Optimization for Oriented Object Detection via Representation Invariance Loss

Qi Ming[1], Zhiqiang Zhou[1,*] Lingjuan Miao[1], Xue Yang[2], Yunpeng Dong[1]
[1] School of Automation, Beijing Institute of Technology
[2] Department of Computer Science and Engineering, Shanghai Jiao Tong University

chaser.ming@gmail.com

Abstract

Arbitrary-oriented objects exist widely in natural scenes, and thus the oriented object detection has received extensive attention in recent years. The mainstream rotation detectors use oriented bounding boxes (OBB) or quadrilateral bounding boxes (QBB) to represent the rotating objects. However, these methods suffer from the representation ambiguity for oriented object definition, which leads to suboptimal regression optimization and the inconsistency between the loss metric and the localization accuracy of the predictions. In this paper, we propose a Representation Invariance Loss (RIL) to optimize the bounding box regression for the rotating objects. Specifically, RIL treats multiple representations of an oriented object as multiple equivalent local minima, and hence transforms bounding box regression into an adaptive matching process with these local minima. Then, the Hungarian matching algorithm is adopted to obtain the optimal regression strategy. We also propose a normalized rotation loss to alleviate the weak correlation between different variables and their unbalanced loss contribution in OBB representation. Extensive experiments on remote sensing datasets and scene text datasets show that our method achieves consistent and substantial improvement. The source code and trained models are available at https://github.com/ming71/RIDet.

1. Introduction

Arbitrary-oriented object detection has a wide range of application scenarios, such as scene text detection [18, 11, 6], face detection [27], object detection in remote sensing images [34, 8, 35, 30, 39, 2, 38], and 3D object detection [45]. In recent years, with the breakthroughs made by convolutional neural networks (CNNs) in the field of object detection [4, 26, 15, 24], a series of CNN-based rotation detectors have been proposed to achieve high-performance oriented object detection [36, 20, 19, 16, 31].
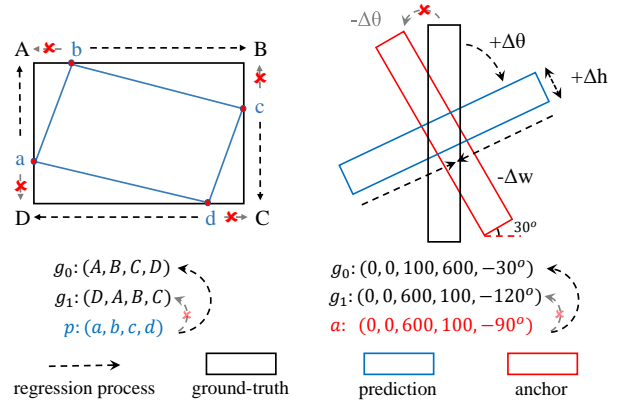
---

*Corresponding author

Figure 1. Examples of suboptimal regression process caused by the representation ambiguity under QBB (left) and OBB (right). g_0 is the given GT representation, and g_1 is one of the ambiguous representations of GT.

Unlike generic object detection that uses the horizontal bounding box to represent the objects, rotation detectors usually adopt the oriented bounding box (OBB) [2, 20, 19] or the quadrilateral bounding box (QBB) [31, 11, 16] to describe the rotating objects, which induces the representation ambiguity. Whether using OBB or QBB, representation ambiguity indicates that an object can be represented in many different forms. These ambiguous representations constitute the representation space of a ground-truth (GT) object g, denoted as $\Omega(g) = \{g_0, g_1, g_2, ...\}$. Ideally, all representations in $\Omega(g)$ are equivalent local optimal solutions in the regression optimization process. However, the loss function of the current rotation detectors can only converge to the given GT representation g_0, and the rest $g_k \in \Omega(g)$ will cause a sharp increase in regression loss. The incorrect loss metric cannot truly reflect the localozation quality of the predictions, which makes the regression process hard to converge and further degrades the detection performance. We next separately analyze the influence of representation ambiguity on the two forms of bounding box definitions.

QBB directly uses four vertices of the quadrilateral to

represent the oriented objects. The unordered vertices have $P_4^4 = 24$ permutations, so theoretically there are 24 quadrilateral representations for one box in the representation space Ω. These diverse representations provide more easily searchable local minima for regression optimization theoretically, but the existing methods do not consider all feasible instances, only to learn a given fixed-sequence quadrilateral representation. In this case, ambiguous representations are not effectively utilized but causes additional learning costs. As shown in the left side of Fig. 1, the predicted quadrilateral is close to the ground-truth box, and the ideal regression strategy is $\{(a \rightarrow D), (b \rightarrow A), (c \rightarrow B), (d \rightarrow C)\}$. But most current detectors will follow the established sequence for regression: $\{(a \rightarrow A), (b \rightarrow B), (c \rightarrow C), (d \rightarrow D)\}$. The suboptimal regression strategy would lead to a sharp loss boost, and thereby increasing the difficulty of network convergence.

Due to boundary problems [39, 36] of angle definition of the oriented rectangle box, we suggest that the representation ambiguity still exists in OBB. On the one hand, the periodicity of angle brings a lot of equivalent representations. On the other hand, the interchangeability between width and height also derives many representations whose angles exceed the boundary. The current detectors cannot handle these ambiguous representations with out-of-bounds angles well. For example, the GT box in the right of Fig. 1 is given as $g_0 : (0, 0, 100, 600, -30°)$. It can also be represented as $g_1 : (0, 0, 600, 100, -120°)$ if without restriction of the angle boundary. For the preset anchor $a : (0, 0, 600, 100, -90°)$, the ideal optimization target is g_1, which only requires a small counterclockwise angle offset. However, due to angle limitation in OBB, it can be only regressed to the suboptimal representation g_0, which would make the deviation of width, height, and angle very large, and the regression loss hard to converge. Besides, the contribution of different variables to the regression accuracy is unequal in OBB, and thus careful weighting is required to balance the loss terms of different variables.

Multiple representations due to representation ambiguity are essentially not bad for the regression task, because they would produce multiple equivalent local minima, and falling into any one of them would reach equivalently good convergence status, which would make the optimization process more flexible, and alleviate the problem of sharp boost and discontinuity in regression loss. Current work, unfortunately, has not fully realized the above benefits and the rotation detectors usually suffer from suboptimal regression strategies that affect the network convergence and degrade detection performance.

In this paper, we propose a Representation Invariance Loss (RIL) that treats the bounding box regression as an adaptive matching process with the local minima of multiple representations, and then searches for the opti-

mal regression strategy dynamically through the Hungarian algorithm during training. Specifically, for QBB, RIL transforms the point regression problem into a dynamic sequential-free vertices matching process, and then Hungarian loss is used to improve convergence and achieve superior performance. For OBB, the RIL allows the predictions with the out-of-bounds angles and adaptively searches for the optimal regression strategy from all ambiguous representations. Besides, a normalized rotation loss is adopted to alleviate the inconsistency between different variables.

