R^3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object

Xue Yang[1,2], Junchi Yan[1,2,*], Ziming Feng[3], Tao He[4,5]

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[3] China Merchants Bank Credit Card Center
[4] Anhui COWAROBOT CO., Ltd.
[5] Anhui Provincial Key Laboratory of Multimodal Cognitive Computation
{yangxue-2019-sjtu, yanjunchi}@sjtu.edu.cn zimingfzm@cmbchina.com tommie.he@cowarobot.com

Abstract

Rotation detection is a challenging task due to the difficulties of locating the multi-angle objects and separating them effectively from the background. Though considerable progress has been made, for practical settings, there still exist challenges for rotating objects with large aspect ratio, dense distribution and category extremely imbalance. In this paper, we propose an end-to-end refined single-stage rotation detector for fast and accurate object detection by using a progressive regression approach from coarse to fine granularity. Considering the shortcoming of feature misalignment in existing refined single-stage detector, we design a feature refinement module to improve detection performance by getting more accurate features. The key idea of feature refinement module is to re-encode the position information of the current refined bounding box to the corresponding feature points through pixel-wise feature interpolation to realize feature reconstruction and alignment. For more accurate rotation estimation, an approximate SkewIoU loss is proposed to solve the problem that the calculation of SkewIoU is not derivable. Experiments on three popular remote sensing public datasets DOTA, HRSC2016, UCAS-AOD as well as one scene text dataset ICDAR2015 show the effectiveness of our approach. The source code is available at https://github.com/Thinklab-SJTU/R3Det_Tensorflow and is also integrated in our open source rotation detection benchmark: https://github.com/yangxue0827/RotationDetection.

Introduction

Object detection is one of the fundamental tasks in computer vision, and many high-performance general-purpose object detectors have been proposed. Current popular detection methods can be in general divided into two types: two-stage object detectors (Girshick et al. 2014; Girshick 2015; Ren et al. 2015; Dai et al. 2016; Lin et al. 2017a) and single-stage object detectors (Liu et al. 2016; Redmon et al. 2016; Lin et al. 2017b). Two-stage methods have achieved promising results on various benchmarks, while the single-stage approach maintains faster detection speed.

However, current general horizontal detectors have fundamental limitations for many practical applications. For instance, scene text detection, retail scene detection and remote sensing object detection whereby the objects can appear in various orientations. Therefore, many rotation detectors based on a general detection framework have been proposed in the above fields. In particular, three challenges are pronounced, as analyzed as follows:

1) **Large aspect ratio.** The Skew Intersection over Union (SkewIoU) score between large aspect ratio objects is sensitive to change in angle, as sketched in Figure 3b.

2) **Densely arranged.** As illustrated in Figure 6, many objects usually appear in densely arranged forms.

3) **Arbitrary orientations.** Objects in images can appear in various orientations, which requires the detector to have accurate direction estimation capabilities.

This paper is devoted to design an accurate and fast rotation detector. To maintain high detection accuracy and speed for large aspect ratio objects, we have adopted a refined single-stage rotation detector. **First**, we find that rotating anchors can perform better in dense scenes, while horizontal anchors can achieve higher recalls in fewer quantities. Therefore, a progressive regression form from coarse to fine is adopted in the refined single-stage detector, that is, the horizontal anchors are used in the first stage for faster speed and higher recall, and then the refined rotating anchors are used in the subsequent refinement stages to adapt to intensive scenarios. **Second**, we also notice that existing refined single-stage detectors (Zhang et al. 2018a; Chi et al. 2019) have feature misalignment problems[1], which greatly limits the reliability of classification and regression during the refined stages. We design a feature refinement module FRM that uses the feature interpolation to obtain the position information correspond to the refined anchors and reconstruct the whole feature map by pixel-wise manner to achieve feature alignment. FRM can also reduce the number of refined bounding box after the first stage, thus speeding up the model. Experimental results have shown that feature refinement is sensitive to location and its improvement in detection results is very noticeable. **Finally**, an approximate SkewIoU loss is devised to address the indifferentiable problem of SkewIoU calculation for more accurate rotation estimation. Combing these three techniques as a whole, our approach achieves state-of-the-art performance with consid-

*Corresponding author is Junchi Yan.

[1] Mainly refers to misalignment between region of interest (RoI) and the feature, see details in Figure 4c.

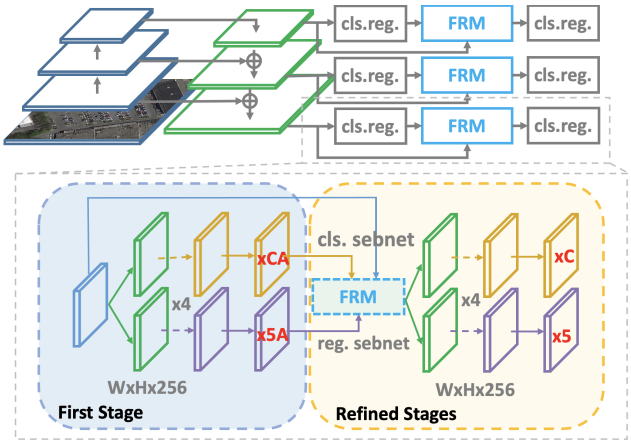


Figure 1: The architecture of the proposed Refined Rotation Single-Stage Detector (RetinaNet as an embodiment). The refinement stage can be repeated by multiple times. 'A' indicates the number of anchors on each feature point, and 'C' indicates the number of categories.

erable speed on four public rotating sensitive datasets including DOTA, HRSC2016, UCAS-AOD, and ICDAR2015. Specifically, this work makes the following contributions:

1) For large aspect ratio object detection, an accurate and fast rotation singe-stage detector is devised in a refined manner, for high-precision detection. In contrast to the recent learning based methods (Chen et al. 2018; Jang et al. 2019; Zhang et al. 2019) for feature alignment, which lacks an explicit mechanism to compensate the misalignment, we propose a direct and effective pure computing based approach which is further extended to handle the rotation case. To our best knowledge, it is the first work for solving the feature misalignment problem for rotation detection.

