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

Xue Yang[1,2], Qingqing Liu[3], Junchi Yan[1,2], Ang Li[4]
[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]Center South University. [4]Nanjing University of Science and Technology .
{yangxue-2019-sjtu, yanjunchi}@sjtu.edu.cn liuqingqing@csu.edu.cn angless@njust.edu.cn

Abstract

Rotation detection is a challenging task due to the difficulties of locating the multi-angle objects and separating them accurately and quickly from the background. Though considerable progress has been made, 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 positioning objects. Considering the shortcoming of feature misalignment in the current refined single-stage detector, we design a feature refinement module to improve detection performance, which is especially effective in the long tail data set. 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 feature interpolation to realize feature reconstruction and alignment. Extensive experiments on two remote sensing public datasets DOTA, HRSC2016 as well as scene text data ICDAR2015 show the state-of-the-art accuracy and speed of our detector. Source code and the models will be made public available upon the publish of the paper.

1. Introduction

Object detection is one of the fundamental tasks in computer vision, and many high-performance general-purpose object detections have been proposed. The current popular detection methods can be divided into two types: two-stage object detectors [11, 10, 29, 7, 22] and single-stage object detectors [24, 28, 23]. Two-stage methods have achieved promising results on a few benchmarks, while the single-stage approach maintains faster detection speeds.

However, the current general horizontal detectors are no longer sufficient for many practical applications. For in-

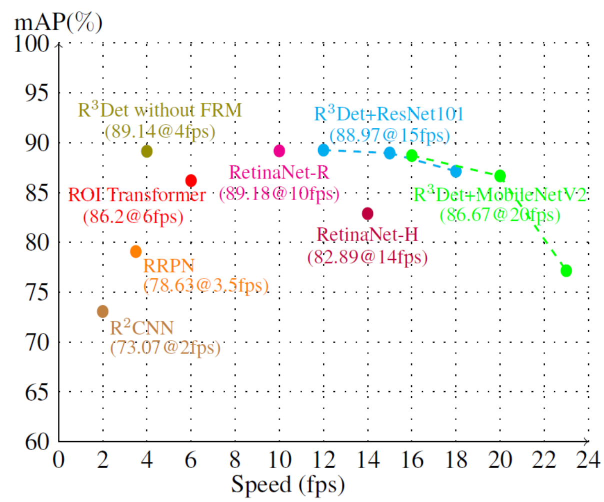


Figure 1: Performance versus speed on HRSC2016 [26] dataset. As can be seen, our algorithm significantly surpasses competitors in accuracy, whilst running very fast. Specific information is listed in Table 4.

stance, scene text detection and remote sensing object detection whereby the objects can be in any direction and position. Therefore, many rotation detectors based on a general detection framework have been proposed in the field of scene text and remote sensing. In particular, three challenges are pronounced for images in the above two fields, as analyzed as follows:

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

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

3. **Category unbalance.** Many multi-category rotated datasets are long-tailed datasets whose categories are ex-

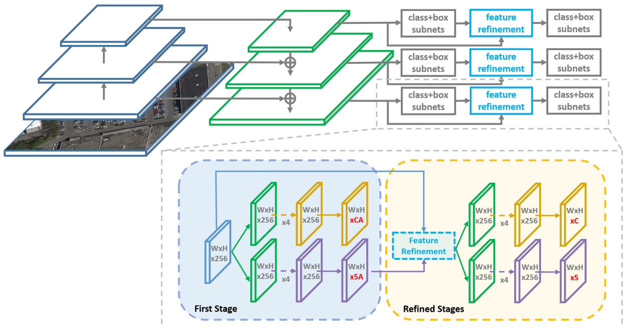
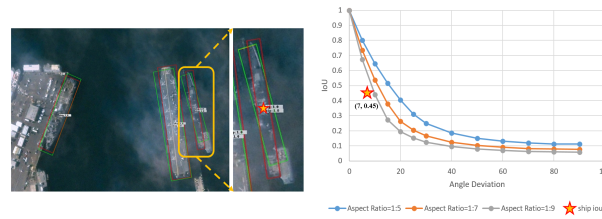


Figure 2: The architecture of the proposed Refined Rotation Single-Stage Detector (RetinaNet as a embodiment). The refinement stage can be repeated by multiple times. Only the bounding box with the highest score of each feature point is preserved in the refinement stage to increase the speed of the model. 'A' indicates the number of anchors on each feature point, and 'C' indicates the number of categories.



(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.

tremely unbalanced, as sketched in Figure 7a.