Rotation detectors trained with RIL treat the ambiguous representations as the equivalent local minima of the regression loss. These local minima provide better convergence than the unique minima defined by the GT representation. Therefore, during the training process, the regressor can dynamically select the optimal local minima in the representation space as the regression target according to the loss function. Based on the proposed RIL, we further build two Representation Invariant Detectors (RIDet) : RIDet-O and RIDet-Q, to prove the superiority of the proposed method. In summary, the main contribution is as follows:

1) We point out representation ambiguity in arbitrary-oriented object detection and analyzed the problems caused by multiple representations from the perspective of regression optimization.

2) The novel representation invariance loss (RIL) is proposed to transforms the bounding box regression task into an optimal matching process. RIL not only solves the inconsistency between the loss and localization quality caused by ambiguous representations, but also utilizes these local minima to improve network convergence and achieve better performance.

3) RIL can be easily applied to existing methods without any additional overhead. Extensive experiments on multiple rotation detectors and datasets prove the superiority of our approach.

2. Related Work

2.1. Arbitrary-Oriented Object Detection

Arbitrary-oriented objects appear widely in natural scenes, so oriented object detection has gradually become a hot topic in computer vision. In recent years, many advanced frameworks have been proposed to achieve accurate rotating object detection [20, 16, 37, 5, 33, 31].

Unlike generic object detection that uses the horizontal bounding box to represent objects, OBB and QBB are usually adopted for oriented object detection. For example, RoI Transformer [2] learns a rotated region of the horizontal region of interest to improve detection performance. RRPN [18] generates inclined proposals with text orientation information to fit into the text region more accurately. R^3Det [33], S^2A-Net [5], and CFC-Net [19] obtain oriented
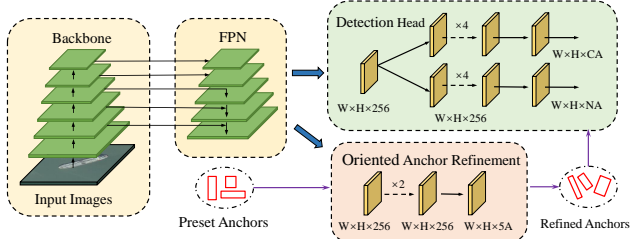
Figure 2. The overall structure of Cas-RetinaNet. C represents the number of classes, and A denotes the number of predefined anchors. N is determined by the parameters of the representation, $N = 5$ for OBB, while $N = 8$ for QBB.

anchors from horizontal anchors to achieve better spatial alignment with oriented objects. TextBoxes++ [11] predicts the vertex offsets from predefined anchor boxes to detect arbitrary-oriented text. Xu *et al.* [31] proposed a novel framework to regress four ratios on each corresponding side to detect multi-orientation objects.

There are also some other arbitrary-oriented object representation methods, such as polar coordinates [46], middle lines [29], ellipse representation [14], and box boundary-aware vectors [41]. Nevertheless, these methods are relatively complex and not general, so the mainstream methods are still OBB and QBB.

2.2. Representation Problems of Oriented Objects

Inappropriate representation of rotating objects will give rise to many problems. Some previous work summarized the issues as the boundary problem [39, 36] in OBB representation, and suggested that this will cause a discontinuous regression loss and hinder the convergence of network training. To solve the issues, Circular Smooth Label [36] and Densely Coded Labels [32] transform the angle regression task into a classification task to avoid the discontinuity of regression loss. SCRDet [39] adopts IoU-Smooth-L$_1$ loss to alleviate the sharp increase of loss caused by angle boundary. RSDet [23] utilizes a modulation loss function to solve the rotation-sensitive error of out-of-bounds angles.

There are also some methods adopt different oriented object representations to avoid the confusion issue of sequential label points under QBB representation [16, 31, 3]. Specifically, SBD [16] predicts key edges of the vertices, and uses an extra combination learning to determine the quadrilateral. Gliding Vertex [31] and TOSO [3] predict the horizontal bounding box of the object and the relative gliding offset to represent the quadrilateral bounding box.

However, these methods simply attemp to eliminate the problems caused by the nonunique definitions of oriented objects, but ignoring that these ambiguous representations are essentially equivalent local optima for regression optimization. Unlike the previous work, our method utilizes diverse representations to search for optimal regression strat-
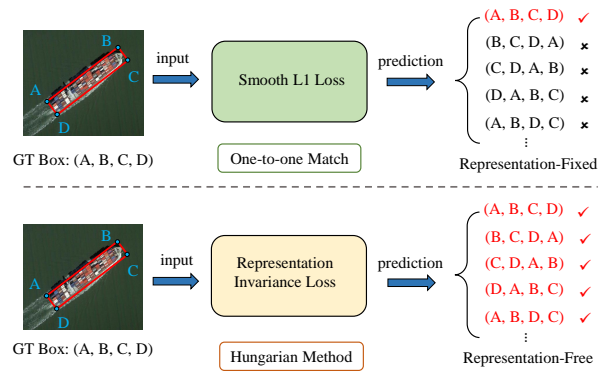

Figure 3. Comparison of regression results of RIL and smooth-L$_1$ loss under QBB. RIL treats ambiguity representations as equivalent local minima, allowing the detector to regress and output sequential-free predictions.

egy. In this case, the representation ambiguity not only no longer causes the various representation problems mentioned above, but also helps to achieve better performance.

3. Representation Invariant Detector

In this section, we first build a baseline and then design Representation Invariance Loss (RIL) for OBB methods and QBB methods, respectively. Then, we apply RIL to the baseline detector to construct Representation Invariant Detector (RIDet).

3.1. Cas-RetinaNet for Orineted Object Detection

We build the Cascaded RetinaNet (Cas-RetinaNet) as the baseline to achieve high-efficiency oriented object detection. Cas-RetinaNet uses horizontal preset anchors, and the Oriented Anchor Refinement Module (O-ARM) is adopted to obtain high-quality training samples. The overall structure of our model is shown in Fig. 2. The Cas-RetinaNet is built based on RetinaNet [13], in which FPN is applied to extract multi-scale features. The detection head performs classification and regression based on the high-quality anchors generated by O-ARM.

