IEEE Access
Multidisciplinary ⁞ Rapid Review ⁞ Open Access Journal

Received October 23, 2020, accepted November 4, 2020, date of publication November 11, 2020,
date of current version November 24, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.3037350*

# SAR: Single-Stage Anchor-Free Rotating Object Detection

**JUNYAN LU**[1,2,3], **TIE LI**[4], **JINGYU MA**[3], **ZHUQIANG LI**[3], **AND HONGGUANG JIA**[1,2,3]
[1]Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]Key Laboratory of Satellite Remote Sensing Application Technology of Jilin Province, Chang Guang Satellite Technology Company Ltd., Chang-chun 130000, China
[4]Shanghai Electro-Mechanical Engineering Institute, Shanghai 201109, China

Corresponding author: Hongguang Jia (jiahg@ciomp.ac.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61903031 and in part by the Fundamental Research Funds for the Central Universities under Grant FRF-TP-18-035A1.

**ABSTRACT** As object detection is widely adopted in aerial images, scene texts and other fields, rotating object detection plays an important role and draws attention since it can provide highly accurate orientation and scale information. In this article, we propose a novel and simple baseline to effectively conduct rotating object detection. First, we design a brand-new representation for rotating objects by using a circle cut horizontal rectangle (CCH). The CCH ensures that the regression parameters will not exceed the defined domain and avoids vertex sorting, thus solving some problems in current common representations, including the boundary problem and order problem, and improving the robustness. Second, we design a lightweight head based on the CCH to add the rotating regression to classic benchmark in an almost cost-free manner and propose a single-stage anchor-free rotating (SAR) object detection convolutional neural network. Finally, we demonstrate the details of our method by applying it to data sets with different scenarios. The experiments confirm that our method achieves competitive accuracy and state-of-the-art speed in aerial image and scene text detection.

**INDEX TERMS** Rotating object detection, representation, circle cut horizontal, single-stage, anchor-free.

## I. INTRODUCTION

Object detection is a classic task in computer vision, and considerable progress has been achieved based on the convolutional neural network (CNN) from deep learning. Generally, the main methods of this research field do not consider the orientation of an object; that is, they use a horizontal annotation to represent the object. In recent years, with the application of object detection in remote sensing images [1]–[5], [35]–[37], scene text [6]–[11], face recognition [12]–[14] and other fields, rotating object detection has played an important role and drawn attention. Because of the advantage of providing highly accurate orientation and scale information, rotating object detection can effectively improve the detection quality in many cases, such as objects with large aspect ratios and dense, small objects, and can be applied to further applications, such as type recognition and change detection.

Rotating object detection methods are mainly based on classic object detection benchmarks [15]–[19] and are driven by the innovation and optimization of their network structure, loss function, training tricks, etc. On this basis, state-of-the-art rotating object detection methods have achieved promising results. Classic object detection benchmarks use the center, width and height $(x, y, w, h)$ or the top left and bottom right corners $(x_{min}, y_{min}, x_{max}, y_{max})$ of a horizontal rectangle to annotate objects, which most rotating object detection methods imitate. Specifically, current rotating object detection methods mainly use two representations for the bounding box of rotating objects: the angle-based 5-parameter rotating rectangle $(x, y, w, h, \theta)$ [2]–[6], [10], [11], where $\theta$ is the angle of rotation, or the vertex-based 8-parameter arbitrary quadrilateral $(x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4)$ [7], [20]–[23], where $(x_i, y_i)$ is the vertex coordinate that represents multioriented objects. Fig. 1 illustrates the rotating rectangle representation in OpenCV and the arbitrary quadrilateral representation in the DOTA [31] data set. Most of the representations in related works use these two representations or are similar.

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

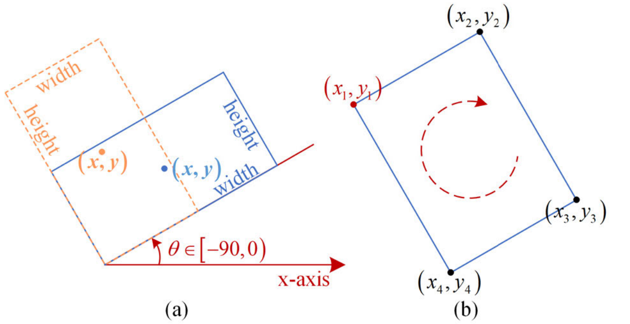

**FIGURE 1.** **(a) Rotating rectangle representation in OpenCV. The ground truth is annotated as $(x, y, w, h, \theta)$, where $(x, y)$ denotes the center point of the rotating rectangle; $w$ denotes the sides with an acute angle with the positive x-axis; $h$ denotes the other sides; and $\theta$ is the acute angle, and its range in OpenCV is defined as $[-90, 0)$. (b) Arbitrary quadrilateral representation in DOTA. The ground truth is annotated as $(x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4)$ in DOTA, where $(x_i, y_i)$ denotes each vertex of the arbitrary quadrilateral. The vertices are arranged in clockwise order, where $(x_1, y_1)$ is the top left corner.**

However, these representations have faced some limitations. The angle-based 5-parameter representation is less robust, and a minor angular deviation will result in a large error, especially for objects with large aspect ratios. For the vertex-based 8-parameter representation, there are no constraints between the vertices of the arbitrary quadrilateral, unlike a rectangle, which may increase the training difficulty and lead to poor results. Moreover, Yang *et al.* [24] propose that using the above two representations will encounter the boundary problem. Specifically, for the angle-based representation, when the ideal regression pattern exceeds the domain of the angle, the regression loss will increase sharply. At this time, the regressor has to abandon the ideal pattern and adopt a more complicated and unreasonable pattern for prediction to avoid this sharp increase in the loss; thus, the regression results near the boundary will be very poor. The situation is most obvious when the object is close to the horizontal or vertical because the boundary of the angle's domain is usually $k\pi$ or $k\pi/2$. The vertex-based representation has an order problem ([24] also regards it as a boundary problem) when the ideal regression pattern gives a prediction but its vertex order does not correspond to the ground truth. At this time, the loss will also be very large, and the approach taken by the regressor is the same, which leads to poor results.

Addressing the above issues, we propose a novel and simple baseline in this article to effectively conduct rotating object detection. Specifically, we design a brand-new representation for rotating objects by using a circle cut horizontal rectangle (CCH), as shown in Fig. 2. For a specific object, we assume that its rotating bounding box is a standard rectangle (the red one, which we call the r-box in the following), and it must have a unique horizontal minimum enclosing rectangle (the blue one, which we call the h-box in the following). Furthermore, the vertices of the r-box must be on a concentric circle of the h-box, and obviously, the circle is unique as well. In other words, when we determine an h-box, as long as we determine the corresponding rotating parameters (such as the radius of the circle, etc.), we can uniquely determine the r-box. Based on this, we propose predicting the classification and h-box of the object by using classic benchmarks, and we use a newly added lightweight CCH head to predict the rotating parameters used to determine the circle and r-box. Note that the h-box is not conceptually equivalent to the bounding box in classic benchmarks, but they are essentially the same. The r-box is obtained on the basis of the h-box, and the regressions of the rotating parameter and horizontal parameter are completely decoupled; therefore, the horizontal regression of classic benchmark is performed independently and completely, that is, the method can add the rotating regression to classic benchmark without almost any modification of the details to the original process and strategy, so as to realize the rotating object detection. Additionally, Because of the one-to-one correspondence between the r-box and h-box, our method is suitable for scenarios where a classic benchmark is applicable, including when the bounding boxes overlapping or one is inside another.

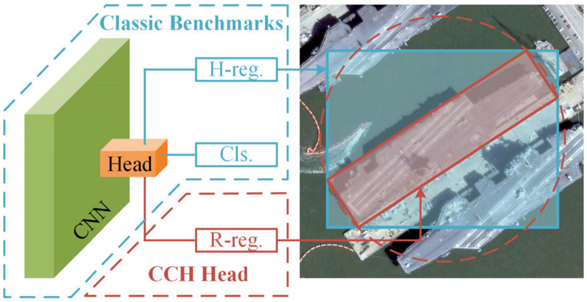

**FIGURE 2.** **Framework of our method. Cls. and H-reg. refer to the classifier and regressor in the head of the classic benchmark, respectively. R-reg. refers to the proposed CCH head used to regress the rotating parameters.**

Compared to the angle-based method, the CCH ensures that the regression will not cross the boundary, which fundamentally solves the boundary problem while enhancing the robustness. Compared to the vertex-based method, the prediction of the CCH is a rectangle that makes the result regular while avoiding the confusion of the order problem. We will discuss the above points in detail in the specific definition of the CCH in section III.

