding box projection for regression uncertainty in oriented object detection,” *IEEE Access*, 9, 58768–58779 (2021).
16. J. R. Yi et al., “Oriented object detection in aerial images with box boundary-aware vectors,” in *IEEE Winter Conf. Appl. Comput. Vision*, pp. 2150–2159 (2021).
17. X. Yang and J. C. Yan, “Arbitrary-oriented object detection with circular smooth label,” in *Eur. Conf. Comput. Vision*, pp. 677–694 (2020).
18. X. Yang et al., “Dense label encoding for boundary discontinuity free rotation detection,” in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 15819–15829 (2021).
19. X. Yang et al., “Rethinking rotated object detection with Gaussian Wasserstein distance loss,” in *Int. Conf. Mach. Learn.*, pp. 11830–11841 (2021).
20. X. Yang et al., “Learning high-precision bounding box for rotated object detection via Kullback-Leibler divergence,” arXiv:2106.01883 (2022).
21. R. Girshick, “Fast R-CNN,” in *IEEE Int. Conf. Comput. Vision*, pp. 1440–1448 (2015).
22. K. M. He et al., “Mask R-CNN,” in *IEEE Int. Conf. Comput. Vision*, pp. 2961–2969 (2017).
23. J. Redmon and A. Farhadi, “YOLO9000: better, faster, stronger,” in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 7263–7271 (2017).
24. J. Redmon and A. Farhadi, “YOLOv3: an incremental improvement,” arXiv:1804.02767 (2018).
25. A. Bochkovskiy et al., “YOLOv4: optimal speed and accuracy of object detection,” arXiv:2004.10934 (2020).
26. W. Liu et al., “SSD: single shot multibox detector,” in *Eur. Conf. Comput. Vision*, pp. 21–37 (2016).
27. H. Law and J. Deng, “CornerNet: detecting objects as paired keypoints,” in *Eur. Conf. Comput. Vision*, pp. 734–750 (2018).
28. K. W. Duan et al., “CenterNet: keypoint triplets for object detection,” in *Int. Conf. Comput. Vision*, pp. 6568–6577 (2019).
29. J. H. Yu et al., “Unitbox: an advanced object detection network,” in *24th ACM Int. Conf. Multimedia*, pp. 516–520 (2016).
30. H. Rezatofighi et al., “Generalized intersection over union: a metric and a loss for bounding box regression,” in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 658–666 (2019).
31. Z. H. Zheng, et al., “Distance-IoU loss: faster and better learning for bounding box regression,” in *AAAI Conf. Artif. Intell.*, pp. 12993–13000 (2020).
32. Y. F. Zhang et al., “Focal and efficient IOU loss for accurate bounding box regression,” arXiv:2101.08158 (2021).
33. X. Wu et al., “ORSIm detector: a novel object detection framework in optical remote sensing imagery using spatial-frequency channel features,” *IEEE Trans. Geosci. Remote Sens.*, 57, 5146–5158 (2019).
34. X. Wu et al., “Fourier-based rotation-invariant feature boosting: an efficient framework for geospatial object detection,” *IEEE Geosci. Remote Sens. Lett.* **17**, 302–306 (2020).
35. Y. Y. Jiang et al., “R2CNN: rotational region CNN for orientation robust scene text detection,” arXiv:1706.09579 (2017).
36. J. Q. Ma et al., “Arbitrary-oriented scene text detection via rotation proposals,” *IEEE Trans. Multimedia* **20**(11), 3111–3122 (2018).
37. M. H. Liao, B. G. Shi, and X. Bai, “Textboxes++: a single-shot oriented scene text detector,” *IEEE Trans. Image Process.* **27**(8), 3676–3690 (2018).
38. M. H. Liao et al., “Rotation-sensitive regression for oriented scene text detection,” in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 5909–5918 (2018).
39. X. Y. Zhou et al., “EAST: an efficient and accurate scene text detector,” in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 5551–5560 (2017).
40. Q. Ming et al., “CFC-Net: a critical feature capturing network for arbitrary-oriented object detection in remote-sensing images,” *IEEE Trans. Geosci. Remote Sens.* 1–14 (2021).
41. X. Yang et al., “SCRDet: towards more robust detection for small, cluttered and rotated objects,” in *IEEE Int. Conf. Comput. Vision*, pp. 8232–8241 (2019).
42. Q. Ming et al., “Optimization for oriented object detection via representation invariance loss,” *IEEE Geosci. Remote Sens. Lett.* **19**, 1–5 (2021).
43. S. Liu et al., “Path aggregation network for instance segmentation,” in *IEEE Conf. on Comput. Vision and Pattern Recognit. (CVPR)*, pp. 8759–8768 (2018).
44. H. G. Zhu et al., “Orientation robust object detection in aerial images using deep convolutional neural network,” in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 3735–3739 (2015).
45. S. M. Azimi et al., “Towards multi-class object detection in unconstrained remote sensing imagery,” in *Asian Conf. Comput. Vision*, pp. 150–165 (2018).
46. D. Karatzas et al., “Icdar 2015 competition on robust reading,” in *13th Int. Conf. Doc. Anal. Recognit.*, pp. 1156–1160 (2015).
47. G. J. Zhang, S. J. Lu, and W. Zhang, “CAD-Net: a context-aware detection network for objects in remote sensing imagery,” *IEEE Trans. Geosci. Remote Sens.* **57**(12), 10015–10024 (2019).
48. J. W. Wang et al., “Mask OBB: a semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images,” *Remote Sens.* **11**(24), 2930 (2019).
49. X. Yang et al., “Scrdet++: detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing,” arXiv:2004.13316 (2020).
50. X. J. Pan et al., “Dynamic refinement network for oriented and densely packed object detection,” in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 11207–11216 (2020).
51. Q. Ming et al., “Dynamic anchor learning for arbitrary-oriented object detection,” in *AAAI Conf. Artif. Intell.*, Vol. 35, No. 3, pp. 2355–2363 (2021).
52. P. B. Zhao et al., “PolarDet: a fast, more precise detector for rotated target in aerial images,” *Int. J. Remote Sens.* **42**(15), 5831–5861 (2021).
53. J. M. Han et al., “Align deep features for oriented object detection,” *IEEE Trans. Geosci. Remote Sens.* **60**, 1–11 (2021).
54. K. Zhang, Q. Q. Zeng, and X. Y. Yu, “ROSD: refined oriented staged detector for object detection in aerial image,” *IEEE Access.* **9**, 66560–66569 (2021).
55. P. Sun et al., “R4Det: refined single-stage detector with feature recursion and refinement for rotating object detection in aerial images,” *Image Vision Comput.* **103**, 104036 (2020).
56. X. Yang et al., “Automatic ship detection in remote sensing images from Google Earth of complex scenes based on multi-scale rotation dense feature pyramid networks,” *Remote Sens.* **10**(1), 132 (2018).
57. L. Liu, Z. X. Pan, and B. Lei, “Learning a rotation invariant detector with rotatable bounding box,” arXiv:1711.09405 (2017).
58. S. Z. Bao et al., “Single shot anchor refinement network for oriented object detection in optical remote sensing imagery,” *IEEE Access* **7**, 87150–87161 (2019).
59. C. Z. Li et al., “Feature-attentioned object detection in remote sensing imagery,” in *IEEE Int. Conf. Image Process.*, pp. 3886–3890 (2019).
Yu et al.: Oriented object detection in aerial images based on area ratio of parallelogram