2) For densely arranged objects, we develop an efficient coarse-to-fine progressive regression approach to better exploring the two forms of anchors in a more flexible manner, tailored to each detection stage. Compared with the previous methods (Ma et al. 2018; Yang et al. 2018b; Fu et al. 2018; Yang et al. 2018a, 2019a) using one single anchor form, our method is more flexible and efficient.

3) For arbitrarily-rotated objects, a derivable approximate SkewIoU loss is devised for more accurate rotation estimation. Compared with the over-approximation of SkewIoU loss in recent work (Chen et al. 2020), our method retains the accurate SkewIoU amplitude and only approximates the gradient direction of SkewIoU loss.

Related Work

Two-Stage Object Detectors. Most existing two-stage methods are region based. In a region based framework, category-independent region proposals are generated from an image in the first stage, followed with feature extraction from these regions, and then category-specific classifiers and regressors are used for classification and regression in the second stage. Finally, the detection results are obtained by

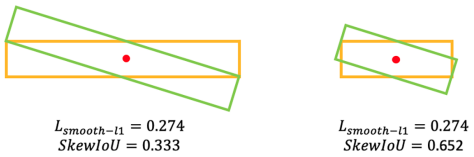


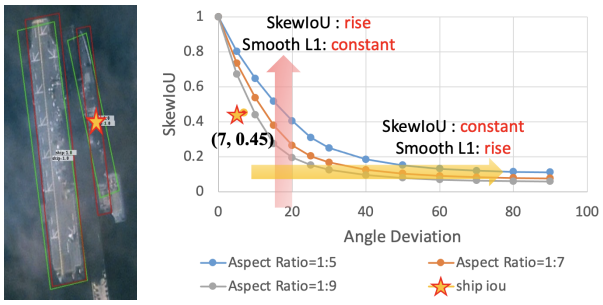
Figure 2: Comparison between SkewIoU and Smooth L1.

using post-processing methods such as non-maximum suppression (NMS). Faster-RCNN (Ren et al. 2015), R-FCN (Dai et al. 2016), and FPN (Lin et al. 2017a) are classic structures in a two-stage approach.

Single-Stage Object Detectors. For their efficiency, single-stage detection methods are receiving more and more attention. Redmon et al. (Redmon et al. 2016) propose YOLO, a unified detector casting object detection as a regression problem from image pixels to spatially separated bounding boxes and associated class probabilities. To preserve real-time speed without sacrificing too much detection accuracy, Liu et al. (Liu et al. 2016) propose SSD. The work (Lin et al. 2017b) solves the class imbalance problem by proposing RetinaNet with Focal loss and further improves the accuracy of single-stage detector.

Rotation Object Detectiors. Remote sensing, scene text and retail scene are the main application scenarios of the rotation detector. Due to the complexity of the remote sensing image scene and the large number of small, cluttered and rotated objects, two-stage rotation detectors are still dominant for their robustness. Among them, ICN (Azimi et al. 2018), ROI-Transformer (Ding et al. 2019), SCRDet (Yang et al. 2019b) and Gliding Vertex (Xu et al. 2020) are state-of-the-art detectors. However, they use a more complicated structure causing speed bottleneck. For scene text detection, there are many efficient rotation detection methods, including both two-stage methods (R^2CNN (Jiang et al. 2017), RRPN (Ma et al. 2018), FOTS (Liu et al. 2018)), as well as single-stage methods (EAST (Zhou et al. 2017), TextBoxes++ (Liao, Shi, and Bai 2018)). For retail scene detection, DRN (Pan et al. 2020) and PIoU (Chen et al. 2020) Loss are the latest two rotation detectors used in retail scene detection, and two rotation retail datasets are proposed.

Refined Object Detectors. To achieve better detection accuracy, many cascaded or refined detectors are proposed. The Cascade RCNN (Cai and Vasconcelos 2018), HTC (Chen et al. 2019), and FSCascade (Li, Yang, and Zhang 2019) perform multiple classifications and regressions in the second stage, which greatly improved the detection accuracy. The same idea is also used in single-stage detectors, such as RefineDet (Zhang et al. 2018a). Unlike the two-stage detectors, which use RoI Pooling (Girshick 2015) or RoI Align (He et al. 2017) for feature alignment. The currently refined single-stage detector is not well resolved in this respect. An important requirement of the refined single-stage detector is to maintain a full convolutional structure, which can retain the advantage of speed, but methods such as RoI Align cannot satisfy it whereby fully-connected layers have to be introduced. Although some works (Chen et al.



(a) (b)

Figure 3: The SkewIoU scores vary with the angle deviation. The red and green rectangles represent the ground truth and the prediction bounding box, respectively.

2018; Jang et al. 2019; Zhang et al. 2019) use deformable convolution (Dai et al. 2017) for feature alignment, whose offset parameters are often obtained by learning the offset between the pre-defined anchor box and the refined anchor. These deformable-based feature alignment methods are too implicit and can not ensure that features are truely aligned. Feature misalignment still limits the performance of the refined single-stage detector. Compared to these methods, our method can clearly find the corresponding feature area by calculation and achieve the purpose of feature alignment by feature map reconstruction.

The Proposed Method

We give an overview of our method as sketched in Figure 1. The embodiment is a refined single-stage rotation detector based on the RetinaNet (Lin et al. 2017b), namely Refined Rotation RetinaNet (R^{3}Det). The refinement stage (which can be added and repeated by multiple times) is added to the network to refine the bounding box, and the feature refinement module FRM is added during the refinement stage to reconstruct the feature map.