The training loss of Cas-RetinaNet consists of three parts: classification loss, anchor refining loss, and bounding box regression loss. The loss function is as follows:

$$
L = \frac{1}{N} \sum_i FL\left(p_i, p_i^*\right) + \frac{1}{N_{p_1}} \sum_i L_{ref}\left(\boldsymbol{t_i}, \boldsymbol{t_i^*}\right) \\
+ \frac{1}{N_{p_2}} \sum_i L_{RI}\left(\boldsymbol{b_i}, \boldsymbol{t_i^*}\right),
\tag{1}
$$

in which FL denotes focal loss [13] for classification task, and L_{ref} is smooth-L$_1$ loss for oriented anchor refinement. p^* and $\boldsymbol{t^*}$ are ground-truth labels for classification, anchor refinement and box regression, while p, $\boldsymbol{t}$, $\boldsymbol{b}$ are corresponding predictions, respectively. N is the number of training
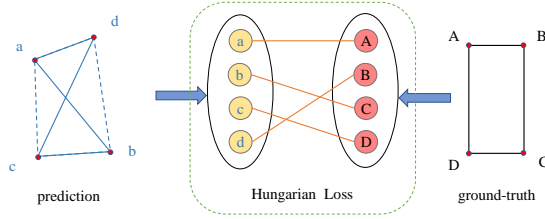
Figure 4. Illustration of matching between GT quadrilateral box and predicted quadrilateral box.
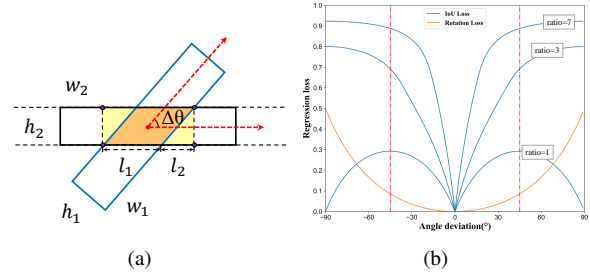

Figure 5. Illustration of normalized angle mapping loss. (a) shows the definition of normalized angle mapping loss. (b) reveals its superiority for measuring angular deviation.

samples, and N_{p_1}, N_{p_2} denote the number of positives in the O-RAM stage and the detection stage, respectively. L_{RI} is the representation invariance loss, which will be elaborated in the next sections.

3.2. Quadrilateral Regression as Point Assignment

Quadrilateral bounding boxes for most arbitrary-oriented objects are convex polygons, such as scene text [40, 9], remote sensing objects [30, 17, 47], human faces [27], and retail products [21]. Therefore, the sequence of vertices is redundant to define a convex quadrilateral. Even so, the current detectors still learn the one-to-one matching between the given GT representation and the predictions as illustrated in the upper part of Fig. 3.

Inspired by the Hungarian method for solving the assignment problem [10], we treat the quadrilateral regression as a dynamic sequential-free points assignment problem. As shown in Fig. 4, we denote the GT quadrilateral as an unordered point set as $g = \{A, B, C, D\}$ and the predicted quadrilateral as $p = \{a, b, c, d\}$. The bounding box regression process can be transformed to the distance optimization between two point sets. We further use Hungarian loss to measure the distance between two sets, and thus the representation invariance loss (RIL) for QBB is as follows:

$$L_{RI}(p, g) = \min_{\pi \in \Pi} \sum_{v \in p, \ v^* \in g_\pi} L_{reg}(v, v^*), \quad (2)$$

in which Π is permutations of $\{1, 2, 3, 4\}$, it represents all feasible sequence of the quadrilateral. π is a permutation in Π, and g_π represents a certain kind of quadrilateral. L_{reg} denotes smooth-L$_1$ loss here.

With the RIL defined above, the network tends to search for the permutation with the minimal loss value among all ambiguous representations for bounding box optimization. For example, the vertices of the prediction in Fig. 4 are very close to the GT quadrilateral, but the unmatched point sequence leads to a large loss for the common rotation detectors. Our RIL can perform sequential-free auto-matching to achieve the optimal regression process and accurately predict the position of the vertices. Note that the training process is constantly changing, so the optimal regression strategy is not fixed, and it may switch between different se-

quences in Π during training. RIL can adaptively select the most suitable representation according to the current optimization situation, and thus can achieve better convergence and performance.

The Hungarian method, which is representation invariant, no longer focuses on the sequence of points, and adaptively searches for the nearby local minimum in all possible permutations to achieve better convergence quality. Moreover, each four-point set determines at most one circumscribed convex quadrilateral. Therefore, in the inference stage, the network allows multiple predictions corresponding to ambiguous representations of the GT objects, as shown in Fig. 3.

3.3. Out-of-bounds Matching for OBB

For the OBB, the boundary of the angle definition constrains the search space of regression loss, and thus the ambiguous representations cannot be effectively utilized to optimize the regression process. Based on the above observations, a Hungarian method based out-of-bounds matching strategy is proposed to solve the boundary constrains. Specifically, we transform the oriented bounding box regression into the adaptive optimal matching between the predictions and the representation space of GT.

Given an object g denoted as (cx, cy, w, h, θ), its representation space $\Omega(g)$ is as follows:

$$\Omega(g) = \{g_0, g_1, g_2, ..., g_i\}, \quad (3)$$

in which the instances are defined as:

$$g_i = \begin{cases} (cx, cy, w, h, \theta + \frac{i\pi}{2}), & i \in \{2k | k \in \mathbb{Z}\} \\ (cx, cy, h, w, \theta + \frac{i\pi}{2}), & i \in \{2k+1 | k \in \mathbb{Z}\} \end{cases} \quad (4)$$

Further the RIL for OBB is as follows:

$$L_{RI}(p, g) = \min_{g_i \in \Omega(g)} L_{nrl}(p, g_i), \quad (5)$$

where p is the prediction, and L_{nrl} denotes the normalized rotation loss for regression which will be elaborated next.

RIL guides the prediction to match the optimal instance in $\Omega(g)$ to search for the nearby local optimum. Noted that
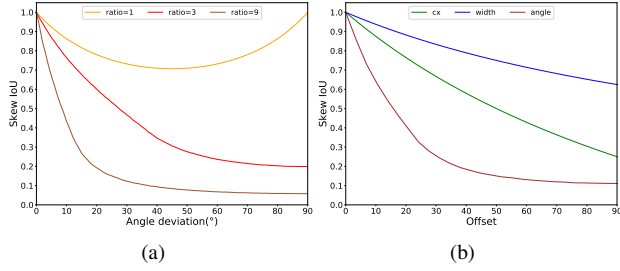
Figure 6. Illustration of inconsistency between loss contribution and IoU variation. (a) shows the same offsets cause the inconsistent IoU changes due to the different aspect ratios of the GT box. (b) reveals that the same contribution of different variables to the loss cause different IoU changes.

the periodicity of the angle leads to an infinite number of possible representations, it is impossible to exhaust all of them for the best matching. Therefore, we further propose a normalized loss to reduce the dimensionality of $\Omega(g)$.

Given the anchor $(x_1, y_1, w_1, h_1, \theta_1)$ and GT box $(x_1, y_1, w_2, h_2, \theta_2)$ with overlapping center points as shown in Fig. 5, the orientations of the two bounding boxes form an acute angle $\Delta\theta$. This angle can be approximated by the projection of its area and the direction of the reference axis. Taking the orientation of the GT as the reference line, l_1 and l_2 are the projections of the overlapping parallelograms, and the intersection area S_0 and projected area S_1 are as follows:

$$l_1 = \frac{h_1}{\sin\theta}, \quad l_2 = \frac{h_2}{\tan\theta},$$
$$S_0 = l_1 \cdot h_2, \quad S_1 = (l_1 + l_2) \cdot h_2. \tag{6}$$

Then the IoU mapping of the angle is as:

$$L_\theta = \frac{S_0}{S_1} - 0.5 = \frac{1}{1 + \alpha \cdot \cos\theta} - 0.5, \tag{7}$$

in which $\alpha = \min(\frac{h_1}{h_2}, \frac{h_2}{h_1})$, and $L_\theta \in [0, 0.5]$. Only when the heights of the anchor and the GT box are equal, and the angle deviation is equal to 0, L_θ reaches the minimum value of 0. With this metric, the representation space of g can be simplified into $\Omega(g) = \{g_0, g_1\}$. In addition, as shown in the right of Fig. 5, L_θ is not sensitive to the aspect ratio of the bounding box itself, and it can converge steadily to objects of different shapes.