In this paper, we mainly discuss how to design an accurate and fast rotation detector. To maintain high positioning accuracy and detection 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 combination strategy of two forms of anchors is adopted in the refinement single-stage detector, that is, the horizontal anchors are used in the first stage for faster speed and more proposals, and then the refined rotating anchors are used in the refinement stages to adapt to intensive scenarios. Second, we also notice that the current refined single-stage detectors have feature misalignment problems [38, 6], 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 of the refined anchors and re-reconstruct the feature map to achieve the purpose of feature alignment. It is worth noting that FRM can also reduce the number of proposals in the refined stages, 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, especially for small sample categories. Combing these three techniques as a whole, our approach achieves state-of-the-art performance with high speed on three public rotating sensitive datasets including DOTA, HRSC2016 and ICDAR2015.

This work makes the following contributions:

1. For large aspect ratio objects, an accurate and fast rotation singe-stage detector is devised in a refined manner. Specifically, it is the process of using multiple bounding boxes and feature refinement.

2. For densely arranged scenes, we consider the advantages of each of the two forms of anchors, and adopt an anchor combination strategy to enable the detector to cope with intensive scenarios with high efficiency.

3. For category unbalance, we propose a FRM that makes the detector features more accurate and reliable during the refinement stages. Experiments show that FRM has greatly improved the category that are not fully learned due to the small number of samples and inaccurate features, such as BD, GTF, BC, SBF, RA, HC, which increased by 4.09%, 2.83%, 3.4%, 4.82%, 1.22%, and 19.26%, respectively.

2. Related Work

Two-Stage Object Detectors. Most of the current two-stage methods are region-based. In a region based framework, category-independent region proposals are generated from an image in the first stage, features are extracted from these regions subsequently, and then category-specific classifiers and regressors are used for classification and regression in the second satge. Finally, the detection results are obtained by using post-processing methods such as non-maximum suppression (NMS). Faster-RCNN [29] is a classic structure in a two-stage approach that can detect object quickly and accurately in an end-to-end manner. Many high-performance detection methods are proposed today, such as R-FCN [7], FPN [22], Light Head RCNN [19], etc.

Single-Stage Object Detectors. For their efficiency, single-stage detection methods are receiving more and more attention. OverFeat [31] is one of the first single-stage detectors based on convolutional neural networks. It performs object detection in a multiscale sliding window fashion via a single forward pass through the CNN. Compared with region based methods, Redmon et al. [28] 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,

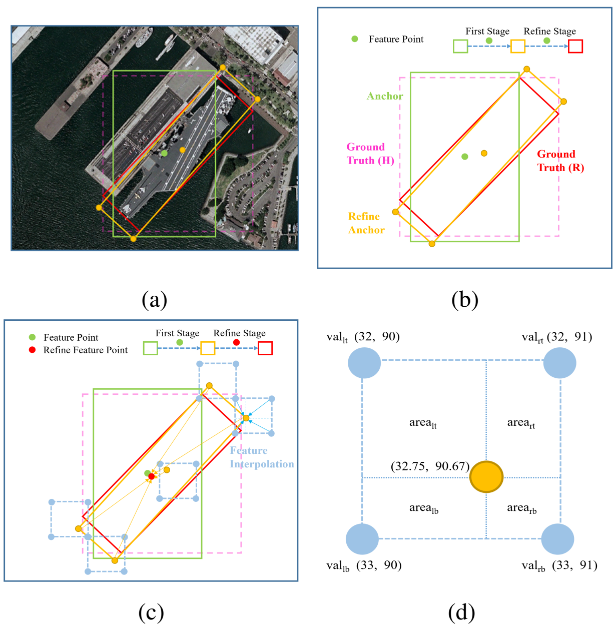


Figure 4: Principle and analysis of feature refinement module. (a) Original image. (b) Refine box without feature interpolation. (c) Refine box with feature interpolation. (d) Feature interpolation.

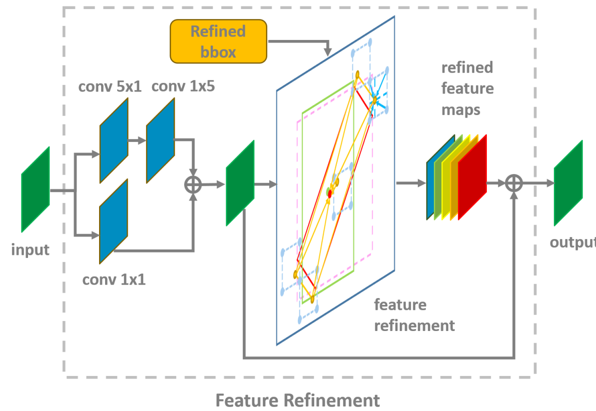


Figure 5: Feature Refinement Module.

Liu et al. [24] propose SSD. The work [23] solves the class imbalance problem by proposing RetinaNet with Focal loss and further improves the accuracy of single-stage detector.