In summary, there are three main contributions of our work:

1) The novel representation CCH improves some issues of the current most popular representations, thereby improving some of the difficulties in rotating object detection accordingly.

2) The lightweight CCH head can add the rotating regression to classic benchmarks in an almost cost-free manner, which can be characterized as plug-and-play.

3) Based on the classic benchmark, we propose a CNN model with the CCH for rotating object detection. Follow-up experiments have confirmed its competitive performance compared to state-of-the-art methods.

## II. RELATED WORKS

Scholars have proposed some solutions to address the issues mentioned above. For the angle-based representation,

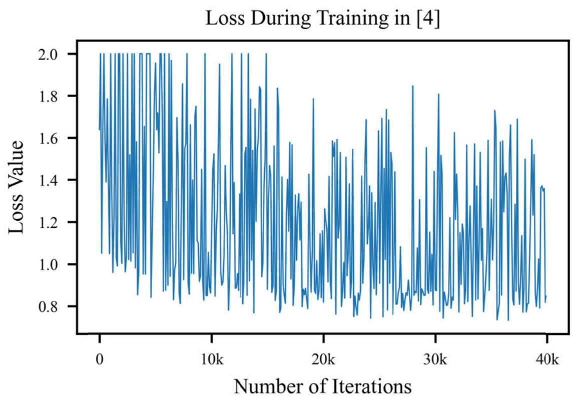

**FIGURE 3.** **We use the method in [4] to train for 40k iterations and draw the loss curve. For visual effects, we sample 400 points at intervals of 100 and truncate the loss values greater than 2.**

Yang *et al.* [4] and Ma *et al.* [10] use the periodicity of an angle to limit $\theta$ to the domain by adding $k\pi$ or $k\pi/2$ before calculating the loss. However, we find that this approach often has difficulty converging in actual training, as shown in Fig. 3, and we believe that the logic of this method is not strict enough in theory. Yang *et al.* [5] propose the IoU-smooth L1 loss by adding an IoU constant factor to the original smooth L1 loss, which eliminates the loss surge under the ideal regression pattern on the boundary condition and retains the original gradient direction to ensure the continuity of the loss function. For the vertex-based representation, Yang *et al.* [23] enumerate the sequence of prediction vertices and calculate the loss with the ground truth in turn, which solves the lack of correspondence of the vertex sequence. However, the above two methods do not solve the root cause of the boundary problem. Instead, various strategies are used in the loss calculation to eliminate or reduce the impact of the boundary problem.

Boundary issues do exist in angle-based and vertex-based representations; therefore, some scholars have attempted to circumvent the boundary problem by proposing new representations. Xu *et al.* [25] propose a gliding vertex representation, which aims to avoid vertex sorting by predicting the offset on each side of the bounding box. However, we believe that the order problem still exists in the representation because the offsets need to be paired with the vertices. More-over, the authors also find that the results near the boundary condition have a large deviation, so they add a regression parameter named the obliquity factor to remedy this. Yang *et al.* [24] point out that the root cause of the boundary problem is that the prediction result exceeds the domain; therefore, they propose a circular smooth label representation, which converts the angular regression task into a classification task (unlike regression, classification will not exceed the domain because the number of classes is limited). We believe this method truly addresses the essence of the boundary problem, but the proposed representation has two drawbacks. First, converting the regression task into a classification task also converts the problem from continuous to discrete, which forfeits prediction precision. If one wants to minimize the precision loss, the classification label needs to be very fine, which will lead to a very thick classifier head. Second, the effect of the representation relies heavily on the selection of its window function, which introduces new hyperparameters that increase the difficulty of tuning.

## III. PROPOSED METHOD

### A. CIRCLE CUT HORIZONTAL RECTANGLE (CCH) ROTATING OBJECT REPRESENTATION

As described in section I, we propose the CCH representation whose specific definition is illustrated in Fig. 4. In the figure, the green rectangle is the h-box, and the blue rectangle is the r-box. Their definitions are the same as in section I, including the circle. In general, there are 8 intersections between the circle and h-box, which can determine 4 r-boxes with different inclination directions and area sizes, as shown in Fig. 4 (d). Therefore, we use three rotating parameters ($\alpha, \beta, \gamma$), which represent the size of the concentric circle radius, the inclination direction of the r-box, and the relative area size of the two r-boxes in the same inclination direction, respectively, to uniquely determine an r-box based on the h-box. The special cases of the CCH are shown in Fig. 5. Here, we regard the horizontal sides of the h-box as the width ($w$) and the vertical sides as the height ($h$). Fig. 4 and Fig. 5 show only the case when $w$ is the short side, and the definition is the same when $h$ is the short side. When $w = h$, either can be considered the short side.

The CCH enjoys the following advantages.

1) The CCH converts part of the regression task into a classification task ($\beta$ and $\gamma$ are discrete binary classifications), and the remaining regression part ($\alpha$) is normalized and limited to [0, 1], ensuring that the prediction will never exceed the domain of the definition, which fundamentally eliminates the boundary problem. Moreover, $\alpha$ is still a continuous variable without discretization, so there is no loss of regression precision.

2) We study the impact of prediction errors on the CCH and angle-based methods. We assume that there are errors in predicting $\alpha$ of the CCH and $\theta$ of the OpenCV representation, while the predictions of other parameters are accurate. In addition, we assume the ground truth of $\theta$ is $-60$ degrees, and the aspect ratio is large ($w : h = 5$). At this time, the ground truth of $\alpha$ is approximately 0.49. Fig. 6 depicts the IoUs between the deviation results and ground truths under the CCH and OpenCV representations when the prediction errors of $\alpha$ and $\theta$ are within plus or minus 30% of their respective domain scopes. It can be seen that the IoU of the CCH is almost always higher than the OpenCV representation. In other words, when the predicted $\alpha$ and $\theta$ values have errors on the same scale, using the CCH to describe the result is more accurate than using the angle-based method. That is, the angle-based representation is highly sensitive to the angle prediction error, while the CCH is more robust. Fig. 7 shows the deviation results and ground truths when the

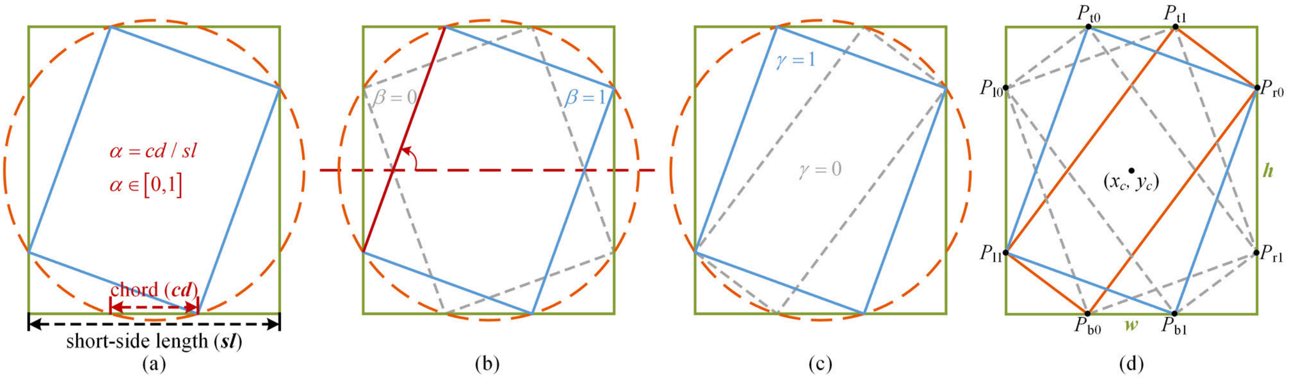

**FIGURE 4.** Definition of the CCH. (a) Definition of $\alpha$. $\alpha$ reflects the size of the concentric circle radius whose domain is [0, 1]. *cd* denotes the chord length of the circle cut by the short side of the h-box (if the circle has intersections with the short side, it must have intersections with the long side). *sl* denotes the short-side length of the h-box. In order to achieve normalization, we define the ratio of *cd* and *sl* as $\alpha$. (b) Definition of $\beta$. $\beta$ represents the inclination direction of the r-box, which is binary. We define $\beta$ as 1 (solid line) when the angle between the long side of the r-box and the positive x-axis is an acute angle (we define the counterclockwise angle as positive); otherwise, it is 0 (dashed line). (c) Definition of $\gamma$. $\gamma$ represents the relative area size of the two r-boxes in the same inclination direction, which is binary. We define the larger as 1 (solid line) and the smaller as 0 (dashed line). (d) All 4 r-boxes in the general case.