OBB representation requires five variables to define the oriented rectangular box. The deviation of IoU caused by the offset of a specic variable varies with the scale and aspect ratio of the GT box, as illustrated in the right of the Fig. 6(a). Besides, the IoU deviations caused by the same offset of different variables are also different, as shown in Fig. 6(b). As a result, there will be inconsistencies between the loss function and the detection metric, which has also been discussed in some previous work [23, 37]. Therefore, it is necessary to normalize the loss of variable offsets, just like L_θ, to obtain a consistent and invariant loss metric.

We further constrain center coordinates and shape of the oriented bounding box to balance the loss contribution between different variables. The center loss L_c is as follows:

$$L_c = \frac{\|c_g - c_a\|_2}{0.5 \cdot \|d\|_2} \tag{8}$$

in which c_g and c_a are the center coordinates of GT and anchor respectively, $d = (w, h)$ denotes the shape of the GT box. During the label assignment phase, only the anchors whose center points fall within the GT box may be selected as positive samples for regression [43]. In this way, L_c is normalized in $[0, 1]$. For example, small objects are more susceptible to inaccurate center point prediction with smooth-L_1 loss. Our normalized center loss is no longer sensitive to the shape of GT box, and thus it is conducive to the detection performance of small objects.

The same scale-sensitive problem also exists in shape regression. As discussed by Redmon *et al.* [24], small deviations in large boxes matter less than in small boxes. Therefore, we utilize the IoU of the horizontal box as the scale-independent shape loss, which is defined as follows:

$$L_s = \min_{g_i \in \Omega(g)} hIoU(p, g_i), \tag{9}$$

in which $hIoU$ calculates the IoU with two overlapping centers without angle consideration, it is defined as follows:

$$hIoU(b_1, b_2) = \frac{\min(w_1, w_2) \cdot \min(h_1, h_2)}{\max(w_1, w_2) \cdot \max(h_1, h_2)}. \tag{10}$$

Obviously, the range of the shape metric is also in $[0, 1]$. The shape loss only evaluates the similarity of the shape and does not specify the width and height, which can speed up the convergence. The prediction of width and height is embedded in the angle loss to achieve an accurate description of the oriented rectangle box jointly.

Then the normalized rotation loss is as follows:

$$L_{nrl} = L_\theta + L_c + L_s, \tag{11}$$

This novel RIL effectively uses entire representation space to optimize the regression process. Simultaneously, the normalized rotation loss can better alleviate the inconsistency of the contribution of different variables to the loss.

4. Experiments

4.1. Datasets

We conduct extensive experiments on multiple datasets to prove the effectiveness of the proposed method, including three remote sensing datasets: HRSC2016 [17], UCAS-AOD [47], DOTA [30], and two scene text datasets: IC-DAR2015 [9], MSRA-TD500 [40].

Backbone	OM	NL	BC	IC15	M500
ResNet-50				75.9	76.1
ResNet-50	✓			76.5 (+0.6)	76.6 (+0.5)
ResNet-50	✓	✓		77.2 (+1.3)	77.5 (+1.4)
ResNet-50	✓	✓	✓	**77.6 (+1.7)**	**78.2 (+2.1)**

Table 1. Evaluation of different components of RIL for OBB. OM, NL, and BC represent optimal matching, normalized loss, and bounded center, respectively. IC15 denotes ICDAR2015 dataset, and M500 is MSRA-TD500 dataset.

Method	Image Res.	BBox	RIL	mAP
RetinaNet [13]	512×512	OBB	×	82.2
			✓	**84.5 (+2.3)**
Cas-RetinaNet [13]	416×416	QBB	×	84.3
			✓	**86.3 (+2.0)**
CFC-Net [19]	416×416	OBB	×	85.2
			✓	**87.3 (+1.9)**
DAL [20]	384×384	OBB	×	87.1
			✓	**88.5 (+1.4)**
S^2A-Net [5]	416×416	OBB	×	88.3
			✓	**89.1 (+0.8)**

Table 2. Performance improvement of RIL on different methods on HRSC2016 dataset.

HRSC2016 [17] is a challenging ship detection dataset, which uses the oriented bounding box to annotate remote sensing ships. It contains 1,061 images with more than 20 categories of various ships. The dataset is divided into training, validation and test set, which contain 436, 181, 444 images respectively. UCAS-AOD [47] is an aerial image dataset for oriented aircraft and car detection. It contains 1,510 images, including 1,000 airplane images and 510 car images. Since no official dataset division strategy is provided, we randomly divide it into the training set, the validation set and the test set with a ratio of 5:2:3. DOTA [30] is the largest public dataset with oriented bounding box annotations for object detection in remote sensing imagery. It contains totally 2,806 large size images (e.g. 4000×4000 pixels). There are 15 categories in total. Half of the images are the training set, 1/6 as the validation set, and 1/3 as the test set. Since that images in DOTA are too large, we crop images into 800×800 patches with the stride set to 200.

ICDAR2015 [9] is the scene text dataset for the incidental scene text challenge in ICDAR 2015 Robust Reading Competition. It contains 1,500 images, of which 1,000 for training and 500 for testing. MSRA-TD500 [40] dataset consists of 300 training images and 200 testing images. It contains text in both English and Chinese. The text regions are annotated with oriented bounding boxes.

4.2. Implementation Deteils

We use the Cas-RetinaNet as the baseline model, and RIL is applied to the baseline to build a Representation Invariant Detector (RIDet). It is further divided into RIDet-O for OBB and RIDet-Q for QBB. We use Pytorch [22] to implement the above methods.

All images are resized to 416×416 or 800×800 for training and testing. Horizontal anchors are preset with aspect ratios of $\{0.5, 1, 2\}$, and scales of $\{2^0, 2^{1/3}, 2^{2/3}\}$. The IoU thresholds used to select training samples are 0.4 in O-RAM and 0.5 in the detection stage. We use Adam optimizer for training with the initial learning rate set to 1e-4. We train the models in 100 epochs for HRSC2016, 24 epochs for DOTA, and 40 epochs for UCAS-AOD, ICDAR2015 and MSRA-TD500. All models are trained on RTX 2080 Ti GPUs with batch size set to 8. Random flip, rotation, and HSV colour

space transformation are adopted for data augmentation.