Rotation RetinaNet

Base Setting. For RetinaNet-based rotation detection, we use five parameters (x, y, w, h, θ) to represent arbitrary-oriented bounding box. Ranging in $[-\pi/2, 0)$, θ denotes the acute or right angle between w of bounding box and x-axis. Therefore, it calls for predicting an additional angle offset in the regression subnet. The regression equation is as follows:

$$\begin{aligned} &t_x = (x - x_a)/w_a, t_y = (y - y_a)/h_a \\ &t_w = \log(w/w_a), t_h = \log(h/h_a), t_\theta = \theta - \theta_a \\ &t_x^{'} = (x^{'} - x_a)/w_a, t_y^{'} = (y^{'} - y_a)/h_a \\ &t_w^{'} = \log(w^{'}/w_a), t_h^{'} = \log(h^{'}/h_a), t_\theta^{'} = \theta^{'} - \theta_a \end{aligned} \tag{1}$$

where x, y, w, h, θ denote the box's center coordinates, width, height and angle, respectively. Variables $x, x_a, x^{'}$ are for the ground-truth box, anchor box, and predicted box, respectively (likewise for y, w, h, θ).

Loss Function. As shown in Figure 2, each box set has the same center point, height and width. The angle difference between the two box sets is the same, but the aspect ratio is different. As a result, the smooth L1 loss value of the two sets is the same (mainly from the angle difference), but the SkewIoU is quite different. The red and orange arrows in Figure 3b show the inconsistency between SkewIoU and smooth L1 Loss. We can draw conclusion that smooth L1 loss function is still not suitable for rotation detection, especially for objects with large aspect ratios, which are sensitive to SkewIoU. What's more, the evaluation metric of rotation detection is also dominated by SkewIoU.

The IoU related loss is an effective regression loss function that can solve above problem and is already widely used in horizontal detection, such as GIoU (Rezatofighi et al. 2019), DIoU (Zheng et al. 2020), etc. However, the SkewIoU calculation function between two rotating boxes is underivable, which means that we cannot directly use the SkewIoU as the regression loss function. Inspired by SCRDet (Yang et al. 2019b), we propose a derivable approximate SkewIoU loss, the multi-task loss is defined as follows:

$$L = \frac{\lambda_1}{N_{pos}} \sum_{n=1}^{N_{pos}} \frac{L_{reg}(v_n^{'}, v_n)}{|L_{reg}(v_n^{'}, v_n)|} |f(IoU)| + \frac{\lambda_2}{N} \sum_{n=1}^{N} L_{cls}(p_n, t_n) \tag{2}$$

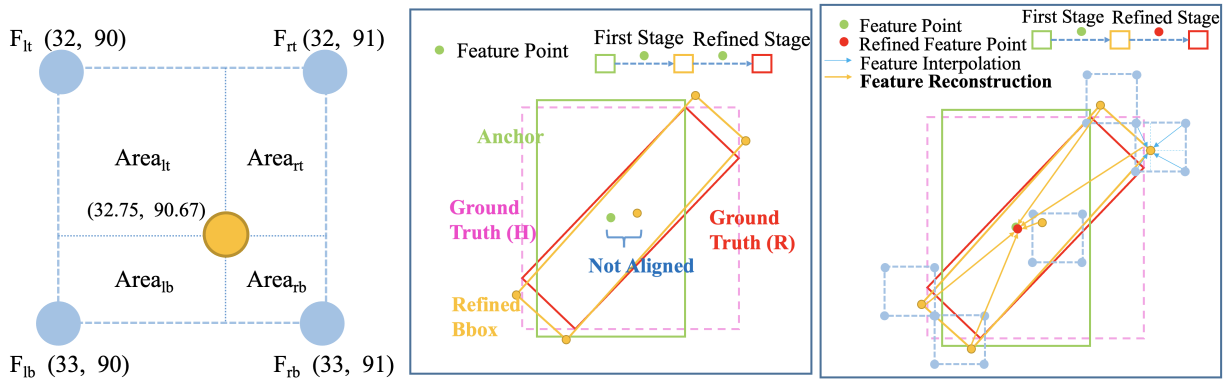
$$L_{reg}(v^{'}, v) = L_{l1}(v_\theta^{'}, v_\theta) - IoU(v_{\{x,y,w,h\}}^{'}, v_{\{x,y,w,h\}}) \tag{3}$$

where N and N_{pos} indicates the number of all and positive anchors. $v^{'}$ represents the predicted offset vectors, v denotes the targets vector of ground-truth. While t_n indicates the label of object, p_n is the probability distribution of various classes calculated by sigmoid function. SkewIoU denotes the overlap of the prediction box and ground-truth. The hyper-parameter λ_1, λ_2 control the trade-off and are set to 1 by default. The classification loss L_{cls} is implemented by focal loss (Lin et al. 2017b). $|.|$ is used to obtain the modulus of the vector and is not involved in gradient back propagation. $f(.)$ represents the loss function related to SkewIoU. $IoU(.)$ represents the horizontal bounding box IoU calculation function.

Compared to the traditional regression loss, the new regression loss can be divided into two parts, $\frac{L_{reg}(v_n^{'}, v_n)}{|L_{reg}(v_n^{'}, v_n)|}$ determines the direction of gradient propagation (a unit vector), which is an important part to ensure that the loss function is derivable. $|f(SkewIoU)|$ is responsible for adjusting the loss value (magnitude of gradient), and it is unnecessary to be derivable (a scalar). Taking into account the inconsistency between SkewIoU and smooth L1 loss, we use Equation 3 as the dominant gradient function for regression loss. Through such a combination, the loss function is derivable, while its size is highly consistent with SkewIoU. Experiments show that the detector based on this approximate SkewIoU loss can achieve considerable gains.

Refined Rotation RetinaNet

Refined Detection. The SkewIoU score is sensitive to the change in angle, and a slight angle shift causes a rapid decrease in the IoU score, as shown in Figure 3. Therefore, the refinement of the prediction box helps to improve the recall rate of the rotation detection. We join multiple refinement



(a) Feature interpolation

(b) Refine box with misaligned feature due to bounding box location changes.

(c) Refine box with aligned features by reconstructing the feature map.

Figure 4: Root cause analysis of feature misalignment and the core idea for our proposed feature refinement module.