Rotation Object Detectors. Remote sensing and scene text 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, the two-stage rotation detector is still dominant. ICN [2], ROI-Transformer [8] and SCRDet [36], achieves the most advanced performance available today. However, they used a more complicated structure, which greatly reduced the speed. As for scene text detection, there are many efficient rotation detection methods, such as two-stage methods: R^2CNN [15], RRPN [**?**], FOTS [25], etc., and EAST [40], TextBoxes [20], etc. based on single-stage methods.

Refined Object Detectors. To achieve better positioning accuracy, many cascaded or refined detectors are proposed. The Cascade RCNN [3], HTC [4], and FSCascade [18] performed multiple classifications and regressions in the second stage, which greatly improved the classification accuracy and positioning accuracy. The same idea is also used in single-stage detectors, such as RefineDet [38]. Unlike the two-stage detectors, which use RoI Pooling [10] or RoI Align [12] for feature alignment, the currently refined single-stage detector is not well resolved in this respect. Althought many papers [5, 14, 37] use deformable convolution for feature alignment, whose offset parameters are often obtained by learning or calculating the offset between the predefined anchor box and the refined anchor. The essence of these deformable-based feature alignment methods is to expand the receptive field, which is too implicit and can not ensure that features are truely aligned. Feature misalignment still limits the performance of the refined single-stage detector.

3. The Proposed Method

We first give an overview of our method as sketched in Figure 2. The embodiment is a single-stage rotation detector based on the RetinaNet [23], namely Refined Rotation RetinaNet (R^3Det). 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. In a single-stage rotating object detection task, continuous refinement of the predicted bounding box can improve the regression accuracy, and feature refinement is a necessary process for this purpose. It should be noted that FRM can also be used on other single-stage detectors (such as SSD), refer to the discussion section.

3.1. Rotation RetinaNet

RetinaNet is one of the most advanced single-stage detectors available today. It consists of two parts: backbone network, classification and regression subnetwork. RetinaNet adopts the Feature Pyramid Network (FPN) [22] as the backbone network. In brief, FPN augments a convolutional network with a top-down pathway and lateral connections so the network efficiently constructs a rich, multi-scale feature pyramid from a single resolution input image. Each level of the pyramid can be used for detecting objects at a different scale. Besides, each layer of the FPN is connected to a classification subnet and a regression subnet for predict-

Method	Backbone	FRM	Data Aug.	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
RetinaNet-H (baseline)	ResNet50	×	×	88.87	74.46	40.11	58.03	63.10	50.61	63.63	**90.89**	77.91	76.38	48.26	55.85	50.67	60.23	34.23	62.22
RetinaNet-R (baseline)	ResNet50	×	×	88.92	67.67	33.55	56.83	66.11	73.28	75.24	90.87	73.95	75.07	43.77	56.72	51.05	55.86	21.46	62.02
R^{3}Det (baseline)	ResNet50	×	×	88.78	70.60	41.97	57.05	68.28	67.31	69.35	90.85	74.31	79.96	46.16	57.12	52.57	58.58	24.26	63.14
R^{3}Det (proposed)	ResNet50	√	×	88.78	74.69	41.94	59.88	68.90	69.77	69.82	90.81	77.71	80.40	50.98	58.34	52.10	58.30	43.52	65.73
R^{3}Det† (proposed)	ResNet50	√	×	88.97	73.80	40.28	57.91	68.77	72.22	72.62	90.80	78.40	76.80	52.94	54.21	50.70	64.31	40.77	65.57
R^{3}Det (proposed)	ResNet50	√	√	89.30	80.29	46.21	65.07	70.51	73.38	77.42	90.83	80.59	82.26	59.29	58.25	57.75	65.90	55.31	70.16
R^{3}Det (proposed)	ResNet101	√	√	**89.54**	**81.99**	48.46	62.52	70.48	74.29	77.54	90.80	81.39	83.54	61.97	**59.82**	65.44	67.46	60.05	71.69
R^{3}Det (proposed)	ResNet152	√	√	89.24	80.81	**51.11**	**65.62**	**70.67**	**76.03**	**78.32**	90.83	**84.89**	**84.42**	**65.10**	57.18	**68.10**	**68.98**	**60.88**	**72.81**

Table 1: Ablative study of each components in our proposed method on the DOTA dataset. The short names for categories are defined as: PL-Plane, BD-Baseball diamond, BR-Bridge, GTF-Ground field track, SV-Small vehicle, LV-Large vehicle, SH-Ship, TC-Tennis court, BC-Basketball court, ST-Storage tank, SBF-Soccer-ball field, RA-Roundabout, HA-Harbor, SP-Swimming pool, and HC-Helicopter. For RetinaNet, 'H' and 'R' represent the horizontal and rotating anchors, respectively. R^{3}Det† indicates that two refinement stages have been added.

ing categories and locations. Note that the object classification subnet and the box regression subnet, though sharing a common structure, use separate parameters. The most important thing is that RetinaNet has proposed focal loss and solved the problem caused by category imbalance, which greatly improved the accuracy of single-stage detector.