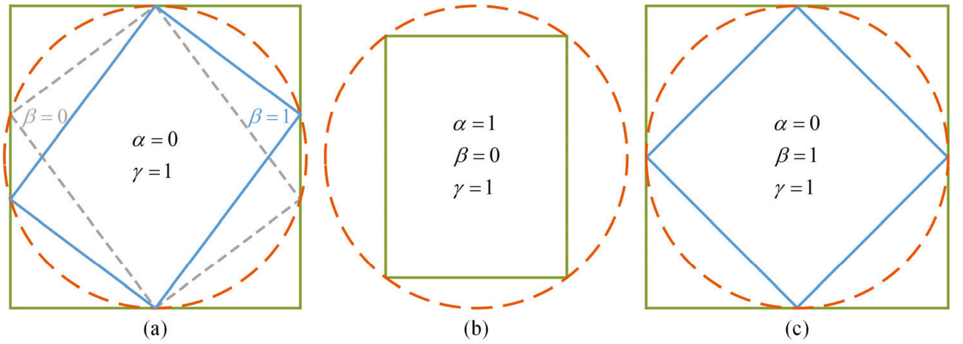

**FIGURE 5.** Special cases of the CCH. (a) The circle is tangent to the short side of the h-box. At this time, $\alpha = 0$, and we define $\gamma = 1$. (b) The circle is the circumcircle of the h-box. At this time, $\alpha = 1$, and we define $\beta = 0$ and $\gamma = 1$. (c) $w = h$ and the circle is tangent to the h-box. At this time, $\alpha = 0$, and we define $\beta = 1$ and $\gamma = 1$.

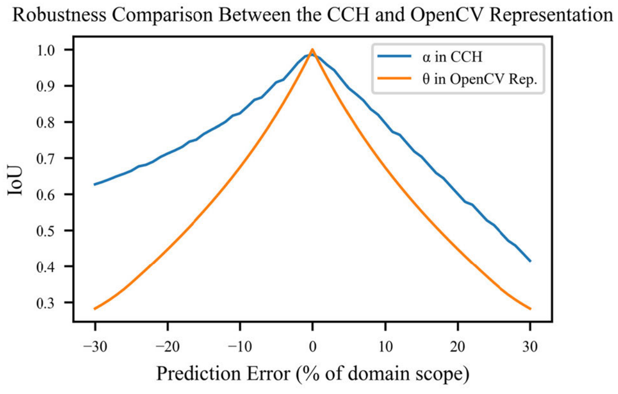

**FIGURE 6.** Robustness comparison between the CCH and OpenCV representation. The horizontal axis represents the prediction error ranges of $\alpha$ and $\theta$, which are within plus or minus 30% of their respective domain scopes, and our sampling interval is 1%. The vertical axis represents the IoUs between the deviation results and ground truths under the CCH and OpenCV representations.

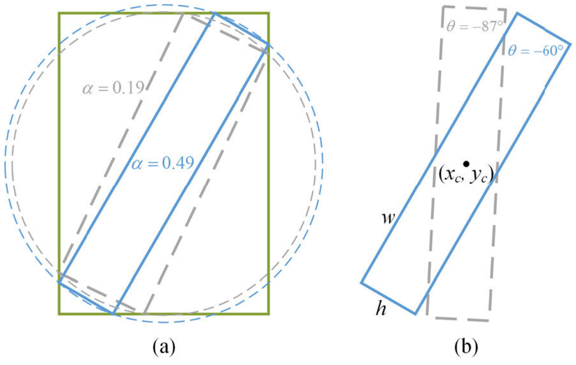

**FIGURE 7.** Deviation results and ground truths when the prediction errors of $\alpha$ and $\theta$ are −30% of their domain scopes. (a) CCH. $\alpha$ of the ground truth (solid line) is approximately 0.49, and $\alpha$ of the deviation result (dashed line) is approximately 0.19 (the domain scope of $\alpha$ is 1). (b) OpenCV representation. $\theta$ of the ground truth (solid line) is −60 degrees, and $\theta$ of the deviation result (dashed line) is −87 degrees (the domain scope of $\theta$ is 90 degrees).

prediction errors are −30% of their domain scopes, which is more intuitive.

3) As mentioned above, since there are no constraints between the vertices of an arbitrary quadrilateral, the

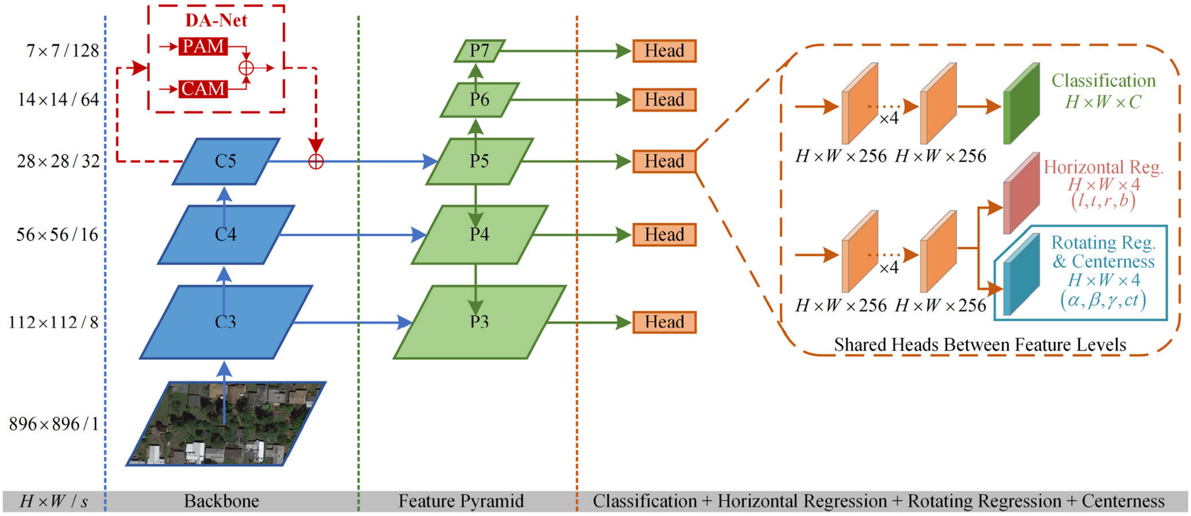

**FIGURE 8.** The network architecture of the SAR, where C3, C4, and C5 denote the feature maps of the backbone, and P3 to P7 are the feature levels of the feature pyramid. $H \times W$ are the height and width of the feature maps, respectively. $/s$ is the downsampling ratio of the feature maps to the level of the input image. The input size is 896 × 896 in this article.

prediction of the vertex-based method may be very irregular. However, the prediction of the CCH is a rectangle, which makes the result regular. Moreover, the CCH representation does not depend on the vertices, and the regression parameters do not need to be paired as in [25], thus avoiding the confusion of the order problem.

4) CCH only adds a very small number of parameters on the head of the network without introducing any hyperparameters. And as stated above, the addition of the rotating regression to classic benchmark is almost free of cost. Taking the FCOS [26] as the benchmark to adopt the CCH, the additional number of parameters is only 11.5k, and the additional number of lines of code is less than 100 lines. Moreover, as the effect of CCH directly depends on the horizontal prediction, it can inherit the improvements of the classic benchmark, including the network structure, loss function, and training tricks.