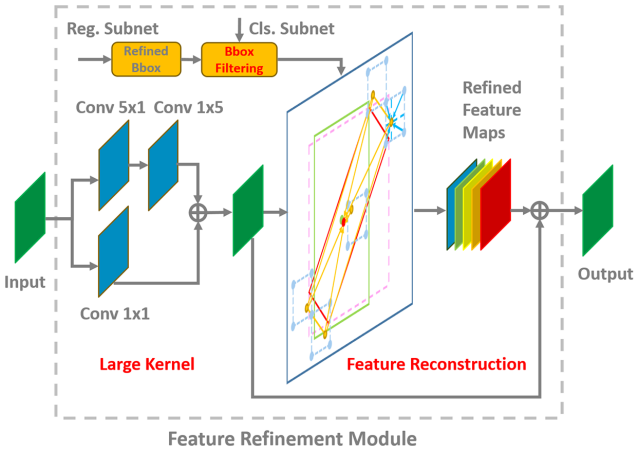


Figure 5: Feature Refinement Module FRM. It mainly includes three parts: refined bounding box filtering (BF), large kernel (LK) and feature reconstruction (FR).

stages with different IoU thresholds. In addition to using the foreground IoU threshold 0.5 and background IoU threshold 0.4 in the first stage, the thresholds of first refinement stage are set 0.6 and 0.5, respectively. If there are multiple refinement stages, the remaining thresholds are 0.7 and 0.6. The overall loss for refined detector is defined as follows:

$$L_{total} = \sum_{i=1}^{N} \alpha_i L_i \tag{4}$$

where L_i is the loss value of the i-th refinement stage and trade-off coefficients α_i are set to 1 by default.

Feature Refinement Module. Many refined detectors still use the same feature map to perform multiple classifications and regressions, without considering the feature misalignment caused by the location changes of the bounding box. Figure 4b depicts the box refining process without feature refinement, resulting in inaccurate features, which can be disadvantageous for those categories that have a large aspect ratio or a small sample size. Here we propose to re-encode the position information of the current refined bounding box (orange rectangle) to the corresponding feature points (red point[2]), thereby reconstructing the entire feature map by pixel-wise manner to achieve the alignment of the features. The whole process is shown in Figure 4c. To accurately obtain the location feature information correspond to the refined bounding box, we adopt the bilinear feature interpolation method, as shown in Figure 4a. Feature interpolation can be formulated as follows:

$$F = F_{lt} * A_{rb} + F_{rt} * A_{lb} + F_{rb} * A_{lt} + F_{lb} * A_{rt} \tag{5}$$

where A denotes the $Area$ in Figure 4a, $F \in \mathbb{R}^{C\times1\times1}$ represents the feature vector of the point on the feature map.

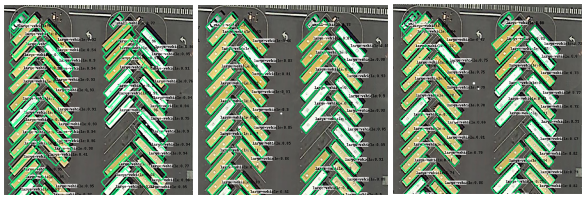
Based on the above result, a feature refinement module is devised, whose structure and pseudo code is shown in Fig-

Algorithm 1 Feature Refinement Module

Input: original feature map F, the bounding box (B) and confidence (S) of the previous stage
Output: reconstructed feature map F'

$B' \leftarrow BoxFilter(B, S)$;
$h, w \leftarrow Shape(F)$, $F' \leftarrow ZerosLike(F)$;
$F \leftarrow Conv_{1\times1}(F) + Conv_{1\times5}(Conv_{5\times1}(F))$
for $i \leftarrow 0$ **to** $h-1$ **do**
  for $j \leftarrow 0$ **to** $w-1$ **do**
    $P \leftarrow GetFivePoints(B'(i,j))$;
    for $p \in P$ **do**
      $p_x \leftarrow Min(p_x, w-1), p_x \leftarrow Max(p_x, 0)$;
      $p_y \leftarrow Min(p_y, h-1), p_y \leftarrow Max(p_y, 0)$;
      $F'(i,j) \leftarrow F'(i,j) + BilinearInte(F, p)$;
    end for
  end for
end for
$F' \leftarrow F' + F$;
return F'

[2]The red and green points should be totally overlapping to each other, while here the red point is intentionally offset in order to distinguishingly visualize the entire process.



(a) RetinaNet-H (b) RetinaNet-R (c) R^3Det^*

Figure 6: Visualization on DOTA. Here 'H' and 'R' represent the horizontal and rotating anchors, respectively.

Method	FRM		SL	SV	LV	SH	mAP
	BF&FR	LK					
RetinaNet-R				64.6	71.0	68.6	62.8
RetinaNet-H				63.5	50.7	65.9	62.8
R^3Det^*		✓		65.0	67.3	67.3	63.5
R^3Det	✓	✓		65.8	72.8	70.1	66.3
$R^3Det^\dagger$	✓	✓		67.5	74.0	70.3	67.7
$R^3Det^\dagger$	✓	✓	✓	68.0	72.7	76.0	69.5

Table 1: Ablative study of each component in our method on the DOTA dataset. $R^3Det^\dagger$ indicates that R^3Det with two refinement stages. BF, LK, FR and SL denote box filtering, large kernel, feature reconstruction and SkewIoU loss.

ure 5 and Algorithm 1, respectively. Specifically, the feature map is added by two-way convolution to obtain a new feature (large kernel, LK). Only the bounding box with the highest score of each feature point is preserved in the refinement stage to increase the speed (box filtering, BF), meanwhile ensuring that each feature point corresponds to only one refined bounding box. The filtering of bounding boxes is a necessary step for feature reconstruction (FR). For each feature point of the feature map, we obtain the corresponding feature vector on the feature map according to the five coordinates of the refined bounding box (one center point and four corner points). A more accurate feature vector is obtained by bilinear interpolation. We add the five feature vectors and replace the current feature vector. After traversing the feature points, we reconstruct the whole feature map. Finally, the reconstructed feature map is added to the original feature map to complete the whole process.

