PIoU Loss: Towards Accurate Oriented Object Detection in Complex Environments

Zhiming Chen[1,2], Kean Chen[2], Weiyao Lin[2⋆], John See[3], Hui Yu[1], Yan Ke[1], and Cong Yang[1⋆]

[1] Clobotics, China
[2] Department of Electronic Engineering, Shanghai Jiao Tong University, China
[3] Faculty of Computing and Informatics, Multimedia University, Malaysia

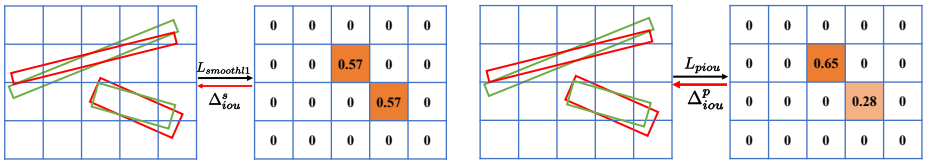
Abstract. Object detection using an oriented bounding box (OBB) can better target rotated objects by reducing the overlap with background areas. Existing OBB approaches are mostly built on horizontal bounding box detectors by introducing an additional angle dimension optimized by a distance loss. However, as the distance loss only minimizes the angle error of the OBB and that it loosely correlates to the IoU, it is insensitive to objects with high aspect ratios. Therefore, a novel loss, Pixels-IoU (PIoU) Loss, is formulated to exploit both the angle and IoU for accurate OBB regression. The PIoU loss is derived from IoU metric with a pixel-wise form, which is simple and suitable for both horizontal and oriented bounding box. To demonstrate its effectiveness, we evaluate the PIoU loss on both anchor-based and anchor-free frameworks. The experimental results show that PIoU loss can dramatically improve the performance of OBB detectors, particularly on objects with high aspect ratios and complex backgrounds. Besides, previous evaluation datasets did not include scenarios where the objects have high aspect ratios, hence a new dataset, Retail50K, is introduced to encourage the community to adapt OBB detectors for more complex environments.

Keywords: Orientated Object Detection; IoU Loss.

1 Introduction

Object detection is a fundamental task in computer vision and many detectors [34, 25, 21, 17] using convolutional neural networks have been proposed in recent years. In spite of their state-of-the-art performance, those detectors have inherent limitations on rotated and densely crowded objects. For example, bounding boxes (BB) of a rotated or perspective-transformed objects usually contain a significant amount of background that could mislead the classifiers. When bounding boxes have high overlapping areas, it is difficult to separate the densely crowded objects. Because of these limitations, researchers have extended existing detectors with oriented bounding boxes (OBB). In particular, as opposed to the BB which is denoted by (c_x, c_y, w, h), an OBB is composed by (c_x, c_y, w, h, θ)

⋆ Corresponding author: cong.yang@clobotics.com, wylin@sjtu.edu.cn



(a) $IOU_1 = 0.35, L_{smoothl1} = 0.57; IOU_2 = 0.72, L_{smoothl1} = 0.57$ (b) $IOU_1 = 0.35, L_{PIoU} = 0.65; IOU_2 = 0.72, L_{PIoU} = 0.28$

Fig. 1. Comparison between PIoU and SmoothL1 [34] losses. (a) Loss values between IoU and SmoothL1 are totally different while their SmoothL1 loss values are the same. (b) The proposed PIoU loss is consistent and correlated with IoU.

where (c_x, c_y), (w, h) and θ are the center point, size and rotation of an OBB, respectively. As a result, OBBs can compactly enclose the target object so that rotated and densely crowded objects can be better detected and classified.

Existing OBB-based approaches are mostly built on anchor-based frameworks by introducing an additional angle dimension optimized by a distance loss [24, 18, 19, 6, 41, 43] on the parameter tuple (c_x, c_y, w, h, θ). While OBB has been primarily used for simple rotated target detection in aerial images [18, 50, 31, 26, 23, 1, 39], the detection performance in more complex and close-up environments is limited. One of the reasons is that the distance loss in those approaches, *e.g.* SmoothL1 Loss [34], mainly focus on minimizing the angle error rather than global IoU. As a result, it is insensitive to targets with high aspect ratios. An intuitive explanation is that object parts far from the center (c_x, c_y) are not properly enclosed even though the angle distance may be small. For example, [19, 6] employ a regression branch to extract rotation-sensitive features and thereby the angle error of the OBB can be modelled in using a transformer. However, as shown in Figure 1(a), the IoU of predicted boxes (green) and that of the ground truth (red) are very different while their losses are the same.