However, the CCH also has some shortcomings. First, the CCH requires the bounding box to be a standard rectangle. This description may not be as flexible and appropriate as the vertex-based representation for some objects (e.g., the bounding box of some objects is obviously more suitable to be represented by a parallelogram). Second, the CCH is actually an implicit representation that needs to obtain the final result through analytical geometric calculations. Truncation errors will occur during the calculation process, and precision loss will occur due to the data type conversion (e.g., double to float). As shown in Fig. 6, note that the IoU of the CCH should be 1 when the prediction error is 0, and the slight error here comes from the above situation.

## B. SINGLE-STAGE ANCHOR-FREE ROTATING (SAR) OBJECT DETECTION NETWORK

### 1) NETWORK STRUCTURE

Inspired by the FCOS [26] benchmark, we propose a single-stage anchor-free rotating (SAR) object detection network based on the CCH, as shown in Fig. 8. The SAR makes two main changes to the FCOS:

a) The head. In the original FCOS paper, the centerness layer is in the classification branch. In the SAR baseline, we move it to the regression branch (the update of the official project of the FCOS states that this can improve accuracy) and add rotating parameter predictions ($\alpha$, $\beta$, $\gamma$) to this layer, where $\alpha$ and the centerness are continuous normalized variables, and $\beta$ and $\gamma$ are binary variables. Finally, the 4 parameters of this layer can be output through a sigmoid layer.

b) The backbone. The complexity of remote sensing images leads to a large amount of background noise in their feature maps, which affects the effect of feature extraction. Therefore, we consider adding an attention mechanism based on the DA-Net [27] to the backbone to reduce the background noise and enhance the feature extraction. The original paper removes the downsampling operations after C3 and employs dilated convolutions in the last two ResNet [28] blocks so that the stride of C5 is 8. We do not adopt the above operations in this article; therefore, the stride of C5 is still 32. C5 has gone through the position attention module (PAM) and channel attention module (CAM) of the DA-Net block, and the outputs are added elementwise; then, the result is added to C5 elementwise to obtain P5.

### 2) NETWORK OUTPUTS

$F_i$ are the feature maps at layer $i$ of the feature pyramid (P3-P7 in Fig. 8). The SAR obtains the final output after $F_i$ passes the shared head. In the baseline, the output layer contains three branches: a classification branch, a horizontal regression branch and a rotating regression & centerness branch.

a) Classification branch. Each location on this layer is a $C$-dimensional vector $\boldsymbol{p}$, where C denotes the number of classes.

b) Horizontal regression branch. Each location on this layer is a 4D horizontal parameter vector $\boldsymbol{th} = (l, t, r, b)$, where $l$, $t$, $r$ and $b$ are the distances from the location to the four respective sides of the h-box, which is the same as the definition in the FCOS.

c) Rotating regression & centerness branch. Each location on this layer is a 4D vector $\boldsymbol{tr} = (\alpha, \beta, \gamma, ct)$, where $(\alpha, \beta, \gamma)$ are the rotating parameters in the CCH and $ct$ is the same as the definition of centerness in the FCOS. Note that we use the generalized focal loss [29] in the subsequent ablation experiments; at that time, the branch no longer contains the centerness and vector $\boldsymbol{tr} = (\alpha, \beta, \gamma)$.

### 3) LOSS FUNCTION

We define the loss function as follows:

$$L = \frac{1}{N_{pos}} \sum_{x,y} L_{cls}\left(\boldsymbol{p}_{x,y}, c^*_{x,y}\right) + \frac{\lambda}{N_{pos}} \sum_{x,y} \mathbb{1}_{\left\{c^*_{x,y}>0\right\}} \times \left[L_{hreg}\left(\boldsymbol{th}_{x,y}, \boldsymbol{th}^*_{x,y}\right) + L_{rreg}\left(\boldsymbol{tr}_{x,y}, \boldsymbol{tr}^*_{x,y}\right)\right] \quad (1)$$

where $(x, y)$ denotes each location on $F_i$, which is the same as the definition in the FCOS regarding it is a positive or negative sample. $N_{pos}$ denotes the number of positive samples. $c^*$ is the class label when the location is a positive sample, otherwise $c^* = 0$ when the location is a negative sample. $L_{cls}$ denotes the classification loss. For comparison, we use the focal loss [18] and generalized focal loss [29] as $L_{cls}$, respectively. $L_{hreg}$ denotes the horizontal regression loss, where $\boldsymbol{th}^*$ is the horizontal regression target; and we use Giou [30] as $L_{hreg}$. $L_{rreg}$ denotes the summation of the rotating regression loss and centerness loss, where $\boldsymbol{tr}^*$ is the rotating regression and centerness target; and we use *BCEWithLogitsLoss* in PyTorch to calculate $L_{rreg}$. $\mathbb{1}_{\left\{c^*_{x,y}>0\right\}}$ is the indicator function, which is 1 if $c^* > 0$ and 0 otherwise. $\lambda$ is the balance weight, which is 1 in this article.

## IV. EXPERIMENTS

### A. DATA SETS AND EVALUATION PROTOCOLS

DOTA [31] is a large-scale object detection data set that contains 2,806 aerial images with sizes ranging from approximately $800 \times 800$ to $4,000 \times 4,000$ pixels. DOTA uses arbitrary quadrilaterals to annotate 188,282 instances in 15 classes: plane (PL), ship (SH), storage tank (ST), baseball diamond (BD), tennis court (TC), basketball court (BC), ground track field (GTF), harbor (HA), bridge (BR), small vehicle (SV), large vehicle (LV), helicopter (HC), roundabout (RA), soccer ball field (SBF), and swimming pool (SP). We crop the images into $896 \times 896$ patches by a sliding window with a step size of 512, and we do not use the instances marked as difficult. We finally obtain 25,696 patches, and we randomly divide 1/2, 1/6 and 1/3 of the patches into training, validation and test sets, respectively. DOTA adopts mAP as the official evaluation protocol.

HRSC2016 [32] is a data set dedicated to ship detection that contains 1,061 aerial images downloaded from Google Earth. Its ground truths are annotated by using a rotating rectangle. We adopt the original division to obtain the training, validation and test sets with 436, 181 and 444 images, respectively. We use mAP as the evaluation protocol of HRSC2016.

ICDAR2015 [33] is a scene text detection data set and challenge that annotates the ground truths by an arbitrary quadrilateral. We adopt the original division to obtain the training and test sets with 1,000 and 500 images, respectively, and we randomly take 30% of the training set as the validation set. The official evaluation protocol of the challenge is the F-measure.

RCTW-17 [34] is a competition on reading Chinese text in images. It provides a large-scale data set with 8,033 images whose ground truths are annotated by an arbitrary quadrilateral. We randomly divide 1/2, 1/6, and 1/3 of the images as the training, validation and test sets, respectively, and we do not use the instances marked as difficult. The official evaluation protocol of the challenge is the F-measure.

### B. IMPLEMENTATION DETAILS

#### 1) GROUND TRUTH GENERATION

We refer to Fig. 4 (d) and follow the steps below to generate the ground truth labels:

a) If the annotation is vertex-based, obtain the minimum enclosing rectangle (r-box) of the vertices (one can use the methods in libraries such as OpenCV); if the annotation is angle-based, it is an r-box.

b) Obtain the horizontal minimum enclosing rectangle of the r-box (which is the h-box), and the center point of the h-box is $(x_c, y_c)$.

c) The intersections of the r-box and h-box are denoted as $P_{\mathrm{l}}$, $P_{\mathrm{t}}$, $P_{\mathrm{r}}$, and $P_{\mathrm{b}}$. If $P_{\mathrm{l}}$ or $P_{\mathrm{r}}$ is above the center point, it is recorded as 0; otherwise, it is recorded as 1. If $P_{\mathrm{t}}$ or $P_{\mathrm{b}}$ is to the left of the center point, it is recorded as 0; otherwise, it is recorded as 1.

d) The calculation of $\alpha$ is as follows:

$$\alpha = \begin{cases} 2\left|p_t x - x_c\right|/w, & w < h \\ 2\left|p_l y - y_c\right|/h, & w \geq h \end{cases} \quad (2)$$

where $p_t x$ and $p_l y$ represent the abscissa and ordinate of $P_{\mathrm{t}}$ and $P_{\mathrm{l}}$, respectively.

e) Determine $\beta$ and $\gamma$ according to the 0-1 order of the r-box vertices.