The refinement stage can be added and repeated by multiple times. The feature reconstruction process of each refinement stage is simulated as follows:

$$F_{i+1} = FRM(B_i, S_i, \{P_2, ..., P_7\}) \tag{6}$$

where F_{i+1} represents the feature map of the $i+1$ stage, B_i, S_i represent the bounding box and confidence score of the ith stage prediction, respectively.

Discussion for comparison with RoIAlign. The core to solve feature misalignment for FRM is feature reconstruction. Compared with RoI Align that has been adopted in many two-stage rotation detectors including R^2CNN and RRPN, FRM has the following differences that contribute to R^3Det's higher efficiency, as shown in Table 6.

1) RoI Align has more sampling points (the default number is $7 \times 7 \times 4 = 196$), and reducing the sampling point greatly affects the performance of the detector. FRM only samples five feature points, about one-fortieth of RoI Align, which gives FRM a huge speed advantage.

2) RoI Align need to obtain the feature corresponding to RoI **(instance level)** before classification and regression. In contrast, FRM first obtains the features corresponding to the feature points, and then reconstructs the entire feature map **(image level)**. As a result, the FRM based method can maintain a full convolution structure that leads to higher efficiency and fewer parameters, compared with the RoI Align based method that involves a fully-connected structure.

Method	FRM		IC15	HR16	UCAS-AOD
	BF&FR	LK			
R^3Det^*		✓	83.27	94.98	95.03
R^3Det	✓	✓	84.96	96.01	96.17

Table 2: Comparison between R^3Det^* and R^3Det. IC15 and HR16 represent ICDAR2015 and HRSC2016.

Experiments

Datasets and Protocls

DOTA (Xia et al. 2018) contains 15 common categories, 2,806 images and 188,282 instances. The proportions of the training set, validation set, and testing set are 1/2, 1/6, and 1/3, respectively. We divide the images into 600^2 subimages with an overlap of 150 pixels and scale it to 800^2.

UCAS-AOD (Zhu et al. 2015) contains 1,510 aerial images of approximately $659 \times 1,280$ pixels, with two categories of 14,596 instances in total. In line with (Azimi et al. 2018; Xia et al. 2018), we randomly select 1,110 for training and 400 for testing. HRSC2016 (Liu et al. 2017) contains images from two scenarios including ships on sea and ships close inshore. The training, validation and test set include 436, 181 and 444 images.

ICDAR2015 (Karatzas et al. 2015) contains a total of 1,500 pictures, 1000 of which are used for training and the remaining are for testing.

For all datasets, the models are trained by 20 epochs in total, and learning rate is reduced tenfold at 12 epochs and 16 epochs, respectively. The initial learning rates for RetinaNet is 5e-4. The number of image iterations per epoch for DOTA, ICDAR2015, HRSC2016 and UCAS-AOD are 54k, 10k, 5k and 5k, and doubled if data augmentation and multi-scale training are used. The experiments in this paper are initialized by ResNet50 (He et al. 2016) by default unless otherwise specified. Weight decay and momentum are 0.0001 and 0.9, respectively. We employ MomentumOptimizer over 4 GPUs with a total of 4 images per minibatch (1 images per GPU). The anchors have areas of 32^2 to 512^2 on pyramid levels P3 to P7, respectively. At each pyramid level we use anchors at seven aspect ratios $\{1, 1/2, 2, 1/3, 3, 5, 1/5\}$ and three scales $\{2^0, 2^{1/3}, 2^{2/3}\}$. We also add six angles $\{-90°, -75°, -60°, -45°, -30°, -15°\}$ for rotating anchor-based method (RetinaNet-R).

#Stages	Test	BR	SV	LV	SH	HA	mAP
1	1	39.3	63.5	50.7	65.9	51.9	62.8
2	2	42.7	65.8	72.8	70.1	56.1	66.3
3	3	45.1	67.1	73.7	70.2	57.0	67.2
4	4	44.2	65.3	73.0	70.2	55.7	67.0
3	$\overline{2-3}$	45.1	67.5	74.0	70.3	57.3	67.7

Table 3: Ablation study for number of stages on DOTA. Note $\overline{2-3}$ indicates the ensemble result, which is the collection of all outputs from the refinement stages.

Method	Baseline	f_1	f_2	f_3
RetinaNet-H	62.8	NAN	65.1	65.3
R^3Det†	67.7	NAN	69.0	69.5

Table 4: Experiments with different SkewIoU functions. f_1, f_2 and f_3 represent $-\ln(SkewIoU)$, $1 - SkewIoU$ and $\exp(1 - SkewIoU) - 1$, respectively.

Robust Baseline Methods

ReitnaNet-H: The advantage of a horizontal anchor is that it can use less anchor but match more positive samples by calculating the IoU with the horizontal circumscribing rectangle of the ground truth, but it introduces a large number of non-object or regions of other objects. For an object with a large aspect ratio, its prediction rotating bounding box tends to be inaccurate, as shown in Figure 6a.

ReitnaNet-R: In contrast, in Figure 6b, the rotating anchor avoids the introduction of noise regions by adding angle parameters and has better detection performance in dense scenes. However, the number of anchor has multiplied, about 6 times in this paper, thus making the model less efficient.

R^{3}Det*: This is a refined detector without feature refinement. Considering the number of original anchors determines the speed of the model, we adopt a progressive regression form from coarse to fine. Specifically, we first use horizontal anchor to reduce the number of anchors and increase the object recall rate in the first stage, and then use the rotating refined anchor to overcome the problems caused by dense scenes in subsequent stages, as shown in Figure 6c.

RetinaNet-H and RetinaNet-R have similar overall mAP according to Table 1, while with their respective characteristics. The horizontal anchor-based approach clearly has an advantage in speed, while the rotating anchor-based method has better regression capabilities in dense object scenarios and objects with large aspect ratio, such as small vehicle, large vehicle, and ship. R^{3}Det* achieves 63.5% performance, better than RetinaNet-H and RetinaNet-R. Although the category of dense and large aspect ratio has been improved a lot, it is still not as good as RetinaNet-R (such as LV and SH). RetianNet-R's advantages in this regard will also be reflected in Table 6.