60. M. H. Liao et al., “Real-time scene text detection with differentiable binarization,” in *AAAI Conf. Artif. Intell.*, Vol. 34, No. 7, 11474–11481 (2020).

**Xinyi Yu** is an associate professor at Zhejiang University of Technology. He received his PhD in mechanical design and theory from Harbin Institute of Technology in 2009. As the person in charge, he presided over 13 projects, accepted and published more than 30 robotics-related articles, and authorized more than 10 invention patents. His current research interests include intelligent learning and robotic systems, multi-machine collaborative control, human–machine integration and intelligent manufacturing system.

**Mi Lin** is currently pursuing his MS degree in control science and engineering at Zhejiang University of Technology in Hangzhou, China. His research focuses on object detection in aerial images and 3D object detection based on LiDAR.

**Jiangping Lu** is currently working toward his MS degree in control science and engineering from Zhejiang University of Technology, Hangzhou, China. His research focuses on oriented object detection in aerial images.

**Linlin Ou** is a professor at Zhejiang University of Technology. She received her doctorate degree in control science and engineering from Shanghai Jiaotong University in 2006. She has published and accepted nearly 35 academic papers and authorized more than 20 invention patents. Her current research interests include intelligent learning and robotic systems, multi-machine collaborative control, and human–machine integration.