To solve the problem above, we introduce a novel loss function, named *Pixels-IoU (PIoU) Loss*, to increase both the angle and IoU accuracy for OBB regression. In particular, as shown in Figure 1(b), the PIoU loss directly reflects the IoU and its local optimum compared to standard distance loss. The rationale behind this is that the IoU loss normally achieves better performance than the distance loss [45, 35]. However, the IoU calculation between OBBs is more complex than BBs since the shape of intersecting OBBs could be any polygon of less than eight sides. For this reason, the PIoU, a continuous and derivable function, is proposed to jointly correlate the five parameters of OBB for checking the position (inside or outside IoU) and the contribution of each pixel. The PIoU loss can be easily calculated by accumulating the contribution of interior overlapping pixels. To demonstrate its effectiveness, the PIoU loss is evaluated on both anchor-based and anchor-free frameworks in the experiments.

To overcome the limitations of existing OBB-based approaches, we encourage the community to adopt more robust OBB detectors in a shift from conventional aerial imagery to more complex domains. We collected a new benchmark dataset, *Retail50K*, to reflect the challenges of detecting oriented targets with high aspect

ratios, heavy occlusions, and complex backgrounds. Experiments show that the proposed frameworks with PIoU loss not only have promising performances on aerial images, but they can also effectively handle new challenges in Retail50K.

The contributions of this work are summarized as follows: (1) We propose a novel loss function, PIoU loss, to improve the performance of oriented object detection in highly challenging conditions such as high aspect ratios and complex backgrounds. (2) We introduce a new dataset, Retail50K, to spur the computer vision community towards innovating and adapting existing OBB detectors to cope with more complex environments. (3) Our experiments demonstrate that the proposed PIoU loss can effectively improve the performances for both anchor-based and anchor-free OBB detectors in different datasets.

2 Related Work

2.1 Oriented Object Detectors

Existing oriented object detectors are mostly extended from generic horizontal bounding box detectors by introducing an additional angle dimension. For instance, [24] presented a rotation-invariant detector based on one-stage SSD [25]. [18] introduced a rotated detector based on two-stage Faster RCNN [34]. [6] designed an RoI transformer to learn the transformation from BB to OBB and thereafter, the rotation-invariant features are extracted. [12] formulated a generative probabilistic model to extract OBB proposals. For each proposal, the location, size and orientation are determined by searching the local maximum likelihood. Other possible ways of extracting OBB include, fitting detected masks [3, 10] and regressing OBB with anchor-free models [49], two new concepts in literature. While these approaches have promising performance on aerial images, they are not well-suited for oriented objects with high aspect ratios and complex environments. For this reason, we hypothesize that a new kind of loss is necessary to obtain improvements under challenging conditions. For the purpose of comparative evaluation, we implement both anchor-based and anchor-free frameworks as baselines in our experiments. We later show how these models, when equipped with PIoU Loss, can yield better results in both retail and aerial data.

2.2 Regression Losses

For bounding box regression, actively used loss functions are Mean Square Error [29] (MSE, L2 loss, the sum of squared distances between target and predicted variables), Mean Absolute Error [38] (MAE, L1 loss, the sum of absolute differences between target and predicted variables), Quantile Loss [2] (an extension of MAE, predicting an interval instead of only point predictions), Huber Loss [13] (basically absolute error, which becomes quadratic when error is small) and Log-Cosh Loss (the logarithm of the hyperbolic cosine of the prediction error) [30]. In practise, losses in common used detectors [32, 25, 34] are extended from the base functions above. However, we can not directly use them since there is an additional angle dimension involved in the OBB descriptor.