To achieve RetinaNet-based rotation detection, we use five parameters (x, y, w, h, θ) to represent arbitrary-oriented rectangle. Ranging in $[-\pi/2, 0)$, θ denotes the acute angle to the x-axis, and for the other side we refer it as w. Therefore, it calls for predicting an additional angular offset in the regression subnet, whose rotation bounding box is:

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

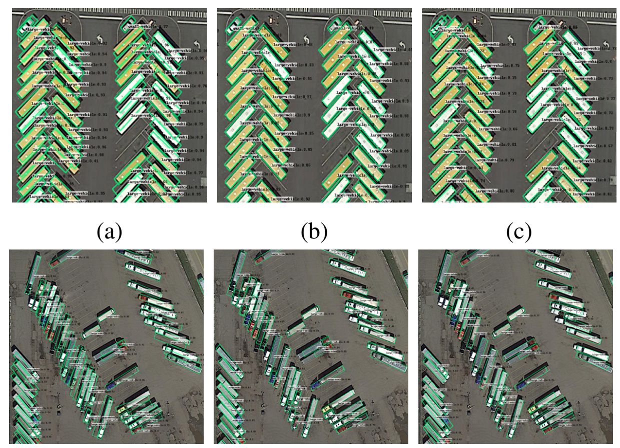
$$\begin{aligned} t_x^{'} &= (x^{'} - x_a)/w_a, t_y^{'} = (y^{'} - y_a)/h_a \\ t_w^{'} &= \log(w^{'}/w_a), t_h^{'} = \log(h^{'}/h_a), t_\theta^{'} = \theta^{'} - \theta_a \end{aligned} \tag{2}$$

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).

The multi-task loss is used which is defined as follows:

$$\begin{aligned} L = &\frac{\lambda_1}{N} \sum_{n=1}^{N} t_n^{'} \sum_{j \in \{x,y,w,h,\theta\}} L_{reg}(v_{nj}^{'}, v_{nj}) \\ &+ \frac{\lambda_2}{N} \sum_{n=1}^{N} L_{cls}(p_n, t_n) \end{aligned} \tag{3}$$

where N indicates the number of anchors, t_n represents the label of object, p_n is the probability distribution of various classes calculated by Sigmoid function. $v_{*j}^{'}$ represents the predicted offset vectors, v_{*j} represents the targets vector of ground-truth. The hyper-parameter λ_1, λ_2 control the trade-off and are set to 1 by default. In addition, the classification loss L_{cls} is focal loss [23]. The regression loss L_{reg} is smooth L1 loss as defined in [10].



(a) (b) (c)

(d) (e) (f)

Figure 6: Visualization of three baselines on the DOTA dataset. (a)(d) RetinaNet-H. (b)(e) RetinaNet-R. (c)(f) R^{3}Det. Here 'H' and 'R' represent the horizontal and rotating anchors, respectively.

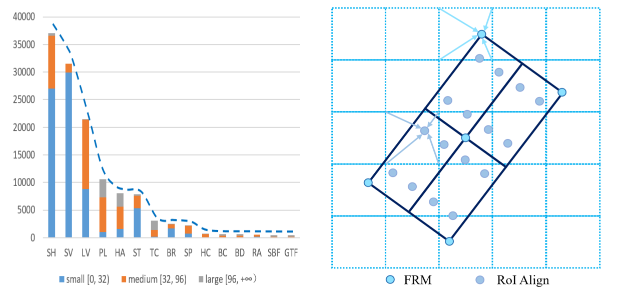
3.2. Refined Rotation RetinaNet

Refined Detection. The Skew Intersection over Union (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 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, remaining refinement stage uses 0.6 and 0.5, respectively. 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



(a) (b)

Figure 7: (a) The quantity of each category in the DOTA. (b) Comparison between RoI Align and FRM.

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, 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), thereby reconstructing the entire feature map to achieve the alignment of the features. The whole process is shown in Figure 4c. To accurately obtain the location feature information of the refined bounding box, we adopt the bilinear feature interpolation method, as shown in Figure 4d. Specifically, feature interpolation can be formulated as follows:

$$\begin{aligned} val = {} & val_{lt} * area_{rb} + val_{rt} * area_{lb} \\ & + val_{rb} * area_{lt} + val_{lb} * area_{rt} \end{aligned} \tag{5}$$

According to the above analysis and method, a feature refinement module is proposed, and its structure is shown in 5. Specifically, the feature map is added by two-way convolution to obtain a new feature, and then a feature-aligned feature map is obtained by performing feature interpolation on the five corner points of the refined boungding box. We only add a small number of parameters in refinement stage to make the comparison as fair as possible, and only the bounding box with the highest score of each feature point is preserved in the refinement stage to increase the speed.