The ground truths in special cases are the same as the description in Fig. 5.

#### 2) TRAINING DETAILS

In the training phase, we use the model pretrained on COCO [40] provided by the official FCOS, which uses ResNet101 [28] as the backbone, and we initialize the newly added layers as in [18]. We train for 60 epochs on all data sets with a minibatch of 16 images. Stochastic gradient descent (SGD) is used as an optimizer, and its weight decay and momentum are set as 0.0001 and 0.9, respectively. The initial learning rate is 0.001 and is reduced tenfold after 36 epochs

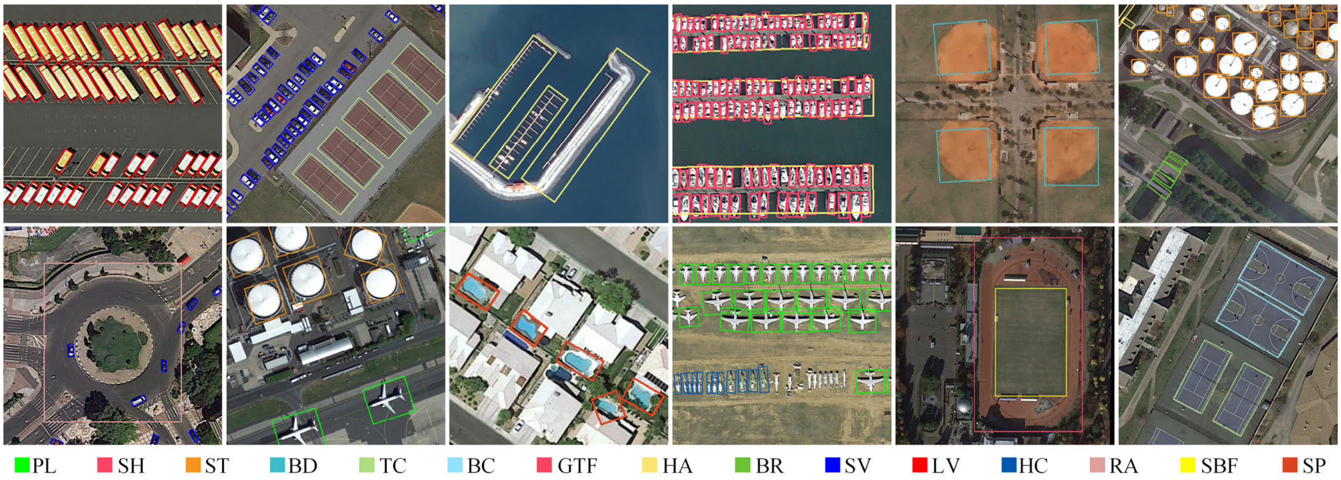

**FIGURE 9.** Several detection results of the SAR are conducted on DOTA [31]. We show all detected objects with classification scores above 0.6.

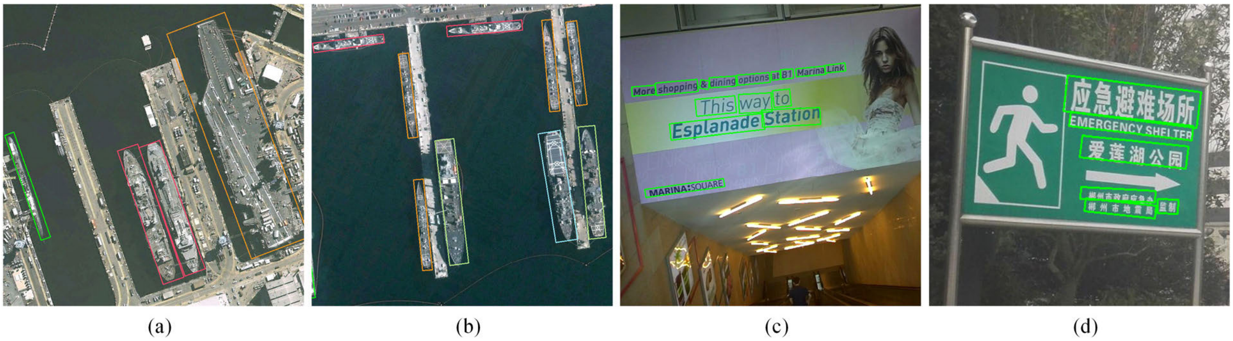

**FIGURE 10.** Several detection results of the SAR are conducted on HRSC2016 [32] in (a)-(b), ICDAR2015 [33] in (c), and RCTW-17 [34] in (d). We show all detected objects with classification scores above 0.6.

and 48 epochs. We use 2 TITAN RTX GPUs (each with 24 GB memory) in our experiments.

### 3) INFERENCE DETAILS

Although $\alpha$ in the CCH is continuous without any loss of precision, however, due to the truncation errors through analytical geometric calculations and the precision loss due to the data type conversion mentioned above, it may cause a negative square root when solving the intersections between the circle and rectangle in special cases. Therefore, we set $\alpha$ to 0 when it is less than 0.05 and set it to 1 when it is greater than 0.95 in the testing phase ($\beta$ and $\gamma$ are determined according to $\alpha$ at that time). The inference generates a set of $(l, t, r, b, \alpha, \beta, \gamma)$. We obtain the h-box according to $(l, t, r, b)$, obtain the radius of the circle according to the h-box and $\alpha$, and then obtain 8 intersections between the circle and h-box. Finally, a rotating rectangle is uniquely determined based on $\beta$ and $\gamma$. The process only involves a basic analytical geometry method, so we do not elaborate on it here.

## *C. RESULTS*

### 1) OBJECT DETECTION IN AERIAL IMAGES

The test results of the SAR on DOTA and HRSC2016 are depicted in Fig. 9 and Figs. 10 (a)-(b), respectively. The SAR is able to identify and locate objects in the complex background of aerial images accurately, even for objects with large aspect ratios and dense small objects. The quantitative comparisons on DOTA and HRSC2016 are illustrated in Table 1 and Table 2, respectively. In terms of accuracy, on DOTA, the mAP of the SAR baseline (FCOS+CCH with ResNet101 as the backbone) is 72.12%, and it can reach 73.48% with an attention mechanism and the generalized focal loss. Among the methods that are better than the SAR in the table, Glid. Ver. [25] is based on the two-stage anchor-based benchmark Faster RCNN [19], and R$^3$Det [3] and RSDet [23] are based on the single-stage anchor-based benchmark RetinaNet [18] while using a deeper backbone (ResNet152 [28]). For comparison, we also use above two benchmarks combined with the CCH for experiments. Specifically, for the RetinaNet, we add the CCH head as the rotating regression branch to its head, and the network architecture is almost the same as in Fig. 8 except that there is no centerness in the rotating regression branch. We implement this based on the FCOS benchmark project, and ResNet152 is used as the backbone. Other corresponding implementation details are the same as in the SAR baseline. The mAP of RetinaNet+CCH is 74.17%, which surpasses the methods [3] and [23]. For the Faster RCNN, we also add the CCH head as the rotating regression branch to its head, as shown in Fig. 11;