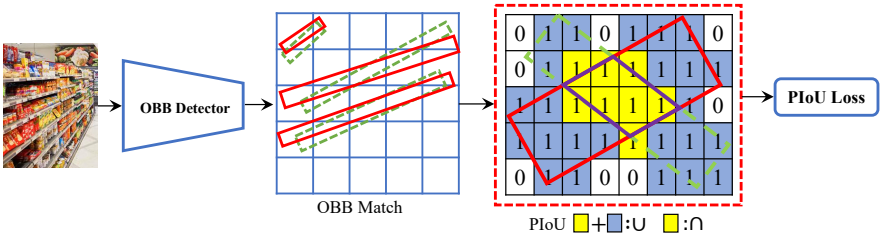
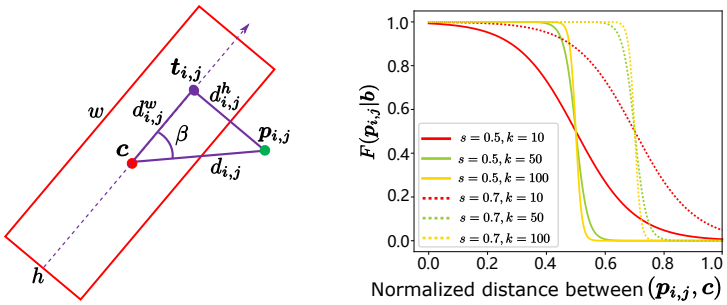


Fig. 2. Our proposed PIoU is a general concept that is applicable to most OBB-based frameworks. All possible predicted (green) and g/t (red) OBB pairs are matched to compute their PIoU. Building on that, the final PIoU loss is calculated using Eq. 14.

Besides the base functions, there have been several works that introduce IoU losses for horizontal bounding box. For instance, [45] propose an IoU loss which regresses the four bounds of a predicted box as a whole unit. [35] extends the idea of [45] by introducing a Generalized Intersection over Union loss (GIoU loss) for bounding box regression. The main purpose of GIoU is to get rid of the case that two polygons do not have an intersection. [37] introduce a novel bounding box regression loss based on a set of IoU upper bounds. However, when using oriented bounding box, those approaches become much more complicated thus are hard to implement, while the proposed PIoU loss is much simpler and suitable for both horizontal and oriented box. It should be noted that the proposed PIoU loss is different from [48] in which the IoU is computed based on axis alignment and polygon intersection, our method is more straightforward, i.e. IoU is calculated directly by accumulating the contribution of interior overlapping pixels. Moreover, the proposed PIoU loss is also different from Mask Loss in Mask RCNN [10]. Mask loss is calculated by the average binary cross-entropy with per-pixel sigmoid (also called Sigmoid Cross-Entropy Loss). Different from it, our proposed loss is calculated based on positive IoU to preserve intersection and union areas between two boxes. In each area, the contribution of pixels are modeled and accumulated depending on their spatial information. Thus, PIoU loss is more general and sensitive to OBB overlaps.

3 Pixels-IoU (PIoU) Loss

In this section, we present in detail the PIoU Loss. For a given OBB $\boldsymbol{b}$ encoded by (c_x, c_y, w, h, θ), an ideal loss function should effectively guide the network to maximize the IoU and thereby the error of $\boldsymbol{b}$ can be minimized. Towards this goal, we first explain the IoU method. Generally speaking, an IoU function should accurately compute the area of an OBB as well as its intersection with another box. Since OBB and the intersection area are constructed by pixels in image space, their areas are approximated by the number of interior pixels. Specifically, as shown in Figure 3(a), $\boldsymbol{t}_{i,j}$ (the purple point) is the intersection point between the mid-vertical line and its perpendicular line to pixel $\boldsymbol{p}_{i,j}$ (the green point). As a result, a triangle is constructed by OBB center $\boldsymbol{c}$ (the red



(a) (b)

Fig. 3. General idea of the IoU function. (a) Components involved in determining the relative position (inside or outside) between a pixel $\boldsymbol{p}$ (green point) and an OBB $\boldsymbol{b}$ (red rectangle). Best viewed in color. (b) Distribution of the kernelized pixel contribution $F(\boldsymbol{p}_{i,j}|\boldsymbol{b})$ with different distances between $\boldsymbol{p}_{i,j}$ and box center $\boldsymbol{c}$. We see that $F(\boldsymbol{p}_{i,j}|\boldsymbol{b})$ is continuous and differentiable due to Eq. 9. Moreover, it approximately reflects the value distribution in Eq. 1 when the pixels $\boldsymbol{p}_{i,j}$ are inside and outside $\boldsymbol{b}$.