4. Experiments

Tests are implemented by TensorFlow [1] on a server with GeForce RTX 2080 Ti and 11G memory. We perform experiments on both aerial benchmarks and scene text benchmarks to verify the generality of our techniques.

4.1. Datasets and Protocls

The benchmark DOTA [34] is for object detection in aerial images. It contains 2,806 aerial images from different sensors and platforms. The image size ranges from around 800×800 to $4,000 \times 4,000$ pixels and contains objects exhibiting a wide variety of scales, orientations, and shapes. These images are then annotated by experts using 15 common object categories. The fully annotated DOTA benchmark contains 188,282 instances, each of which is labeled by an arbitrary quadrilateral. There are two detection tasks for DOTA: horizontal bounding boxes (HBB) and oriented bounding boxes (OBB). Half of the original images are randomly selected as the training set, 1/6 as the validation set, and 1/3 as the testing set. We divide the images into 600×600 subimages with an overlap of 150 pixels and scale it to 800×800. With all these processes, we obtain about 27,000 patches. The model is trained by 135k iterations in total, and the learning rate changes during the 81k and 108k iterations from 5e-4 to 5e-6.

mAP	Feature Refinement Interpolation Formula	status
65.73	$val_{lt} * area_{rb} + val_{rt} * area_{lb} + val_{rb} * area_{lt} + val_{lb} * area_{rt}$	right
64.28	$val_{lt} * area_{lt} + val_{rt} * area_{rt} + val_{rb} * area_{rb} + val_{lb} * area_{lb}$	wrong
64.37	$val_{lt} * area_{lb} + val_{rt} * area_{rb} + val_{rb} * area_{rt} + val_{lb} * area_{lt}$	wrong

Table 2: Experiments with different interpolation formulas. Feature interpolation has position-sensitive properties.

The HRSC2016 dataset [26] contains images from two scenarios including ships on sea and ships close inshore. All the images are collected from six famous harbors. The image sizes range from 300×300 to $1,500 \times 900$. The training, validation and test set include 436 images, 181 images and 444 images, respectively. For all experiments we use an image scale of 800×800 for training and testing. we train the model with 5e-4 learning rate for the first 30k iterations, then 5e-5 and 5e-6 for the other two 10k iterations.

ICDAR2015 is used in Challenge 4 of ICDAR 2015 Robust Reading Competition [16]. It includes a total of 1500 pictures, 1000 of which are used for training and the remaining are for testing. The text regions are annotated by 4 vertices of the quadrangle. We use its origin image size 720×1280 for training and testing. The ICDAR2015 dataset uses the same learning strategy and changes the learning rate size in 15k iterations, 20k iterations, and 25k iterations, respectively.

We experiment with ResNet-FPN and MobileNetv2-FPN [30] backbones. All backbones are pre-trained on ImageNet [17]. Besides, weight decay and momentum are 0.0001 and 0.9, respectively. We employ MomentumOptimizer over 8 GPUs with a total of 8 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.

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
Two-stage methods																
R-FCN [7]	37.80	38.21	3.64	37.26	6.74	2.60	5.59	22.85	46.93	66.04	33.37	47.15	10.60	25.19	17.96	26.79
FR-H [29]	47.16	61.00	9.80	51.74	14.87	12.80	6.88	56.26	59.97	57.32	47.83	48.70	8.23	37.25	23.05	32.29
FR-O [34]	79.09	69.12	17.17	63.49	34.20	37.16	36.20	89.19	69.60	58.96	49.4	52.52	46.69	44.80	46.30	52.93
R-DFPN [35]	80.92	65.82	33.77	58.94	55.77	50.94	54.78	90.33	66.34	68.66	48.73	51.76	55.10	51.32	35.88	57.94
R^2CNN [15]	80.94	65.67	35.34	67.44	59.92	50.91	55.81	90.67	66.92	72.39	55.06	52.23	55.14	53.35	48.22	60.67
RRPN [27]	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 [2]	81.40	74.30	47.70	70.30	64.90	67.80	70.00	90.80	79.10	78.20	53.60	62.90	67.00	64.20	50.20	68.20
RoI-Transformer [8]	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
SCRDet [36]	**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 methods																
SSD [9]	39.83	9.09	0.64	13.18	0.26	0.39	1.11	16.24	27.57	9.23	27.16	9.09	3.03	1.05	1.01	10.59
YOLOv2 [28]	39.57	20.29	36.58	23.42	8.85	2.09	4.82	44.34	38.35	34.65	16.02	37.62	47.23	25.5	7.45	21.39
R^3Det+ResNet101	89.54	**81.99**	48.46	62.52	70.48	74.29	77.54	90.80	81.39	83.54	61.97	59.82	65.44	67.46	60.05	71.69
R^3Det+ResNet152	89.24	80.81	51.11	65.62	**70.67**	**76.03**	78.32	90.83	84.89	84.42	**65.10**	57.18	**68.10**	**68.98**	60.88	**72.81**

Table 3: Detection accuracy on different objects and overall performance with the state-of-the-art methods on DOTA.