Ablation Study

Feature Refinement Module. Table 1 shows that R^{3}Det* can improve performance by about 0.8% which is not significant. We believe that the main reason is that the feature misalignment problem. FRM reconstructs the feature map based on the refined anchor, which increases the overall performance by 2.8% to 66.3% according to Table 1. In order to further verify the effectiveness of FRM, we have also verified it in other datasets, including the text dataset ICDAR2015, and remote sensing dataset HRSC2016 and UCAS-AOD. FRM still shows a stronger performance advantage. As shown in Table 2, the FRM-based method is improved by 1.69%, 1.03%, and 1.14% respectively under the same experimental configuration.

Number of Refinement Stages. Refinement strategy can significantly improve the performance of rotation detection, especially the introduction of feature refinement. Table 1 explores the relationship between the number of refinements and model performance. R^{3}Det† has joined the two refinement stages and bring more gain. To further explore the impact of the number of stages, several experimental results are summarized in Table 3. Experiments show that three or more refinements will not bring additional improvements to overall performance. We also find that ensemble multi-stage results can further improve detection performance.

Approximate SkewIoU Loss. We use two different detectors and three different SkewIoU functions to verify the effectiveness of the approximate SkewIoU, as shown in Table 4. RetinaNet-based detectors will have a large number of low-SkewIoU prediction bounding box in the early stage of training, and will produce very large loss after the $\log$ function, and training is prone to non-convergence. Compared with the linear function, the derivative of the $\exp$-based function is related to SkewIoU, that is, more attention is paid to the training of difficult samples, so it has a higher performance improvement. Compare with PIoU, we can achieve considerable gains on a higher baseline and far exceed PIoU in final performance, 73.8% versus 60.5% as shown in Table 5.

Comparison with the State-of-the-Art

Results on DOTA. The results on DOTA are shown in Table 5. The compared methods include i) one-stage methods, such as PIoU, P-RSDet (Zhou et al. 2020), O^2-DNet (Wei et al. 2019) , DRN ii) two stage methods, such as RoI-Transformer, SCRDet, CAD-Net (Zhang, Lu, and Zhang 2019), Gliding Vertex, Mask OBB (Wang et al. 2019), FFA (Fu et al. 2020), APE (Zhu, Du, and Wu 2020). Two-stage detectors are still dominant in DOTA, and the latest two-stage detection methods, such as ROI Transformer, SCRDet, and APE have performed well. However, they all use complex model structures in exchange for performance improvements, which are extremely low in terms of detection efficiency. The advantage of the two-stage method on the DOTA dataset lies in the multi-stage regression and the use of low-level feature maps (P2) that are friendly to small objects. Compared to all published single-stage methods, our method achieves the best performance without using multi-scale training and testing, at 73.8%. By using a stronger backbone and multi-scale training and testing, as used in the most advanced two-stage method CenterMask OBB, R^{3}Det performs competitive performance, about 76.5%.

Method	BB	MS	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
Two-stage																		
RoI-Trans.	R-101	✓	88.6	78.5	43.4	75.9	68.8	73.7	83.6	90.7	77.3	81.5	58.4	53.5	62.8	58.9	47.7	69.6
CAD-Net	R-101		87.8	82.4	49.4	73.5	71.1	63.5	76.7	**90.9**	79.2	73.3	48.4	60.9	62.0	67.0	62.2	69.9
SCRDet	R-101	✓	90.0	80.7	52.1	68.4	68.4	60.3	72.4	90.9	**87.9**	**86.9**	65.0	66.7	66.3	68.2	65.2	72.6
Gliding Ver.	R-101		89.6	85.0	52.3	**77.3**	73.0	73.1	86.8	90.7	79.0	86.8	59.6	**70.9**	72.9	70.86	57.3	75.0
Mask OBB	RX-101	✓	89.6	**86.0**	54.2	72.9	76.5	74.2	85.6	89.9	83.8	86.5	54.9	69.6	73.9	69.1	63.3	75.3
FFA	R-101	✓	**90.1**	82.7	54.2	75.2	71.0	79.9	83.5	90.7	83.9	84.6	61.2	68.0	70.7	76.0	63.7	75.7
APE	RX-101		90.0	83.6	53.4	76.0	74.0	77.2	79.5	90.8	87.2	84.5	**67.7**	60.3	**74.6**	71.8	65.6	75.8
One-stage																		
PIoU	DLA-34		80.9	69.7	24.1	60.2	38.3	64.4	64.8	90.9	77.2	70.4	46.5	37.1	57.1	61.9	64.0	60.5
P-RSDet	R-101	✓	89.0	73.7	47.3	72.0	70.6	73.7	72.8	90.8	80.1	81.3	59.5	57.9	60.8	65.2	52.6	69.8
O^2-DNet	H-104	✓	89.3	82.1	47.3	61.2	71.3	74.0	78.6	90.8	82.2	81.4	60.9	60.2	58.2	67.0	61.0	71.0
DRN	H-104	✓	89.7	82.3	47.2	64.1	76.2	74.4	85.8	90.6	86.2	84.9	57.7	61.9	69.3	69.6	58.5	73.2
$R^3Det^\dagger$	R-101		88.8	83.1	50.9	67.3	76.2	80.4	86.7	90.8	84.7	83.2	62.0	61.4	66.9	70.6	53.9	73.8
R^3Det	R-152	✓	89.8	83.8	48.1	66.8	**78.8**	**83.3**	**87.8**	90.8	85.4	85.5	65.7	62.7	67.5	**78.6**	**72.6**	**76.5**

Table 5: Detection accuracy on DOTA. $R^3Det^\dagger$ indicates that two refinement stages have been added. R-101 denotes ResNet-101 (likewise for R-50, R-152), RX-101 and H-104 represent ResNeXt101 (Xie et al. 2017) and Hourglass-104 (Newell, Yang, and Deng 2016). MS indicates that multi-scale training or testing is used. BB means Backbone.