**TABLE 1.** Quantitative Comparison with Related Works on DOTA.

| Method | PL | SH | ST | BD | TC | BC | GTF | HA | BR | SV | LV | HC | RA | SBF | SP | mAP | FPS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FR-O$^{(II)(A)}$ [31] | 79.09 | 36.20 | 58.96 | 69.12 | 89.19 | 69.60 | 63.49 | 46.69 | 17.17 | 34.20 | 37.16 | 46.30 | 52.52 | 49.4 | 44.80 | 52.93 | - |
| IENet [35] | 80.20 | 52.58 | 78.51 | 64.54 | 81.45 | 44.66 | 32.07 | 64.40 | 39.82 | 49.71 | 65.01 | 36.75 | 56.73 | 46.54 | 64.24 | 57.14 | - |
| R-DFPN$^{(II)(A)}$ [4] | 80.92 | 54.78 | 68.66 | 65.82 | 90.33 | 66.34 | 58.94 | 55.10 | 33.77 | 55.77 | 50.94 | 35.88 | 51.76 | 48.73 | 51.32 | 57.94 | - |
| R$^2$CNN$^{(II)(A)}$ [6] | 80.94 | 55.81 | 72.39 | 65.67 | 90.67 | 66.92 | 67.44 | 55.14 | 35.34 | 59.92 | 50.91 | 48.22 | 52.23 | 55.06 | 53.35 | 60.67 | - |
| RRPN$^{(II)(A)}$ [10] | 88.52 | 57.25 | 67.38 | 71.20 | 90.81 | 72.84 | 59.30 | 53.08 | 31.66 | 51.85 | 56.19 | 53.58 | 52.84 | 56.69 | 51.94 | 61.01 | - |
| ICN$^{(II)(A)}$ [1] | 81.40 | 70.00 | 78.20 | 74.30 | 90.80 | 79.10 | 70.30 | 67.00 | 47.70 | 64.90 | 67.80 | 50.20 | 62.90 | 53.60 | 64.20 | 68.20 | - |
| RADet$^{(II)(A)(X)}$ [36] | 79.45 | 68.86 | 74.97 | 76.99 | 89.70 | 78.14 | 65.83 | 66.14 | 48.05 | 65.46 | 74.40 | 62.16 | 64.63 | 49.92 | **71.58** | 69.09 | - |
| RoI-Trans.$^{(II)(A)}$ [2] | 88.64 | 83.59 | 81.46 | 78.52 | 90.74 | 77.27 | 75.92 | 62.83 | 43.44 | 68.81 | 73.68 | 47.67 | 53.54 | 58.39 | 58.93 | 69.56 | 5.9 |
| CADNet$^{(II)(A)}$ [37] | 87.80 | 76.70 | 73.30 | 82.40 | 90.90 | 79.20 | 73.50 | 62.00 | 49.40 | 71.10 | 63.50 | 62.20 | 60.90 | 48.40 | 67.00 | 69.90 | - |
| SCRDet$^{(II)(A)}$ [5] | 89.98 | 72.41 | 86.86 | 80.65 | 90.85 | 87.94 | 68.36 | 66.25 | 52.09 | 68.36 | 60.32 | 65.21 | 66.68 | **65.02** | 68.24 | 72.61 | - |
| R$^3$Det$^{(A)(152)}$ [3] | 89.49 | 78.21 | 84.23 | 81.17 | 90.81 | 85.26 | 66.10 | 68.16 | 50.53 | 70.92 | 78.66 | **67.17** | 63.77 | 61.81 | 69.83 | 73.74 | - |
| RSDet$^{(A)(152)}$ [23] | 90.10 | 73.60 | 84.70 | 82.00 | **91.20** | 87.10 | 68.50 | 66.10 | 53.80 | 70.20 | **78.70** | 63.70 | 68.20 | 64.30 | 69.30 | 74.10 | - |
| Glid. Ver.$^{(II)(A)}$ [25] | 89.64 | **86.82** | 86.81 | **85.00** | 90.74 | 79.02 | **77.34** | 72.94 | 52.26 | **73.01** | 73.14 | 57.32 | **70.91** | 59.55 | 70.86 | 75.02 | 10 |
| **SAR baseline** | 90.37 | 79.18 | 79.10 | 82.53 | 86.46 | 86.98 | 69.18 | 65.57 | 49.30 | 67.48 | 71.63 | 61.73 | 65.42 | 59.75 | 67.07 | 72.12 | 16.53 |
| **SAR$^{*}$** | 90.45 | 80.03 | 80.37 | 83.08 | 87.60 | 87.61 | 69.33 | 66.22 | 50.87 | 68.55 | 73.55 | 62.15 | 65.52 | 60.94 | 67.38 | 72.91 | 15.50 |
| **SAR$^{\dagger}$** | 90.19 | 80.91 | 81.64 | 82.16 | 88.51 | **89.04** | 70.15 | 66.14 | 49.23 | 66.95 | 72.90 | 63.79 | 65.99 | 61.47 | 67.58 | 73.11 | **16.56** |
| **SAR$^{*\dagger}$** | **90.89** | 81.24 | 82.43 | 82.67 | 90.27 | 88.19 | 69.90 | 66.20 | 49.75 | 68.07 | 72.31 | 63.48 | 66.43 | 62.08 | 68.34 | 73.48 | 15.53 |
| **RetinaNet+CCH$^{(A)(152)}$** | 89.77 | 80.92 | 83.90 | 82.01 | 90.92 | 81.45 | 72.59 | 72.01 | 50.38 | 71.61 | 74.01 | 64.55 | 66.07 | 62.45 | 69.91 | 74.17 | 11.90 |
| **Faster RCNN+CCH$^{(II)(A)}$** | 89.67 | 84.63 | **87.07** | 79.78 | 90.91 | 88.22 | 68.29 | **75.13** | **54.17** | 71.70 | 77.90 | 64.29 | 66.95 | 60.49 | 70.01 | **75.28** | 10.44 |

*(II)* indicates that the method is two-stage. *(A)* indicates that the method is anchor-based. (*X*) indicates that the backbone network is ResNeXt101 [38]. (*152*) indicates that the backbone network is ResNet152 [28]. The other backbone networks are ResNet101 [28] without specified. * indicates that attention mechanism DA-Net [27] block is used. † indicates that generalized focal loss [29] is used.

**TABLE 2.** Quantitative Comparison with Related Works on HRSC2016.

| Method | Backbone | Input Size | Device | mAP | FPS |
|---|---|---|---|---|---|
| R$^2$CNN [6] | ResNet101 | 800×800 | K80 | 73.07 | 2 |
| RC1 & RC2 [32] | VGG16 | 800×800 | - | 75.70 | <1 |
| RRPN [10] | ResNet101 | 800×800 | - | 79.08 | 3.5 |
| R$^2$PN [39] | VGG16 | - | - | 79.60 | <1 |
| RRD [8] | VGG16 | 384×384 | - | 84.30 | - |
| RoI-Trans. [2] | ResNet101 | 512×800 | - | 86.20 | 6 |
| Glid. Ver. [25] | ResNet101 | - | TITAN Xp | 88.20 | - |
| R$^3$Det [3] | ResNet152 | 800×800 | 2080Ti | **89.33** | 10 |
| **SAR baseline** | ResNet101 | 896×896 | TITAN RTX | 88.11 | **16.52** |
| **SAR$^{*\dagger}$** | ResNet101 | 896×896 | TITAN RTX | 88.45 | 15.53 |

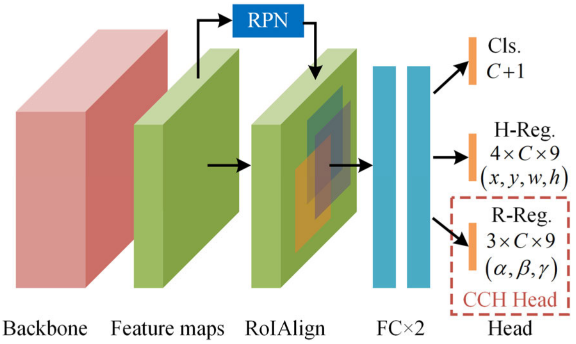

**FIGURE 11.** Framework of Faster RCNN with the CCH. Cls. and H-reg. refer to the classifier and regressor in the original head, respectively. *C* is the number of classes, and 9 is the number of anchors. We implement this based on the "*maskrcnn-benchmark*" project (https://github.com/facebookresearch/maskrcnn-benchmark).

and the backbone is ResNet101. The rotating parameters $(\alpha, \beta, \gamma)$ are trained with the binary cross entropy (BCE) loss, and the loss is added to the loss function in [19]. Other corresponding implementation details are the same as in [19] and the SAR baseline, respectively. The mAP of Faster RCNN+CCH is 75.28%, which is superior to the state-of-the-art method [25]. On HRSC2016, the SAR achieves an 88.45% mAP, and the only superior comparative method R$^3$Det [3] uses a deeper backbone (ResNet152). In terms of speed, the SAR achieves state-of-the-art performance among the comparative methods. Without any parallel acceleration technology, the average test time is 60.5 ms per image (16.53 FPS) with the SAR baseline and 64.5 ms per image (15.50 FPS) with an attention mechanism when the input size is 896 × 896. Note that the runtime for the NMS and coordinate calculation of the CCH is included, while the runtime for data loading is not. From Table 2, it can be seen that our input image size is the largest. However, compared with the fastest method, the SAR boosts the speed by more than 60%. The speed advantage of the SAR is mainly due to two points. First, the single-stage anchor-free benchmark FCOS, whose characteristics make the horizontal detection part fast, is the main reason for our speed advantage. Second, the CCH head almost does not change any strategy or process of the FCOS, and it is quite lightweight. The official test time of the FCOS is 57 ms with the input images resized to have their shorter side as 800 pixels and their longer side less than or equal to 1333 pixels, and the tests use an NVIDIA Tesla V100 GPU. Thus, excluding the hardware difference, the CCH only adds a small amount of overhead to the algorithm.