Method	FRM	Backbone	Image Size	Data Aug.	mAP	Speed
R^2CNN [15]	-	ResNet101	800*800	×	73.07	2fps
RC1 & RC2 [26]	-	VGG16	-	-	75.7	slow
RRPN [27]	-	ResNet101	800*800	×	79.08	3.5fps
R^2PN [39]	-	VGG16	-	√	79.6	slow
RetinaNet-H	-	ResNet101	800*800	√	82.89	14fps
RRD [21]	-	VGG16	384*384	-	84.3	slow
RetinaNet-R	-	ResNet101	800*800	√	89.18	10fps
RoI-Transformer [8]	-	ResNet101	512*800	×	86.20	6fps
R^3Det (proposed)	×	ResNet101	800*800	√	89.14	4fps
	√	ResNet152	800*800	√	**89.33**	10fps
	√	ResNet101	300*300	√	87.14	18fps
	√	ResNet101	600*600	√	88.97	15fps
	√	ResNet101	800*800	√	89.26	12fps
	√	MobileNetV2	300*300	√	77.16	**23fps**
	√	MobileNetV2	600*600	√	86.67	20fps
	√	MobileNetV2	800*800	√	88.71	16fps

Table 4: Comparison of the accuracy and speed of different methods on the HRSC2016 dataset.

4.2. Robust Baseline Methods

It is necessary to establish some robust baseline methods before verifying the effectiveness of the proposed method. From the perspective of the anchor, we analyze the effect of two forms of anchor on the speed and accuracy of the detection method, and finally construct a compromised robust baseline method.

The anchor setting is critical for region-based detection models. Both the horizontal anchor and the rotating anchor can achieve the purpose of rotation detection, but they have their own advantages and disadvantages. 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 and Figure 6d. In contrast, in Figure 6b and Figure 6e, 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 anchors has multiplied, making the model less efficient.

Method	FRM	Recall	Precision	F-measure	Res.	Device	FPS
CTPN [33]	-	51.56	74.22	60.85	-	-	-
SegLink [32]	-	76.80	73.10	75.00	-	-	-
RRPN [27]	-	82.17	73.23	77.44	-	-	slow
EAST [40]	-	78.33	83.27	80.72	720p	Titan X	13.2
Deep direct regression [13]	-	80.00	82.00	81.00	-	-	slow
R^2CNN [15]	-	79.68	85.62	82.54	720p	K80	0.44
FOTS RT [25]	-	**85.95**	79.83	82.78	720p	Titan X	**24**
R^3Det (proposed)	×	81.64	84.97	83.27	720p	2080 Ti	4
	√	83.54	**86.43**	**84.96**	720p	2080 Ti	13.5

Table 5: Comparison of the accuracy and speed of different methods on the ICDAR2015 dataset.

The performance of the single-stage detection method based on two forms of anchor (RetinaNet-H and RetinaNet-R) on the DOTA data set OBB task is shown in Table 1. In general, they have similar overall performance (62.22% versus 62.02%), 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, such as small vehicle, large vehicle, and ship. To more effectively verify the validity of the feature refinement module, we also build a refined rotation detector, which does not refine the feature. Since the number of anchors does not decrease before and after the refinement stage, the number of original anchors determines the speed of the model. Taking into account the speed and accuracy, we adopt an anchor combination strategy. Specifically, we first use horizontal anchors to reduce the number of proposals and increase the object recall rate, and then use the rotating refined anchor to overcome the problems caused by dense scenes, as shown in 6c and Figure 6f. In the end, the refined rotation detector achieves 63.14% performance, and better than RetinaNet-H and RetinaNet-R.