point), $\boldsymbol{p}_{i,j}$ and $\boldsymbol{t}_{i,j}$. The length of each triangle side is denoted by $d^w_{i,j}$, $d^h_{i,j}$ and $d_{i,j}$. To judge the relative location (inside or outside) between $\boldsymbol{p}_{i,j}$ and $\boldsymbol{b}$, we define the binary constraints as follows:

$$\delta(\boldsymbol{p}_{i,j}|\boldsymbol{b}) = \begin{cases} 1, & d^w_{i,j} \leq \frac{w}{2}, d^h_{i,j} \leq \frac{h}{2} \\ 0, & otherwise \end{cases} \tag{1}$$

where d_{ij} denotes the L2-norm distance between pixel (i, j) and OBB center (c_x, c_y), d_w and d_h denotes the distance d along horizontal and vertical direction respectively:

$$d_{ij} = d(i, j) = \sqrt{(c_x - i)^2 + (c_y - j)^2} \tag{2}$$

$$d^w_{ij} = |d_{ij} \cos\beta| \tag{3}$$

$$d^h_{ij} = |d_{ij} \sin\beta| \tag{4}$$

$$\beta = \begin{cases} \theta + \arccos\dfrac{c_x - i}{d_{ij}}, & c_y - j \geq 0 \\ \theta - \arccos\dfrac{c_x - i}{d_{ij}}, & c_y - j < 0 \end{cases} \tag{5}$$

Let $B_{\boldsymbol{b},\boldsymbol{b}'}$ denotes the smallest horizontal bounding box that covers both $\boldsymbol{b}$ and $\boldsymbol{b}'$. We can then compute the intersection area $S_{\boldsymbol{b}\cap\boldsymbol{b}'}$ and union area $S_{\boldsymbol{b}\cup\boldsymbol{b}'}$ between two OBBs $\boldsymbol{b}$ and $\boldsymbol{b}'$ using the statistics of all pixels in $B_{\boldsymbol{b},\boldsymbol{b}'}$:

$$S_{\boldsymbol{b}\cap\boldsymbol{b}'} = \sum_{\boldsymbol{p}_{i,j}\in B_{\boldsymbol{b},\boldsymbol{b}'}} \delta(\boldsymbol{p}_{i,j}|\boldsymbol{b})\delta(\boldsymbol{p}_{i,j}|\boldsymbol{b}') \tag{6}$$

$$S_{\boldsymbol{b}\cup\boldsymbol{b}'} = \sum_{\boldsymbol{p}_{i,j}\in B_{\boldsymbol{b},\boldsymbol{b}'}} \delta(\boldsymbol{p}_{i,j}|\boldsymbol{b}) + \delta(\boldsymbol{p}_{i,j}|\boldsymbol{b}') - \delta(\boldsymbol{p}_{i,j}|\boldsymbol{b})\delta(\boldsymbol{p}_{i,j}|\boldsymbol{b}') \tag{7}$$

The final IoU of $\boldsymbol{b}$ and $\boldsymbol{b}'$ can be calculated by dividing $S_{\boldsymbol{b}\cap\boldsymbol{b}'}$ and $S_{\boldsymbol{b}\cup\boldsymbol{b}'}$. However, we observe that Eq. 1 is not a continuous and differentiable function. As a result, back propagation (BP) cannot utilize an IoU-based loss for training. To solve this problem, we approximate Eq. 1 as $F(\boldsymbol{p}_{i,j}|\boldsymbol{b})$ taking on the product of two kernels:

$$F(\boldsymbol{p}_{i,j}|\boldsymbol{b}) = K(d^w_{i,j}, w)K(d^h_{i,j}, h) \tag{8}$$

Particularly, the kernel function $K(d, s)$ is calculated by:

$$K(d, s) = 1 - \frac{1}{1 + e^{-k(d-s)}} \tag{9}$$

where k is an adjustable factor to control the sensitivity of the target pixel $\boldsymbol{p}_{i,j}$. The key idea of Eq. 8 is to obtain the contribution of pixel $\boldsymbol{p}_{i,j}$ using the kernel function in Eq. 9. Since the employed kernel is calculated by the relative position (distance and angle of the triangle in Figure 3(a)) between $\boldsymbol{p}_{i,j}$ and $\boldsymbol{b}$, the intersection area $S_{\boldsymbol{b}\cap\boldsymbol{b}'}$ and union area $S_{\boldsymbol{b}\cup\boldsymbol{b}'}$ are inherently sensitive to both OBB rotation and size. In Figure 3(b), we find that $F(\boldsymbol{p}_{i,j}|\boldsymbol{b})$ is continuous and differentiable. More importantly, it functions similarly to the characteristics of Eq. 1 such that $F(\boldsymbol{p}_{i,j}|\boldsymbol{b})$ is close to 1.0 when the pixel $\boldsymbol{p}_{i,j}$ is inside and otherwise when $F(\boldsymbol{p}_{i,j}|\boldsymbol{b}) \sim 0$. Following Eq. 8, the intersection area $S_{\boldsymbol{b}\cap\boldsymbol{b}'}$ and union area $S_{\boldsymbol{b}\cup\boldsymbol{b}'}$ between $\boldsymbol{b}$ and $\boldsymbol{b}'$ are approximated by:

$$S_{\boldsymbol{b}\cap\boldsymbol{b}'} \approx \sum_{\boldsymbol{p}_{i,j}\in B_{\boldsymbol{b},\boldsymbol{b}'}} F(\boldsymbol{p}_{i,j}|\boldsymbol{b})F(\boldsymbol{p}_{i,j}|\boldsymbol{b}') \tag{10}$$

$$S_{\boldsymbol{b}\cup\boldsymbol{b}'} \approx \sum_{\boldsymbol{p}_{i,j}\in B_{\boldsymbol{b},\boldsymbol{b}'}} F(\boldsymbol{p}_{i,j}|\boldsymbol{b}) + F(\boldsymbol{p}_{i,j}|\boldsymbol{b}') - F(\boldsymbol{p}_{i,j}|\boldsymbol{b})F(\boldsymbol{p}_{i,j}|\boldsymbol{b}') \tag{11}$$

In practice, to reduce the computational complexity of Eq. 11, $S_{\boldsymbol{b}\cup\boldsymbol{b}'}$ can be approximated by a simpler form:

$$S_{\boldsymbol{b}\cup\boldsymbol{b}'} = w \times h + w' \times h' - S_{\boldsymbol{b}\cap\boldsymbol{b}'} \tag{12}$$

where (w, h) and (w', h') are the size of OBBs $\boldsymbol{b}$ and $\boldsymbol{b}'$, respectively. Our experiment in Section 5.2 shows that Eq. 12 can effectively reduce the complexity of Eq. 10 while preserving the overall detection performance. With these terms, our proposed Pixels-IoU ($PIoU$) is computed as:

$$PIoU(\boldsymbol{b}, \boldsymbol{b}') = \frac{S_{\boldsymbol{b}\cap\boldsymbol{b}'}}{S_{\boldsymbol{b}\cup\boldsymbol{b}'}} \tag{13}$$

Let $\boldsymbol{b}$ denotes the predicted box and $\boldsymbol{b}'$ denotes the ground-truth box. A pair $(\boldsymbol{b}, \boldsymbol{b}')$ is regarded as positive if the predicted box $\boldsymbol{b}$ is based on a positive anchor and $\boldsymbol{b}'$ is the matched ground-truth box (an anchor is matched with a ground-truth box if the IoU between them is larger them 0.5). We use M to denote the

Table 1. Comparison between different datasets with OBB annotations. ≈ indicate estimates based on selected annotated samples as full access was not possible.

Dataset	Scenario	Median Ratio	Images	Instances
SZTAKI [1]	Aerial	≈1:3	9	665
VEDAI [31]	Aerial	1:3	1268	2950
UCAS-AOD [50]	Aerial	1:1.3	1510	14596
HRSC2016 [26]	Aerial	1:5	1061	2976
Vehicle [23]	Aerial	1:2	20	14235
DOTA [39]	Aerial	1:2.5	2806	188282
SHIP [18]	Aerial	≈1:5	640	-
OOP [12]	PASCAL	≈1:1	4952	-
Proposed	**Retail**	**1:20**	**47000**	**48000**

set of all positive pairs. With the goal to maximize the PIoU between $\boldsymbol{b}$ and $\boldsymbol{b}'$, the proposed PIoU Loss is calculated by:

$$L_{piou} = \frac{-\sum_{(\boldsymbol{b},\boldsymbol{b}')\in M} \ln PIoU(\boldsymbol{b},\boldsymbol{b}')}{|M|} \tag{14}$$

Theoretically, Eq. 14 still works if there is no intersection between $\boldsymbol{b}$ and $\boldsymbol{b}'$. This is because $PIoU(\boldsymbol{b},\boldsymbol{b}') > 0$ based on Eq. 9 and the gradients still exist in this case. Moreover, the proposed PIoU also works for horizontal bounding box regression. Specifically, we can simply set $\theta = 0$ in Eq. 5 for this purpose. In Section 5, we experimentally validate the usability of PIoU for horizontal bounding box regression.