Method	BB	Size	mAP (07)	mAP (12)	Speed
R^2CNN	R-101	800	73.07	79.73	5fps
RC1 & RC2	VGG16	–	75.7	–	–
RRPN	R-101	800	79.08	85.64	1.5fps
R^2PN	VGG16	–	79.6	–	–
RetinaNet-H	R-101	800	82.89	89.27	14fps
RRD	VGG16	384	84.3	–	–
RoI-Trans.	R-101	800	86.20	–	6fps
Gliding Ver.	R-101	–	88.20	–	–
DRN	H-104	–	–	92.70	–
SBD	R-50	–	–	93.70	–
R^3Det^*	R-101	800	89.14	94.98	4fps
RetinaNet-R	R-101	800	89.18	95.21	8fps
R^3Det	R-101	300	87.14	93.22	18fps
	R-101	600	88.97	94.61	15fps
	R-101	800	**89.26**	**96.01**	12fps
	M-V2	300	77.16	84.31	**23fps**
	M-V2	600	86.67	92.83	20fps
	M-V2	800	88.71	94.45	16fps

Table 6: Evaluation on HRSC2016. Number after mAP i.e. 07 (12) means using the 2007 (2012) evaluation metric. M-V2 denotes MobileNetV2.

Results on HRSC2016 and UCAS-AOD. The HRSC2016 is a challenging dataset that contains lots of large aspect ratio ship instances with arbitrary orientation. We use RRPN and R^2CNN for comparative experiments, which are originally used for scene text detection. Experimental results show that these two methods under-perform in the remote sensing dataset, only 73.07% and 79.08% respectively. Although RoI Transformer achieves 86.20% mAP, its detection speed is still not ideal, and only about 6fps without accounting for the post-processing operations. RetinNet-H, RetinaNet-R and R^3Det^* are the three baseline models used in this paper. RetinaNet-R achieves the best detection results, around 89.14%, which is consistent with the performance of the ship category in the DOTA dataset. This further illustrates that the rotation-based approach has advantages in large aspect ratio object detection. Under ResNet101 backbone, R^3Det can achieve better performance than RC1 & RC2 (Liu et al. 2017), R^2PN (Zhang et al. 2018b), RRD (Liao et al. 2018), Gliding Vertex, DRN, SDB (Liu et al. 2019) and above methods. Besides, our method can achieve 86.67% accuracy and 20fps speed, given MobileNetv2 (Sandler et al. 2018) as backbone with input image size 600. Table 7 illustrates the comparison of performance on UCAS-AOD dataset, our results are the best out of all the existing published methods (YOLOv2 (Redmon and Farhadi 2017), R-DFPN, DRBox (Liu, Pan, and Lei 2017), S^2ARN (Bao et al. 2019), FADet (Li et al. 2019)), at 96.17%.

Method	mAP	Plane	Car
YOLOv2	87.90	96.60	79.20
R-DFPN	89.20	95.90	82.50
DRBox	89.95	94.90	85.00
S^2ARN	94.90	97.60	92.20
RetinaNet-H	95.47	97.34	93.60
FADet	95.71	**98.69**	92.72
R^3Det	**96.17**	98.20	**94.14**

Table 7: Detection accuracy on UCAS-AOD.

Conclusion

We have presented an end-to-end refined single-stage detector designated for rotating objects with large aspect ratio, dense distribution and arbitrary orientations. Seeing the shortcoming of feature misalignment in existing refined single-stage detectors, we design a feature refinement module, whose key idea is to re-encode the position information of the current refined bounding box to the corresponding feature points through pixel-wise feature interpolation to achieve feature reconstruction and alignment. For more accurate rotation estimation, an approximate SkewIoU loss is devised to solve the problem that the calculation of SkewIoU is not derivable. Experiments across different datasets show competitive performance of our method regarding with both accuracy and efficiency.

Acknowledgments

This research was partially supported by China Major State Research Development Program (2018AAA0100704), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), NSFC (U20B2068, 72061127003). Xue Yang is partially supported by Wu Wen Jun Honorary Doctoral Scholarship, AI Institute, Shanghai Jiao Tong University.

References

Azimi, S. M.; Vig, E.; Bahmanyar, R.; Körner, M.; and Reinartz, P. 2018. Towards multi-class object detection in unconstrained remote sensing imagery. In *Asian Conference on Computer Vision*, 150–165. Springer.

Bao, S.; Zhong, X.; Zhu, R.; Zhang, X.; Li, Z.; and Li, M. 2019. Single Shot Anchor Refinement Network for Oriented Object Detection in Optical Remote Sensing Imagery. *IEEE Access* 7: 87150–87161.

Cai, Z.; and Vasconcelos, N. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6154–6162.

Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Shi, J.; Ouyang, W.; et al. 2019. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4974–4983.

Chen, X.; Yu, J.; Kong, S.; Wu, Z.; and Wen, L. 2018. Dual Refinement Networks for Accurate and Fast Object Detection in Real-World Scenes. *arXiv preprint arXiv:1807.08638* .

Chen, Z.; Chen, K.; Lin, W.; See, J.; Yu, H.; Ke, Y.; and Yang, C. 2020. PIoU Loss: Towards Accurate Oriented Object Detection in Complex Environments. *Proceedings of the European Conference on Computer Vision* .

Chi, C.; Zhang, S.; Xing, J.; Lei, Z.; Li, S. Z.; and Zou, X. 2019. Selective refinement network for high performance face detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8231–8238.

Dai, J.; Li, Y.; He, K.; and Sun, J. 2016. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems*, 379–387.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 764–773.

Ding, J.; Xue, N.; Long, Y.; Xia, G.-S.; and Lu, Q. 2019. Learning RoI Transformer for Oriented Object Detection in Aerial Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2849–2858.

Fu, K.; Chang, Z.; Zhang, Y.; Xu, G.; Zhang, K.; and Sun, X. 2020. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing* 161: 294–308.