4.3. Ablation Study

Feature Refinement Module. Although R^3Det without FRM has been successful, it is incremental and only improves performance by about 1%. We believe that the main reason is that the anchor is not consistent with the feature

(a) BC and TC (b) SBF, GTF, TC and SP (c) HA (d) HA and SH

(e) SP (f) RA and SV (g) ST (h) BD and RA

(i) SV and LV (j) PL and HC (k) BR

Figure 8: Examples on DOTA. Our method performs better on those with large aspect ratio , in arbitrary direction, and high density.

map after the box refinement. FRM reconstructed the feature map based on the refined anchor, which increased the overall performance by 2.59% to 65.73% according to Table 1. We count the number of objects for each category, as shown in Figure 7a. Coincidentally, FRM has greatly improved the category that are not fully learned due to the small number of samples and inaccurate features, such as BD, GTF, BC, SBF, RA, HC, which increased by 4.09%, 2.83%, 3.4%, 4.82%, 1.22%, and 19.26%, respectively. We speculate that feature map alignment will facilitate few-shot learning.

Number of Refinement Stages. We have known that adding a refinement stage has a significant improvement in rotation detection, especially the introduction of feature refinement. How about joining multiple refinements? $R^3Det^{\dagger}$ in Table 1 has joined the two refinement stages, and the experiment shows that multiple refinements will not bring additional improvements to overall performance. Despite this,

(a) (b)

(c) (d)

Figure 9: Text detection results on the ICDAR2015 benchmarks.

there are still significant improvements in the three categories of SH, LV and SP.

Data Augmentation and Backbone Data augmentation is one of the important means to improve the model's ability. We improve the performance of the model from 65.57% to 70.16% by random horizontal, vertical flipping, random graying, and random rotation. In addition, we also explore the gain of the backone for the model. Under ResNet101 and ResNet152 as the backbone, we observe a reasonable improvement in table 1 (70.16% $\rightarrow$ 71.69% $\rightarrow$ 72.81%).

Feature Refinement Interpolation Formula. When we randomly disturb the order of the four weights in the interpolation formula, the final performance of the model will be greatly reduced, as shown in Table 2. This phenomenon reflects the location sensitivity of the feature points and explains why the performance of the model can be greatly improved after the feature is correctly refined.

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

The proposed R^3Det with FRM is compared to state-of-the-art object detectors on three datasets: DOTA [8], HRSC2016 [26] and ICDAR2015 [16]. Our model achieves competitive performances and outperforms all other models without any bells or whistles.

Results on DOTA. We compare our results with the state-of-the-arts in DOTA as depicted in Table 3. The results of DOTA reported here are obtained by submitting our predictions to the official DOTA evaluation server[1]. The current two-stage detector is still the most popular method in DOTA dataset research, and the latest two-stage detection methods, such as ICN, ROI Transformer, and SCRDet, have performed well. However, they all use complex model structures in exchange for performance improvements, which are extremely low in terms of detection

[1]https://captain-whu.github.io/DOTA/

Model	Backbone	FRM	DOTA	HRSC2016
SSD-H		-	60.15	82.21
Refined SSD	VGG16	×	62.79	88.48
Refined SSD		√	**65.79**	**89.32**

Table 6: Performance verification of FRM in SSD.

efficiency. For the time being, the single-stage detection method has not achieved satisfactory results in the large remote sensing dataset. The single-stage detection method proposed in this paper achieves comparable performance with the most advanced two-stage method, while maintaining a fast detection speed. The speed analysis is detailed in the next section, and the detection results for each class on DOTA are shown in Figure 8.

Results on HRSC2016. The HRSC2016 contains lots of large aspect ratio ship instances with arbitrary orientation, which poses a huge challenge to the positioning accuracy of the detector. We used RRPN [**?**] and R^2CNN [15] for comparative experiments, which were originally used for scene text detection. Experiments show that these two methods do not have competitive performance in the remote sensing dataset, only 73.07% and 79.08% respectively. Although RoI Transformer [8] achieved 86.20% results without data augmentation, its detection speed is still not ideal, and only about 6fps without calculating post-processing operations. RetinNet-H, RetinaNet-R and R^3Det without FRM 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 target detection. Under ResNet101 backbone, our model achieves state-ofthe-art performances.

Results on ICDAR2015. Scene text detection is also one of the main application scenarios for rotation detection. As you see in Table 5, our method achieves 84.96% while maintaining 13.5fps in the ICDAR2015 dataset, better than most mainstream algorithms. Once again, the validity of the structure proposed in this paper is proved. It also shows that the proposed techniques are general that can be useful for both aerial images and scene text images. The detection results on ICDAR2015 are shown in Figure 9.

4.5. Speed Comparison

It is worth noting that we only add a small number of parameters in refinement stage to make the comparison as fair as possible. When we use the FRM, only the bounding box with the highest score of each feature point is preserved in the refinement stage to increase the speed of the model. We compare the speed and accuracy with the other six methods on the HRSC2016 dataset. The time of post process (i.e. R-NMS) is included. At the same time, we also explore the impact of different backbones and image sizes on the performance of the proposed model. The detailed experimental results are shown in Table 4 and Figure 1. Our method can achieve 86.67% accuracy and 20fps speed, when the backbone is MobileNetv2, and the input image size is 600×600.