4 Retail50K Dataset

OBB detectors have been actively studied for many years and several datasets with such annotations have been proposed [39, 1, 23, 26, 31, 50, 18, 12]. As shown in Table 1, most of them only focused on aerial images (Figure 4 (a),(b)) while a few are annotated based on existing datasets such as MSCOCO [22], PASCAL VOC [7] and ImageNet [5]. These datasets are important to evaluate the detection performance with simple backgrounds and low aspect ratios. For example, aerial images are typically gray and texture-less. The statistics in [39] shows that most datasets of aerial images have a wide range of aspect ratios, but around 90% of these ratios are distributed between 1:1 and 1:4, and very few images contain OBBs with aspect ratios larger than 1:5. Moreover, aspect ratios of OBBs on PASCAL VOC are mostly close to square (1:1). As a result, it is hard to assess the capability of detectors on objects with high aspect ratios and complex backgrounds using existing datasets. Motivated by this, we introduce a new dataset, namely Retail50K, to advance the research of detection of rotated objects in complex environments. We intend to make this publicly available to the community (`https://github.com/clobotics/piou`).

Figure 4 (c) illustrates a sample image from Retail50K dataset. Retail50K is a collection of 47,000 images from different supermarkets. Annotations on those

(a) DOTA (b) HRSC (c) Retail50K

Fig. 4. Sample images and their annotations of three datasets evaluated in our experiments: (a) DOTA [39] (b) HRSC2016 [26] (c) Retail50K. There are two unique characteristics of Retail50K: (1) Complex backgrounds such as occlusions (by price tags), varied colours and textures. (2) OBB with high aspect ratios.

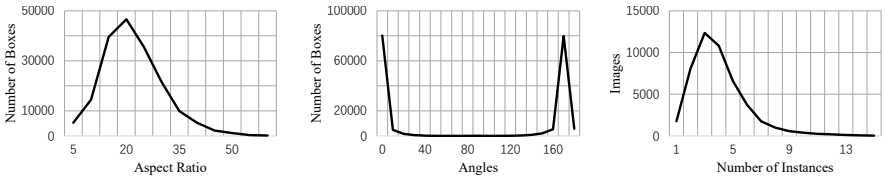


Fig. 5. Statistics of different properties of Retail50K dataset.

images are the layer edges of shelves, fridges and displays. We focus on such retail environments for three reasons: (1) **Complex background.** Shelves and fridges are tightly filled with many different items with a wide variety of colours and textures. Moreover, layer edges are normally occluded by price tags and sale tags. Based on our statistics, the mean occlusion is around 37.5%. It is even more challenging that the appearance of price tags are different in different supermarkets. (2) **High aspect ratio.** Aspect ratio is one of the essential factors for anchor-based models [33]. Bounding boxes in Retail50K dataset not only have large variety in degrees of orientation, but also a wide range of aspect ratios. In particular, the majority of annotations in Retail50K are with high aspect ratios. Therefore, this dataset represents a good combination of challenges that is precisely the type we find in complex retail environments.(3) **Useful in practice.** The trained model based on Retail50K can be used for many applications in retail scenarios such as shelf retail tag detection, automatic shelf demarcation, shelf layer and image yaw angle estimation, etc. It is worth to note that although SKU-110K dataset [9] is also assembled from retail environment such as supermarket shelves, the annotations in this dataset are horizontal bounding boxes (HBB) of shelf products since it mainly focuses on object detection in densely packed scenes. The aspect ratios of its HBB are distributed between 1:1-1:3 and hence, it does not cater to the problem that we want to solve.