4.3. Ablation Study

4.3.1 Evaluation of normalized rotation loss for OBB

Experiments conducted on the two scene texts datasets confirmed the effectiveness of RIL under OBB representation. The experimental results are shown in Tab. 1. The baseline model is the Cas-RetinaNet detailed above. The regression head outputs the oriented rectangular bounding box in the experiments in this section. Note that we do not use complex data augmentation in this part.

RIDet-O achieves the F-measure of 77.6% and 78.2% on the ICDAR2015 and MSRA-TD500, which are improved by 1.7% and 2.1% compared to the baseline model, respectively. The optimal matching strategy allows the out-of-bounds predictions to search for the optimal local minima for regression, which is conducive to network convergence, and thus better performance can be achieved. It improves the F-measure by 0.6% and 0.5%. The normalized rotation loss further anchieves the increase of 0.7% and 0.9% on two scene text datasets, which balances the contribution of different variables and alleviates the inconsistency between loss and detection metric. Finally, selecting the anchors with the center point inside the GT box for training improves by 0.4% and 0.7%. These high-quality positives not only provide better priori about the position of the center points but also help constrain the range of center loss, which has also been demonstrated in ATSS [43].

4.3.2 Evaluation on different models

We further conduct experiments on different detectors to verify the generalization of RIL. The experimental results on HRSC2016 [17] and DOTA [30] are shown in Tab. 2 and Tab. 5, respectively.

Experimental results in Tab. 2 show that RIL improves the mAP by 2.3% on RetinaNet with OBB, and 2.0% on Cas-RetinaNet with QBB on HRSC2016 . It proves that the strategy of transforming the bounding box regression task

BBox	Models	mAP	BBox	Model	mAP
OBB	IoU-Smooth [39]	82.59	QBB	MR Loss [23]	86.50
	MR Loss [23]	83.60		Gliding Vertex [31]	88.20
	DCL [32]	89.46		BBAVectors [41]	88.60
	RIDet-O	**89.63**		RIDet-Q	**89.10**

Table 3. Comparison with some related methods on HRSC2016.

Methods	Backbone	Size	mAP
Two-stage:			
R^{2}CNN [7]	ResNet101	800×800	73.07
RRPN [18]	ResNet101	800×800	79.08
R^{2}PN [44]	VGG16	—	79.60
RoI Trans. [2]	ResNet101	512×800	86.20
Gliding Vertex [31]	ResNet101	512×800	88.20
OPLD [28]	ResNet50	1024×1333	88.44
DCL [32]	ResNet101	800×800	89.46
Single-stage:			
RetinaNet [13]	ResNet50	416×416	80.81
RRD [12]	VGG16	384×384	84.30
RSDet [23]	ResNet50	800×800	86.50
BBAVector [41]	ResNet101	608×608	88.60
R^{3}Det [33]	ResNet101	800×800	89.26
RIDet-O (Ours)	ResNet18	416×416	89.25
RIDet-O (Ours)	ResNet50	416×416	89.47
RIDet-Q (Ours)	ResNet101	800×800	89.10
RIDet-O (Ours)	ResNet101	800×800	**89.63**

Table 4. Comparisons with advance methods on HRSC2016.

into the optimal matching process between predction and the representation space of GT can be beneficial in different representation methods. The performance in Tab. 5 further confirms this point. RIDet-Q and RIDet-O achieved 2.26% and 1.81% improvement on the corresponding baseline detector, respectively. S^{2}A-Net [5] is an advanced rotation detector that achieves 76.38% mAP on the DOTA . Integrating RIL to S^{2}A-Net improves performance by 1.24% to achieve the mAP of 77.62%, which further proves the generalization and superiority of our method.

DAL [20] and CFC-Net [19] are also currently advanced detectors that achieve accurate oriented object detection. But these detectors will still be confused by ambiguous representations of GT. After training with RIL, the detection performance is further improved by 1.9% and 1.4% on DAL [20] and CFC-Net [19], respectively. It indicates that as long as there is representation ambiguity, our method can obtain consistent performance gains.

Extensive experiments on different models and representations prove that our method can effectively optimize the representation learning method for the rotation detectors and achieve stable performance improvements.

4.3.3 Comparison with related methods

We have conducted comparisons of the performance of different models in related work on the HRSC2016 dataset , and the results are shown in Tab. 3.

RIDet-O achieves the highest mAP of 89.63% among the compared methods. IoU-Smooth [39] and MR Loss [23] solve the loss discontinuity caused by multiple representations and achieve the mAP of 82.59% and 83.60%, respectively. However, the weight of skew IoU in IoU-Smooth Loss [39] is not differentiable, and MR Loss [23] only considers limited redundant representations. DCL [32] converts the angle regression into fine-grained angle classification to eliminate the problems caused by the angle boundary, and thus achieving the mAP of 89.46%. But it brings additional computational overhead and cannot be directly used to other detectors. Compared with these works with OBB, our method can be directly applied to mainstream rotation detectors to achieve the optimal regression and balanced training without any inference overhead.

Among the compared quadrilateral detectors, RIDet-Q achieves the highest mAP of 89.10%, which proves the effectiveness of our RIL. MR Loss [23] achieves the mAP

of 86.5%, which also consider the order of the vertices for QBB. The main difference is that MR loss traverses the four ordered point sequences to find the suitable representation, which make it limited to the ordered quadrilateral representation, and the traversal of the possible permutations is both time-consuming and insufficient. The offset prediction in Gliding Vertex [31] depends on the accurate prediction of the corresponding horizontal box, which further leads to optimization difficulties. In contrast, RIDet-Q is a completely sequence-free quadrilateral detector. It solves the problem of representation ambiguity via dynamically searching for optimal equivalent representations, and thus improves the detection performance.

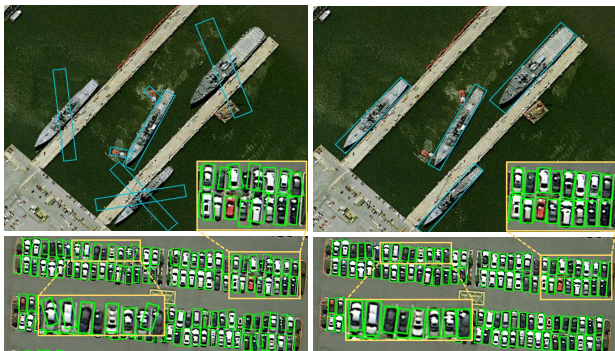## 4.4. Comparison with the State-of-the-Art

4.4.1 Results on HRSC2016

The HRSC2016 dataset [17] contains lots of oriented ships. As shown in Tab. 4, RIDet outperforms other single-stage detectors and even some recent advanced two-stage detectors such as DCL [32], RoI Transformer[2]. RIDet-O and RIDet-O achieve the mAP of 89.10%, and 89.63%, respectively. Especially when the images are resized to 416×416 with lightweight ResNet-18 as the backbone, RIDet-O still achieve competitive performance of 89.25%.

The visualization results in Fig. 7(a) show that the convergence of the angle regression is slow in the baseline model which only learns a specific representation. RIDet treats all feasible representations as equivalent local optimums. Therefore the regression loss converges faster, and better performance can be achieved under the same number of iterations, as shown in Fig. 7(b).