5. Discussion

Comparison between RoI Align and FRM. RoI Align is the key to maintaining feature alignment in a two-stage detector. RoI Align uses bilinear interpolation to compute the exact values of the input features at four regularly sampled locations in each RoI bin, and aggregate the result (using max or average), see Figure 7b for details. No quantization is performed on any coordinates involved in the RoI, its bins, or the sampling points. In contrast, FRM also uses the method of feature interpolation, but only for the five corners of RoI. FRM can be considered as a simplified and fast version of RoI Align and is more suitable for single-stage detectors, also with reference to Figure 7b.

Performance verification of FRM in SSD. We also verify the portability of FRM on different data sets based on SSD, see Table 6 for detailed results. FRM brings 3% and 0.84% gain in the datasets DOTA and HRSC2016, respectively. This indicates that the FRM has excellent model migration capabilities.

6. Conclusion

In this paper, we have presented an end-to-end refined singl-stage detector designated for rotating objects with large aspect ratio, dense distribution and category extremely imbalance, which are common in aerial and scene text image. Considering the shortcoming of feature misalignment in the current refined single-stage detector, we design a feature refinement module to improve detection performance, which is especially effective in the long tail data set. The key idea of FRM is to re-encode the position information of the current refined bounding box to the corresponding feature points through feature interpolation to realize feature reconstruction and alignment. We perform careful ablation experiments and comparative experiments on multiple rotation detection data sets such as DOTA, HRSC2016, and ICDAR2015, and demonstrate that our method achieves the state-of-the-art detection accuracy with high efficiency.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

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

[3] Z. Cai and N. Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

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

[5] X. Chen, J. Yu, S. Kong, Z. Wu, and L. Wen. Dual refinement networks for accurate and fast object detection in real-world scenes. *arXiv preprint arXiv:1807.08638*, 2018.

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

[7] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.

[8] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu. Learning roi transformer for detecting oriented objects in aerial images. In *Proc. CVPR*, 2019.

[9] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017.

[10] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[11] R. Girshick, J. Donahue, T. Darrell, and J. 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.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[13] W. He, X. Y. Zhang, F. Yin, and C. L. Liu. Deep direct regression for multi-oriented scene text detection. 2017.

[14] H.-D. Jang, S. Woo, P. Benz, J. Park, and I. S. Kweon. Propose-and-attend single shot detector. *arXiv preprint arXiv:1907.12736*, 2019.

[15] Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, P. Fu, and Z. Luo. R2cnn: rotational region cnn for orientation robust scene text detection. *arXiv preprint arXiv:1706.09579*, 2017.

[16] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. 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.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[18] A. Li, X. Yang, and C. Zhang. Rethinking classification and localization for cascade r-cnn. *arXiv preprint arXiv:1907.11914*, 2019.

[19] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun. Light-head r-cnn: In defense of two-stage object detector. *arXiv preprint arXiv:1711.07264*, 2017.

[20] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu. Textboxes: A fast text detector with a single deep neural network. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[21] M. Liao, Z. Zhu, B. Shi, G.-s. Xia, and X. 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.

[22] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In *CVPR*, volume 1, page 4, 2017.

[23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[24] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[25] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, and J. Yan. Fots: Fast oriented text spotting with a unified network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5676–5685, 2018.

[26] Z. Liu, L. Yuan, L. Weng, and Y. Yang. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *Proc. ICPRAM*, volume 2, pages 324–331, 2017.

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

[28] J. Redmon, S. Divvala, R. Girshick, and A. 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.

[29] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):1137–1149, 2017.

[30] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. 2018.

[31] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[32] B. Shi, X. Bai, and S. Belongie. Detecting oriented text in natural images by linking segments. 2017.

[33] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao. Detecting text in natural image with connectionist text proposal network. In *European conference on computer vision*, pages 56–72. Springer, 2016.

[34] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proc. CVPR*, 2018.

[35] X. Yang, H. Sun, K. Fu, J. Yang, X. Sun, M. Yan, and Z. Guo. Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks. *Remote Sensing*, 10(1):132, 2018.

[36] X. Yang, J. Yang, J. Yan, Y. Zhang, T. Zhan, Z. Guo, S. Xian, and K. Fu. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In *Proc. ICCV*, 2019.

[37] H. Zhang, H. Chang, B. Ma, S. Shan, and X. Chen. Cascade retinanet: Maintaining consistency for single-stage object detection. *arXiv preprint arXiv:1907.06881*, 2019.

[38] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li. Single-shot refinement neural network for object detection. 2018.

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

[40] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang. East: An efficient and accurate scene text detector. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.