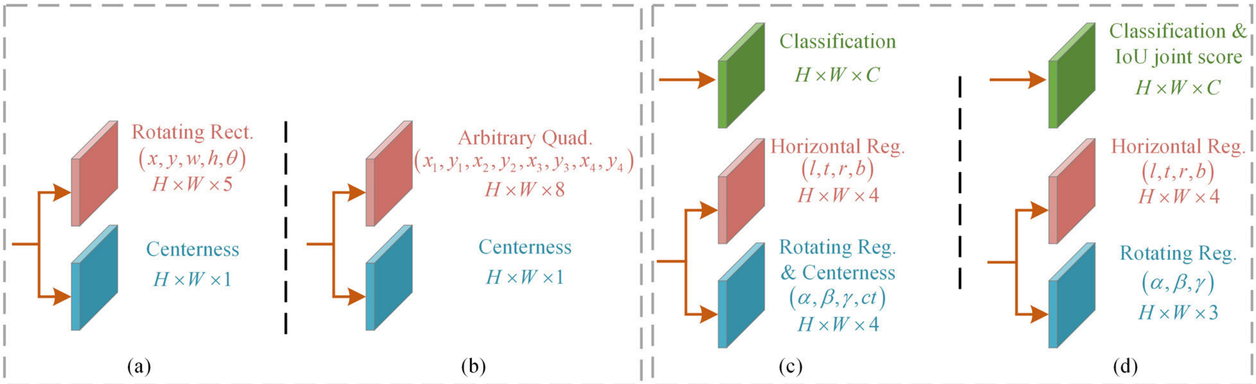

**FIGURE 12.** Regression branch of different representations: the 5-parameter rotating rectangle is in (a), and the 8-parameter arbitrary quadrilateral is in (b). Heads of different classification loss functions: the focal loss is in (c), and the generalized focal loss is in (d).

**TABLE 3.** Quantitative Comparison with Related Works on ICDAR2015.

| Method | Precision | Recall | F-score | FPS |
|---|---|---|---|---|
| CTPN [41] | 74.22 | 51.56 | 60.85 | - |
| SegLink [34] | 73.10 | 76.80 | 75.00 | - |
| RRPN [10] | 73.23 | 82.17 | 77.44 | <1 |
| EAST [11] | 83.27 | 78.33 | 80.72 | 13.2 |
| R$^2$CNN [6] | 85.62 | 79.68 | 82.54 | <1 |
| R$^3$Det [3] | **86.43** | 83.54 | 84.96 | 13.5 |
| FOTS [9] | 85.17 | **91.00** | **87.99** | 7.8 |
| **SAR baseline** | 83.16 | 81.50 | 82.32 | **14.80** |
| **SAR**$^{*\dagger}$ | 84.35 | 83.08 | 83.71 | 13.91 |

**TABLE 4.** Quantitative Comparison with Related Works on RCTW-17.

| Method | Precision | Recall | F-score | FPS |
|---|---|---|---|---|
| Official baseline [43] | 76.00 | 40.40 | 52.80 | 8.9 |
| RRD [8] | 72.40 | 45.30 | 55.70 | 10.0 |
| LOMO [21] | **80.40** | 50.80 | 62.30 | 4.4 |
| Glid. Ver. [25] | 77.00 | 61.00 | 68.10 | 7.8 |
| **SAR baseline** | 78.08 | 60.22 | 68.00 | **16.50** |
| **SAR**$^{*\dagger}$ | 78.79 | **61.40** | **69.02** | 15.55 |

**TABLE 5.** Quantitative Comparison of the Three Representations with the FCOS Benchmark on DOTA.

| Representation | mAP | FPS | Code lines |
|---|---|---|---|
| CCH | **72.12** | **16.53** | **78** |
| 5-param | 63.72 | 16.10 | 599 |
| 8-param | 58.65 | 15.83 | 532 |

### 2) SCENE TEXT DETECTION

The test results of the SAR for oriented scene text detection on ICDAR2015 are shown in Fig. 10 (c), and the quantitative illustrations are given in Table 3. We do not perform data augmentation or use a pretrained model on similar data sets, and we obtain an F-score of 83.71%. On RCTW-17, the test results are shown in Fig. 10 (d), and the quantitative illustrations are given in Table 4. We achieve an F-score of 69.02%, which is the best among several comparative methods. The above results show that the SAR is able to detect scene texts of arbitrary orientations correctly; in addition, the SAR is effective in both aerial image and natural scene image detection.

## *D. ABLATION STUDY*

### 1) CCH VS. ROTATING RECTANGLE & ARBITRARY QUADRILATERAL REPRESENTATIONS

We conduct a comparison of the regressions using the CCH, rotating rectangle and arbitrary quadrilateral representations on DOTA. For the latter two, we still only modify the regression branch of the FCOS head, as shown in Figs. 12 (a)-(b). The training details are the same as in the CCH, and we use the smooth L1 loss as the loss function. A comparison of the three is shown in Fig. 13 and Table 5. The CCH clearly surpasses the other two methods. In addition, the objects in Fig. 13 are almost horizontal and vertical, which confirms that the CCH solves the boundary problem effectively. Prediction near the boundary condition is the difficulty of rotating object detection, while improper representations will cause a sharp increase in the loss and bring unnecessary difficulties to the training. In other words, a proper representation is the fundamental way to solve the boundary problem. Moreover, the last column in the table refers to the amount of code that needs to be modified in the benchmark when using the representation. Due to the change in the regression parameters, rotating rectangle and arbitrary quadrilateral representations have to modify many details to adapt to the original process and strategy. In contrast, the CCH saves plenty of work.

### 2) WITH OR WITHOUT ATTENTION

As shown in Table 1, the use of the DA-Net block increases the mAP by 0.79%, confirming that the attention mechanism can indeed improve the detector's performance in remote sensing images. We believe that this is due to the reduction in complex background noise as well as the enhanced foreground feature extraction.

### 3) EFFECT OF GENERALIZED FOCAL LOSS

The results in Table 1 show that replacing the focal loss with the generalized focal loss [29] can increase the mAP

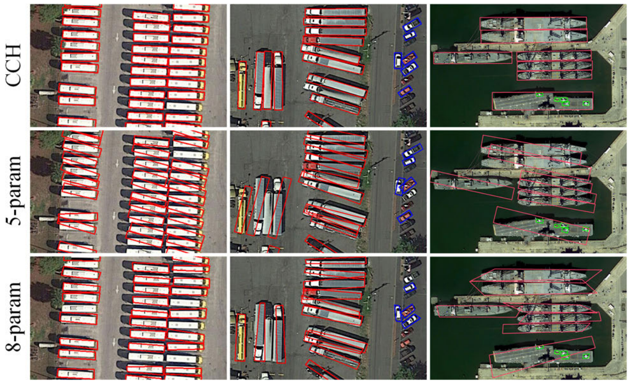

**FIGURE 13. Comparison of the detection results of the three representations. We show all detected objects with classification scores above 0.6.**

of baseline by 0.99%. Indeed, we only adopt the quality focal loss from [29], which merges the classification loss and regression quality estimation loss into one and does not change the other parts of the loss calculation in the baseline. Specifically, we remove the regression quality estimation (centerness) from the rotating regression branch and merge it into the classification branch, as illustrated in Figs. 12 (c)-(d). Note that in the generalized focal loss ablation experiment, the regression quality estimation is the IoU instead of the centerness, which is the same as in [29]. It is more reasonable to make the rotating regression branch focus only on the rotating parameter prediction, and we believe that this is one of the reasons for the improvement, in addition to the analysis in [29].