Methods	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
Two-stage:																	
RRPN [18]	R-101	88.52	71.20	31.66	59.30	51.85	56.19	57.25	90.81	72.84	67.38	56.69	52.84	53.08	51.94	53.58	61.01
ICN [1]	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. [2]	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-Net [42]	R-101	87.80	82.40	49.40	73.50	71.10	63.50	76.70	**90.90**	79.20	73.30	48.40	60.90	62.00	67.00	62.20	69.90
O²-DNet [29]	H-104	89.31	82.14	47.33	61.21	71.32	74.03	78.62	90.76	82.23	81.36	60.93	60.17	58.21	66.98	61.03	71.04
SCRDet [39]	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
Single-stage:																	
DRN [21]	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 [20]	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 [23]	R-101	89.80	**82.90**	48.60	65.20	69.50	70.10	70.20	90.50	85.60	83.40	62.50	63.90	65.6 0	67.20	**68.00**	72.20
BBAVector [41]	R-101	88.35	79.96	50.69	62.18	78.43	78.98	87.94	90.85	83.58	84.35	54.13	60.24	65.22	64.28	55.70	72.32
CFC-Net [19]	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
R³Det [33]	R-152	89.49	81.17	50.53	66.10	70.92	78.66	78.21	90.81	85.26	84.23	61.81	63.77	68.16	69.83	67.17	73.74
S²A-Net [5]	R-50	89.07	82.22	53.63	69.88	**80.94**	**82.12**	88.72	90.73	83.77	86.92	63.78	**67.86**	76.51	73.03	56.60	76.38
Baseline-Q	R-101	88.65	75.53	43.68	66.48	62.92	78.59	79.42	90.84	79.97	74.86	62.95	58.77	68.29	72.06	54.83	70.52
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
Baseline-O	R-101	88.36	75.31	46.33	66.59	76.88	77.88	87.23	90.51	82.73	83.84	60.89	60.73	67.29	72.01	56.91	72.89
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
S²A-Net + RIL	R-50	89.31	80.77	**54.07**	**76.38**	79.81	81.99	**89.13**	90.72	83.58	**87.22**	64.42	67.56	**78.08**	**79.17**	62.07	**77.62**

Table 5. Performance evaluation of OBB task on DOTA dataset. H-104 is Hourglass-104. R-50 and R-101 denote ResNet-50 and ResNet-101, respectively.

(a)   (b)

Figure 7. Comparison of detection results between detectors with smoth-L_1 loss (a) and RIL (b) .

4.4.2 Results on DOTA

We compared the proposed method with other state-of-the-art detectors on the DOTA dataset. The ground-truth labels of the test set are not publicly available, and thus the results in Tab. 5 are obtained by submitting the detection results to the official DOTA evaluation server.

As shown in Tab. 5, RIDet-Q and RIDet-O achieve mAP of 72.78% and 74.70%, respectively. Note that the baseline model we used is the naive one-stage detector Cas-RetinaNet. After using RIL, the performance of RIDet can be comparable to many state-of-the-art methods, such as R³Det [33] and CFC-Net [19] that also adopt anchor refinement module. The S²A-Net [5] trained with RIL achieves a mAP of 77.62%, surpassing the other compared methods. More visualization of detection results on the DOTA dataset

Methods	car	airplane	mAP
YOLOv3 [25]	74.63	89.52	82.08
RetinaNet [13]	84.64	**90.51**	87.57
FR-O [30]	86.87	89.86	88.36
RoI Transformer [2]	87.99	89.90	88.95
RIDet-Q	88.50	89.96	89.23
RIDet-O	**88.88**	90.35	**89.62**

Table 6. Detection performance on UCAS-AOD dataset.

can be found in the appendix.

4.4.3 Results on UCAS-AOD

The UCAS-AOD dataset contains a large number of densely arranged cars and airplanes. The experimental results on the UCAS-AOD dataset are shown in Tab. 6. RIDet-Q and RIDet-O achieve the mAP of 89.23% and 89.62%, respectively. RIDet outperforms other advanced rotation detectors such as CFC-Net [19] and RoI Transformer [2] in our experiments, which proves the superiority of our method. The visualization results of vehicle detection of RIDet-Q are shown in Fig. 7. RIL treats the quadrilateral regression as the point matching problem, which signi cantly improves detection performance.

5. Conclusion

In this paper, we discuss the suboptimal regression problem caused by the ambiguous representations in oriented object detection. The Representation Invariance Loss (RIL) is proposed to solve the representation ambiguity and improve network convergence to improve detection perfor-
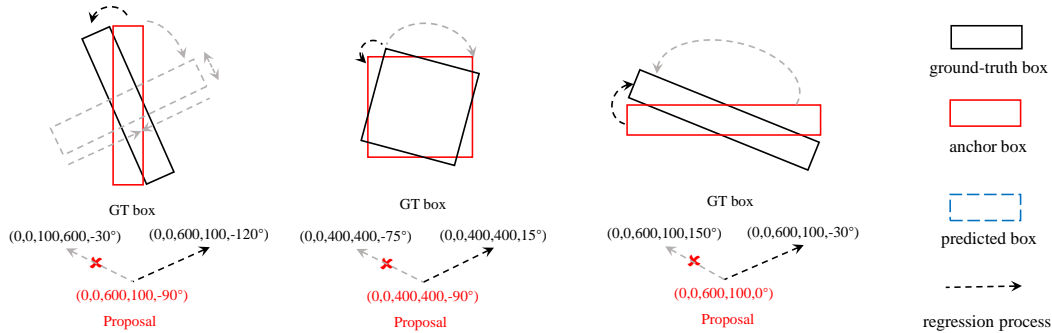
Figure 8. The bounding box regression process of oriented object under QBB (top) and OBB (bottom) characterization. The red box is the GT box, and the blue box is the prediction.

mance. RIL treats multiple representations as equivalent local minima, and then transforms the bounding box regression task into the optimal matching process between predictions and these local minima. Extensive experiments on multiple datasets proved the superiority of our method.

6. Appendix

6.1. The Mapping between Object and Representation

We denote an object as a set $\mathcal{O}$ with one element, and its all possible representations as a set $\mathcal{R}$. Ideally, $f : \mathcal{R} \to \mathcal{O}$ is a bijective function that will not cause ambiguity. However, due to the disorder of the quadrilateral under the QBB and the interchangeability between the width and height under the OBB, this conclusion does not hold. The existence of multiple representations makes $f : \mathcal{R} \to \mathcal{O}$ a surjection. And specific representation can specify the unique object. Therefore, we can use multiple representations as the local minimum for optimization. This conclusion is obviously established in the OBB representation. For QBB representation, this conclusion is only valid when the ground-truth box is a convex quadrilateral. Fortunately, almost all oriented objects that use quadrilateral labels are convex polygons, such as arbitrary-oriented scene text, remote sensing objects, human faces, and retail products. Therefore, our RIL can effectively use disorder vertices to optimize regression.