Fu, K.; Li, Y.; Sun, H.; Yang, X.; Xu, G.; Li, Y.; and Sun, X. 2018. A ship rotation detection model in remote sensing images based on feature fusion pyramid network and deep reinforcement learning. *Remote Sensing* 10(12): 1922.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, 1440–1448.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 580–587.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Jang, H.-D.; Woo, S.; Benz, P.; Park, J.; and Kweon, I. S. 2019. Propose-and-Attend Single Shot Detector. *arXiv preprint arXiv:1907.12736* .

Jiang, Y.; Zhu, X.; Wang, X.; Yang, S.; Li, W.; Wang, H.; Fu, P.; and Luo, Z. 2017. R2CNN: rotational region CNN for orientation robust scene text detection. *arXiv preprint arXiv:1706.09579* .

Karatzas, D.; Gomez-Bigorda, L.; Nicolaou, A.; Ghosh, S.; Bagdanov, A.; Iwamura, M.; Matas, J.; Neumann, L.; Chandrasekhar, V. R.; Lu, S.; et al. 2015. ICDAR 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition*, 1156–1160. IEEE.

Li, A.; Yang, X.; and Zhang, C. 2019. Rethinking Classification and Localization for Cascade R-CNN. *arXiv preprint arXiv:1907.11914* .

Li, C.; Xu, C.; Cui, Z.; Wang, D.; Zhang, T.; and Yang, J. 2019. Feature-Attentioned Object Detection in Remote Sensing Imagery. In *2019 IEEE International Conference on Image Processing*, 3886–3890. IEEE.

Liao, M.; Shi, B.; and Bai, X. 2018. Textboxes++: A single-shot oriented scene text detector. *IEEE Transactions on Image Processing* 27(8): 3676–3690.

Liao, M.; Zhu, Z.; Shi, B.; Xia, G.-s.; and Bai, X. 2018. Rotation-sensitive regression for oriented scene text detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5909–5918.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017a. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2117–2125.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017b. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988.

Liu, L.; Pan, Z.; and Lei, B. 2017. Learning a Rotation Invariant Detector with Rotatable Bounding Box. *arXiv preprint arXiv:1711.09405* .

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision*, 21–37. Springer.

Liu, X.; Liang, D.; Yan, S.; Chen, D.; Qiao, Y.; and Yan, J. 2018. Fots: Fast oriented text spotting with a unified network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5676–5685.

Liu, Y.; Zhang, S.; Jin, L.; Xie, L.; Wu, Y.; and Wang, Z. 2019. Omnidirectional scene text detection with sequential-free box discretization. *arXiv preprint arXiv:1906.02371* .

Liu, Z.; Yuan, L.; Weng, L.; and Yang, Y. 2017. 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, 324–331.

Ma, J.; Shao, W.; Ye, H.; Wang, L.; Wang, H.; Zheng, Y.; and Xue, X. 2018. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia* .

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision*, 483–499. Springer.

Pan, X.; Ren, Y.; Sheng, K.; Dong, W.; Yuan, H.; Guo, X.; Ma, C.; and Xu, C. 2020. Dynamic Refinement Network for Oriented and Densely Packed Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11207–11216.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788.

Redmon, J.; and Farhadi, A. 2017. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7263–7271.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 91–99.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 658–666.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520.

Wang, J.; Ding, J.; Guo, H.; Cheng, W.; Pan, T.; and Yang, W. 2019. Mask obb: A semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images. *Remote Sensing* 11(24): 2930.

Wei, H.; Zhou, L.; Zhang, Y.; Li, H.; Guo, R.; and Wang, H. 2019. Oriented Objects as pairs of Middle Lines. *arXiv preprint arXiv:1912.10694* .

Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; and Zhang, L. 2018. DOTA: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3974–3983.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1492–1500.

Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; Chen, K.; Xia, G.-S.; and Bai, X. 2020. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Yang, J.; Ji, L.; Geng, X.; Yang, X.; and Zhao, Y. 2019a. Building detection in high spatial resolution remote sensing imagery with the U-Rotation Detection Network. *International Journal of Remote Sensing* 40(15): 6036–6058.

Yang, X.; Fu, K.; Sun, H.; Sun, X.; Yan, M.; Diao, W.; and Guo, Z. 2018a. Object Detection with Head Direction in Remote Sensing Images Based on Rotational Region CNN. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, 2507–2510. IEEE.

Yang, X.; Sun, H.; Fu, K.; Yang, J.; Sun, X.; Yan, M.; and Guo, Z. 2018b. 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.

Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; and Fu, K. 2019b. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In *Proceedings of the IEEE International Conference on Computer Vision*, 8232–8241.

Zhang, G.; Lu, S.; and Zhang, W. 2019. CAD-Net: A context-aware detection network for objects in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing* 57(12): 10015–10024.

Zhang, H.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2019. Cascade RetinaNet: Maintaining Consistency for Single-Stage Object Detection. *arXiv preprint arXiv:1907.06881* .

Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; and Li, S. Z. 2018a. Single-Shot Refinement Neural Network for Object Detection.

Zhang, Z.; Guo, W.; Zhu, S.; and Yu, W. 2018b. Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks. *IEEE Geoscience and Remote Sensing Letters* 15(11): 1745–1749.

Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; and Ren, D. 2020. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 12993–13000.

Zhou, L.; Wei, H.; Li, H.; Zhang, Y.; Sun, X.; and Zhao, W. 2020. Objects detection for remote sensing images based on polar coordinates. *arXiv preprint arXiv:2001.02988* .

Zhou, X.; Yao, C.; Wen, H.; Wang, Y.; Zhou, S.; He, W.; and Liang, J. 2017. EAST: An Efficient and Accurate Scene Text Detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zhu, H.; Chen, X.; Dai, W.; Fu, K.; Ye, Q.; and Jiao, J. 2015. Orientation robust object detection in aerial images using deep convolutional neural network. In *2015 IEEE International Conference on Image Processing*, 3735–3739. IEEE.

Zhu, Y.; Du, J.; and Wu, X. 2020. Adaptive period embedding for representing oriented objects in aerial images. *IEEE Transactions on Geoscience and Remote Sensing* .