## V. CONCLUSION

In this article, we propose a novel representation to effectively address the issues common in angle-based and vertex-based representations of rotating object detection. On this basis, we propose a lightweight head that can add the rotating regression to classic object detection benchmarks in an almost cost-free manner and build a single-stage anchor-free rotating object detection baseline. Without bells and whistles (e.g., cascade refinement or the NAS-FPN), our method achieves accurate and robust performance in aerial image and scene text detection while being rather fast. In follow-up research, we would like to explore the combination of our method with the latest innovations of classic benchmarks, including the breakthroughs in backbones, attention mechanisms, cascade refinements, feature alignment, and so on.

## REFERENCES

[1] S. M. Azimi, E. Vig, R. Bahmanyar, M. Körner, and P. Reinartz, "Towards multi-class object detection in unconstrained remote sensing imagery," in *Proc. Asian Conf. Comput. Vis.*, Perth, WA, Australia, Jun. 2018, pp. 150–165.

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

[3] X. Yang, Q. Liu, J. Yan, A. Li, Z. Zhang, and G. Yu, "R3Det: Refined single-stage detector with feature refinement for rotating object," 2019, *arXiv:1908.05612*. [Online]. Available: http://arxiv.org/abs/1908.05612

[4] X. Yang, H. Sun, K. Fu, J. Yang, X. Sun, M. Yan, and Z. Guo, "Automatic ship detection in remote sensing images from Google Earth of complex scenes based on multiscale rotation dense feature pyramid networks," *Remote Sens.*, vol. 10, no. 1, p. 132, Jan. 2018.

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

[6] Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, P. Fu, and Z. Luo, "R2CNN: Rotational region CNN for orientation robust scene text detection," 2017, *arXiv:1706.09579*. [Online]. Available: http://arxiv.org/abs/1706.09579

[7] M. Liao, B. Shi, and X. Bai, "TextBoxes++: A single-shot oriented scene text detector," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3676–3690, Aug. 2018.

[8] M. Liao, Z. Zhu, B. Shi, G.-S. Xia, and X. Bai, "Rotation-sensitive regression for oriented scene text detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5909–5918.

[9] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, and J. Yan, "FOTS: Fast oriented text spotting with a unified network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5676–5685.

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

[11] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "EAST: An efficient and accurate scene text detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5551–5560.

[12] C. Huang, H. Ai, Y. Li, and S. Lao, "High-performance rotation invariant multiview face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 671–686, Apr. 2007.

[13] H. A. Rowley, S. Baluja, and T. Kanade, "Rotation invariant neural network-based face detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1998, pp. 38–44.

[14] X. Shi, S. Shan, M. Kan, S. Wu, and X. Chen, "Real-time rotation-invariant face detection with progressive calibration networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2295–2303.

[15] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Barcelona, Spain, 2016, pp. 379–387.

[16] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[17] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[20] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Multi-oriented and multi-lingual scene text detection with direct regression," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5406–5419, Nov. 2018.

[21] C. Zhang, B. Liang, Z. Huang, M. En, J. Han, E. Ding, and X. Ding, "Look more than once: An accurate detector for text of arbitrary shapes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10552–10561.

[22] Y. Liu, S. Zhang, L. Jin, L. Xie, Y. Wu, and Z. Wang, "Omnidirectional scene text detection with sequential-free box discretization," 2019, *arXiv:1906.02371*. [Online]. Available: http://arxiv.org/abs/1906.02371

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

[24] X. Yang and J. Yan, "Arbitrary-oriented object detection with circular smooth label," 2020, *arXiv:2003.05597*. [Online]. Available: http://arxiv.org/abs/2003.05597

[25] Y. Xu, M. Fu, Q. Wang, Y. Wang, K. Chen, G.-S. Xia, and X. Bai, "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Feb. 18, 2020, doi: 10.1109/TPAMI.2020.2974745.

[26] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9627–9636.

[27] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[29] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," 2020, *arXiv:2006.04388*. [Online]. Available: http://arxiv.org/abs/2006.04388

[30] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Los Angeles, CA, USA, Jun. 2019, pp. 658–666.

[31] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.

[32] Z. Liu, H. Wang, L. Weng, and Y. Yang, "Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 8, pp. 1074–1078, Aug. 2016.

[33] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, and E. Valveny, "ICDAR 2015 competition on robust reading," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015, pp. 1156–1160.

[34] B. Shi, X. Bai, and S. Belongie, "Detecting oriented text in natural images by linking segments," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2550–2558.

[35] Y. Lin, P. Feng, and J. Guan, "IENet: Interacting embranchment one stage anchor free detector for orientation aerial object detection," 2019, *arXiv:1912.00969*. [Online]. Available: http://arxiv.org/abs/1912.00969

[36] Y. Li, Q. Huang, X. Pei, L. Jiao, and R. Shang, "RADet: Refine feature pyramid network and multi-layer attention network for arbitrary-oriented object detection of remote sensing images," *Remote Sens.*, vol. 12, no. 3, p. 389, Jan. 2020.

[37] G. Zhang, S. Lu, and W. Zhang, "CAD-net: A context-aware detection network for objects in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 10015–10024, Dec. 2019.

[38] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 1492–1500.

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

[40] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Zurich, Switzerland, 2014, pp. 740–755.

[41] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, 2016, pp. 56–72.

[42] B. Shi, C. Yao, M. Liao, M. Yang, P. Xu, L. Cui, S. Belongie, S. Lu, and X. Bai, "ICDAR2017 competition on reading chinese text in the wild (RCTW-17)," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, Nov. 2017, pp. 1429–1434.

**JUNYAN LU** was born in Yanbian, Jilin, China, in 1990. He received the B.E. and master's degrees from Beihang University, Beijing, China, in 2012 and 2015, respectively. He is currently pursuing the Ph.D. degree with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun.

His research interests include computer vision, object detection, and remote sensing image interpretation.

**TIE LI** was born in Daqing, Heilongjiang, China, in 1989. He received the B.E. and master's degrees from Beihang University, Beijing, China, in 2012 and 2015, respectively.

He is currently with the Shanghai Electro-Mechanical Engineering Institute. His research interests include computer vision, object detection, and remote sensing image processing.

**JINGYU MA** was born in Kaifeng, Henan, China, in 1992. He received the B.E. degree from Wuhan University, Wuhan, China, in 2013, and the master's degree from the University of Chinese Academy of Sciences, Beijing, in 2017.

He is currently with Chang Guang Satellite Technology Company Ltd. (CGSTL). His research interests include image classification and instance segmentation in remote sensing imagery.

**ZHUQIANG LI** received the B.S. degree from the China University of Geosciences, Beijing, in 2014, and the master's degree from the School of Geography, Beijing Normal University, Beijing, in 2017.

He is currently with Chang Guang Satellite Technology Company Ltd. (CGSTL). His research interests include remote sensing image classification based on deep learning and 3D urban modeling.

**HONGGUANG JIA** received the B.E. degree from the Harbin Institute of Technology, Harbin, China, in 1994, the master's degree from the Changchun Institute of Optics and Mechanics, Changchun, in 1997, and the Ph.D. degree from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, in 2000.

From May 2002 to May 2003, he was an Associate Professor with the Centre National de la Recherche Scientifique (CNRS), France. He is currently the Vice President and the Chief Engineer with Chang Guang Satellite Technology Company Ltd. (CGSTL), and also a Full Professor with the University of Chinese Academy of Sciences. His research interests include UAV technology, precise terminal guidance technology, the semi-physical simulation of aircraft, and small fast electromechanical servo technology.