6.2. Eliminating Representation Ambiguity via Optimal Matching for OBB

The matching-based method can effectively eliminate the problems caused by ambiguous representation, which has been introduced in the main body of the paper. Here is a supplementary explanation for the elimination of ambiguity under the OBB representation.

As shown in the left of Fig. 8, RIL searches for the nearest representation when optimizing the regression loss and
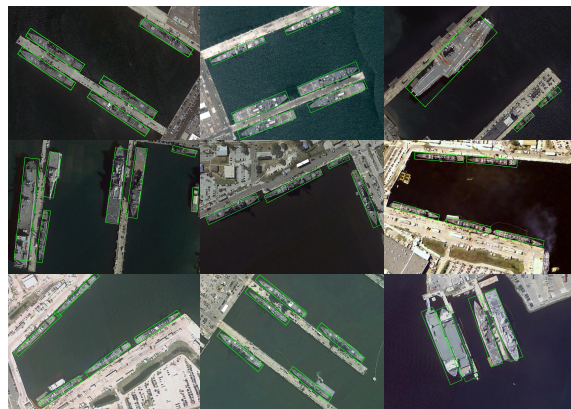

Figure 9. Visualization of detection results on HRSC2016 dataset.

Figure 10. Visualization of detection results on ICDAR2015 dataset.

fits this minimum value with boundary-free predictions. In this way, the loss function will not increase suddenly but will fit the GT box in the fastest way.

The middle of Fig. 8 shows the solution of RIL in the square-like problem [32]. In this case, due to the extreme similarity of width and height, there are two local optima. However, the constraints of definition boundary of the angle hinder free regression of angle, and current detectors cannot

Figure 11. Visualization of detection results on MSRA-TD500 dataset.

Figure 12. Visualization of detection results on UCAS-AOD dataset

be optimized efficiently. RIL can effectively use all redundant representations for optimal matching and resolve this problem.

The illustration on the right in Fig. 8 is similar to the one on the left, so it won't be elaborated here. The root cause of the ambiguity in the 180° representation is the boundedness of the angle definition, while the 90° representation also suffers from the interchangeability between the edges and the angle. It is worth noting that the normalized rotation mapping loss uses the angular deviation between the two boxes to optimize the orientation and constrain it into $[0, 90°]$. Therefore, it can independently solve the representation ambiguity caused by the out-of-bounds definition of angle, such as the situation on the right side of Fig. 8.

6.3. Visualization of Detection Results

We visualized some detection results on different datasets, including three remote sensing data sets: HRSC2016 (shown in Fig. 9), UCAS-AOD (shown in Fig. 12), DOTA (shown in Fig. 13), and two scene text detection datasets: ICDAR2015 (shown in Fig. 10), MSRA-TD500 (shown in Fig. 11).

References

[1] Seyed Majid Azimi, Eleonora Vig, Reza Bahmanyar, Marco Körner, and Peter Reinartz. Towards multi-class object detection in unconstrained remote sensing imagery. In *Asian Conference on Computer Vision*, pages 150–165. Springer, 2018. 8

[2] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning roi transformer for oriented object detection in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2849–2858, 2019. 1, 2, 7, 8

[3] Pengming Feng, Youtian Lin, Jian Guan, Guangjun He, Huifeng Shi, and Jonathon Chambers. Toso: Student's distribution aided one-stage orientation target detection in remote sensing images. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4057–4061. IEEE, 2020. 3

[4] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 1

[5] J. Han, J. Ding, J. Li, and G. S. Xia. Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–11, 2021. 2, 6, 7, 8

[6] Pan He, Weilin Huang, Tong He, Qile Zhu, Yu Qiao, and Xiaolin Li. Single shot text detector with regional attention. In *Proceedings of the IEEE international conference on computer vision*, pages 3047–3055, 2017. 1

[7] Yingying Jiang, Xiangyu Zhu, Xiaobing Wang, Shuli Yang, Wei Li, Hua Wang, Pei Fu, and Zhenbo Luo. R2cnn: rotational region cnn for orientation robust scene text detection. *arXiv preprint arXiv:1706.09579*, 2017. 7

[8] Jiao Jiao, Yue Zhang, Hao Sun, Xue Yang, Xun Gao, Wen Hong, Kun Fu, and Xian Sun. A densely connected end-to-end neural network for multiscale and multiscene sar ship detection. *IEEE Access*, 6:20881–20892, 2018. 1

[9] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1156–1160. IEEE, 2015. 4, 5, 6
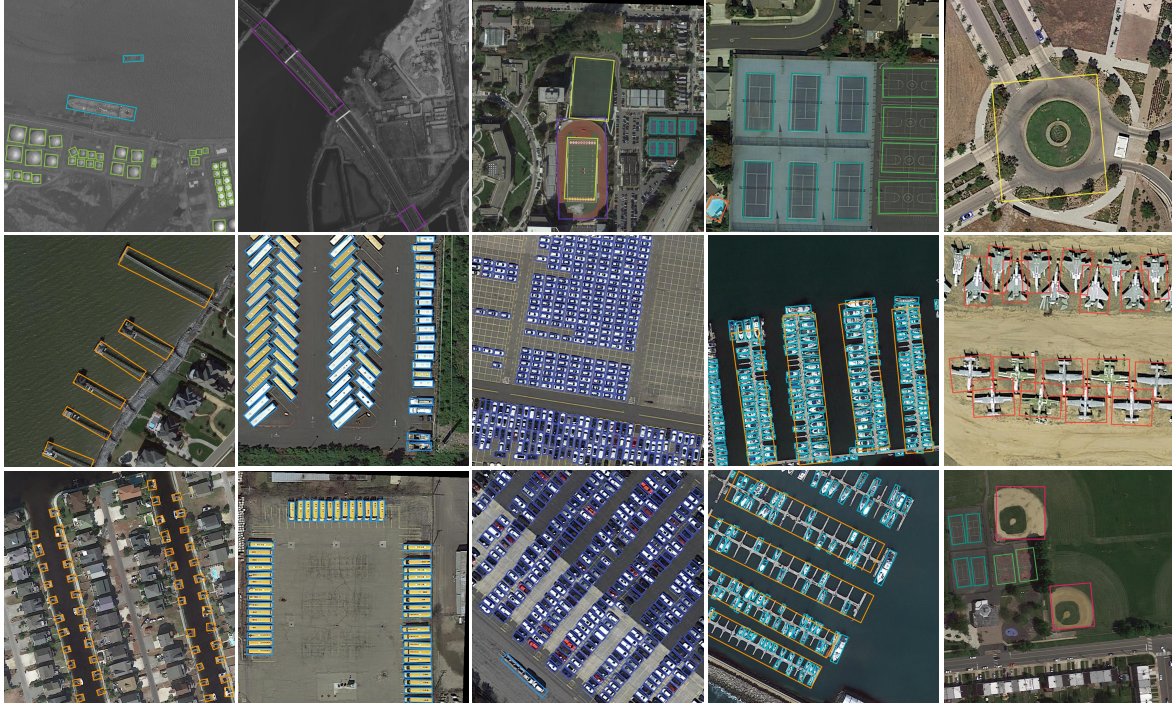
Figure 13. Visualization of detection results on DOTA dataset

[10] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 4

[11] Minghui Liao, Baoguang Shi, and Xiang Bai. Textboxes++: A single-shot oriented scene text detector. *IEEE transactions on image processing*, 27(8):3676–3690, 2018. 1, 3

[12] Minghui Liao, Zhen Zhu, Baoguang Shi, Gui-song Xia, and Xiang Bai. Rotation-sensitive regression for oriented scene text detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5909–5918, 2018. 7

[13] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3, 6, 7, 8

[14] Jin Liu and Haokun Zheng. Efn: Field-based object detection for aerial images. *Remote Sensing*, 12(21):3630, 2020. 3

[15] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 1

[16] Yuliang Liu, Sheng Zhang, Lianwen Jin, Lele Xie, Yaqiang Wu, and Zhepeng Wang. Omnidirectional scene text detection with sequential-free box discretization. *IJCAI*, 2019. 1, 2, 3

[17] Zikun Liu, Liu Yuan, Lubin Weng, and Yiping Yang. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *Proceedings of the International Conference on Pattern Recognition Applications and Methods*, volume 2, pages 324–331, 2017. 4, 5, 6, 7

[18] Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia*, 20(11):3111–3122, 2018. 1, 2, 7, 8

[19] Qi Ming, Lingjuan Miao, Zhiqiang Zhou, and Yunpeng Dong. Cfc-net: A critical feature capturing network for arbitrary-oriented object detection in remote sensing images. *arXiv preprint arXiv:2101.06849*, 2021. 1, 2, 6, 7, 8

[20] Qi Ming, Zhiqiang Zhou, Lingjuan Miao, Hongwei Zhang, and Linhao Li. Dynamic anchor learning for arbitrary-oriented object detection. *arXiv preprint arXiv:2012.04150*, 2020. 1, 2, 6, 7, 8

[21] Xingjia Pan, Yuqiang Ren, Kekai Sheng, Weiming Dong, Haolei Yuan, Xiaowei Guo, Chongyang Ma, and Changsheng Xu. Dynamic refinement network for oriented and densely packed object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11207–11216, 2020. 4, 8

[22] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019. 6

[23] Wen Qian, Xue Yang, Silong Peng, Yue Guo, and Junchi Yan. Learning modulated loss for rotated object detection. *arXiv preprint arXiv:1911.08299*, 2019. 3, 5, 7, 8

[24] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1, 5

[25] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 8

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 1

[27] Xuepeng Shi, Shiguang Shan, Meina Kan, Shuzhe Wu, and Xilin Chen. Real-time rotation-invariant face detection with progressive calibration networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2295–2303, 2018. 1, 4

[28] Qing Song, Fan Yang, Lu Yang, Chun Liu, Mengjie Hu, and Lurui Xia. Learning point-guided localization for detection in remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2020. 7

[29] Haoran Wei, Yue Zhang, Zhonghan Chang, Hao Li, Hongqi Wang, and Xian Sun. Oriented objects as pairs of middle lines. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:268–279, 2020. 3, 8

[30] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2018. 1, 4, 5, 6, 8

[31] Yongchao Xu, Mingtao Fu, Qimeng Wang, Yukang Wang, Kai Chen, Gui-Song Xia, and Xiang Bai. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1, 2, 3, 7

[32] Xue Yang, Liping Hou, Yue Zhou, Wentao Wang, and Junchi Yan. Dense label encoding for boundary discontinuity free rotation detection. *arXiv preprint arXiv:2011.09670*, 2020. 3, 7, 9

[33] Xue Yang, Qingqing Liu, Junchi Yan, Ang Li, Zhiqiang Zhang, and Gang Yu. R3det: Refined single-stage detector with feature refinement for rotating object. *arXiv preprint arXiv:1908.05612*, 2019. 2, 7, 8

[34] Xue Yang, Hao Sun, Kun Fu, Jirui Yang, Xian Sun, Menglong Yan, and Zhi Guo. Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks. *Remote Sensing*, 10(1):132, 2018. 1

[35] Xue Yang, Hao Sun, Xian Sun, Menglong Yan, Zhi Guo, and Kun Fu. Position detection and direction prediction for arbitrary-oriented ships via multitask rotation region convolutional neural network. *IEEE Access*, 6:50839–50849, 2018. 1

[36] Xue Yang and Junchi Yan. Arbitrary-oriented object detection with circular smooth label. In *European Conference on Computer Vision*, pages 677–694. Springer, 2020. 1, 2, 3

[37] Xue Yang, Junchi Yan, Qi Ming, Wentao Wang, Xiaopeng Zhang, and Qi Tian. Rethinking rotated object detection with gaussian wasserstein distance loss. *arXiv preprint arXiv:2101.11952*, 2021. 2, 5

[38] Xue Yang, Junchi Yan, Xiaokang Yang, Jin Tang, Wenglong Liao, and Tao He. Scrdet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing. *arXiv preprint arXiv:2004.13316*, 2020. 1

[39] Xue Yang, Jirui Yang, Junchi Yan, Yue Zhang, Tengfei Zhang, Zhi Guo, Xian Sun, and Kun Fu. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8232–8241, 2019. 1, 2, 3, 7, 8

[40] Cong Yao, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu. Detecting texts of arbitrary orientations in natural images. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1083–1090. IEEE, 2012. 4, 5, 6

[41] Jingru Yi, Pengxiang Wu, Bo Liu, Qiaoying Huang, Hui Qu, and Dimitris Metaxas. Oriented object detection in aerial images with box boundary-aware vectors. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2150–2159, 2020. 3, 7, 8

[42] Gongjie Zhang, Shijian Lu, and Wei Zhang. Cad-net: A context-aware detection network for objects in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 57(12):10015–10024, 2019. 8

[43] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9759–9768, 2020. 5, 6

[44] Zenghui Zhang, Weiwei Guo, Shengnan Zhu, and Wenxian Yu. Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks. *IEEE Geoscience and Remote Sensing Letters*, 15(11):1745–1749, 2018. 7

[45] Yu Zheng, Danyang Zhang, Sinan Xie, Jiwen Lu, and Jie Zhou. Rotation-robust intersection over union for 3d object detection. In *European Conference on Computer Vision*, pages 464–480. Springer, 2020. 1

[46] Lin Zhou, Haoran Wei, Hao Li, Wenzhe Zhao, Yi Zhang, and Yue Zhang. Objects detection for remote sensing images based on polar coordinates. *arXiv preprint arXiv:2001.02988*, 2020. 3

[47] Haigang Zhu, Xiaogang Chen, Weiqun Dai, Kun Fu, Qixiang Ye, and Jianbin Jiao. Orientation robust object detection in aerial images using deep convolutional neural network. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 3735–3739. IEEE, 2015. 4, 5, 6