Images and Categories: Images in Retail50K were collected from 20 supermarket stores in China and USA. Dozens of volunteers acquired data using their personal cellphone cameras. To increase the diversity of data, images were collected in multiple cities from different volunteers. Image quality and view settings were unregulated and so the collected images represent different scales, viewing angles, lighting conditions, noise levels, and other sources of variability. We also

recorded the meta data of the original images such as capture time, volunteer name, shop name and MD5 [40] checksum to filter out duplicated images. Unlike existing datasets that contain multiple categories [39, 22, 7, 5], there is only one category in Retail50K dataset. For better comparisons across datasets, we also employ DOTA [39] (15 categories) and HRSC2016 [26] (the aspect ratio of objects is between that of Retail50K and DOTA) in our experiments (Figure 4).
Annotation and Properties: In Retail50K dataset, bounding box annotations were provided by 5 skilled annotators. To improve their efficiency, a handbook of labelling rules was provided during the training process. Candidate images were grouped into 165 labelling tasks based on their meta-data so that peer reviews can be applied. Finally, considering the complicated background and various orientations of layer edges, we perform the annotations using arbitrary quadrilateral bounding boxes (AQBB). Briefly, AQBB is denoted by the vertices of the bounding polygon in clockwise order. Due to high efficiency and empirical success, AQBB is widely used in many benchmarks such as text detection [15], object detection in aerial images [18], etc. Based on AQBB, we can easily compute the required OBB format which is denoted by (c_x, c_y, w, h, θ).

Since images were collected with personal cellphone cameras, the original images have different resolutions; hence they were uniformly resized into 600×800 before annotation took place. Figure 5 shows some statistics of Retail50K. We see that the dataset contains a wide range of aspect ratios and orientations (Figure 5 (a) and (b)). In particular, Retail50K is more challenging as compared to existing datasets [23, 39, 18] since it contains rich annotations with extremely high aspect ratios (higher than 1:10). Similar to natural-image datasets such as ImageNet (average 2) and MSCOCO (average 7.7), most images in our dataset contain around 2-6 instances with complex backgrounds (Figure 5 (c)). For experiments, we selected half of the original images as the training set, 1/6 as validation set, and 1/3 as the testing set.

5 Experiments

5.1 Experimental Settings

We evaluate the proposed PIoU loss with anchor-based and anchor-free OBB-detectors (RefineDet, CenterNet) under different parameters, backbones. We also compare the proposed method with other state-of-the-art OBB-detection methods in different benchmark datasets (*i.e.* DOTA [39], HRSC2016 [26], PASCAL VOC [7]) and the proposed Retail50K dataset. The training and testing tasks are accomplished on a desktop machine with Intel(R) Core(TM) i7-6850K CPU @ 3.60GHzs, 64 GB installed memory, a GeForce GTX 1080TI GPU (11 GB global memory), and Ubuntu 16.04 LTS. With this machine, the batch size is set to 8 and 1 for training and testing, respectively.
Anchor-based OBB Detector: For anchor-based object detection, we train RefineDet [46] by updating its loss using the proposed PIoU method. Since the detector is optimized by classification and regression losses, we can easily replace

the regression one with PIoU loss L_{piou} while keeping the original Softmax Loss L_{cls} for classification. We use ResNet [11] and VGG [36] as the backbone models. The oriented anchors are generated by rotating the horizontal anchors by $k\pi/6$ for $0 \leq k < 6$. We adopt the data augmentation strategies introduced in [25] except cropping, while including rotation (i.e. rotate the image by a random angle sampled in $[0, \pi/6]$). In training phase, the input image is resized to 512×512. We adopt the mini-batch training on 2 GPUs with 8 images per GPU. SGD is adopted to optimize the models with momentum set to 0.9 and weight decay set to 0.0005. All evaluated models are trained for 120 epochs with an initial learning rate of 0.001 which is then divided by 10 at 60 epochs and again at 90 epochs. Other experimental settings are the same as those in [46].

Anchor-free OBB Detector: To extend anchor-free frameworks for detecting OBB, we modify CenterNet [49] by adding an angle dimension regressed by L1-Loss in its overall training objective as our baseline. To evaluate the proposed loss function, in similar fashion as anchor-based approach, we can replace the regression one with PIoU loss L_{piou} while keeping the other classification loss L_{cls} the same. Be noted that CenterNet uses a heatmap to locate the center of objects. Thus, we do not back-propagate the gradient of the object's center when computing the PIoU loss. We use DLA [44] and ResNet [11] as the backbone models. The data augmentation strategies is the same as those for RefineDet-OBB (shown before). In training phase, the input image is resized to 512×512. We adopt the mini-batch training on 2 GPUs with 16 images per GPU. ADAM is adopted to optimize the models. All evaluated models are trained for 120 epochs with an initial learning rate of 0.0005 which is then divided by 10 at 60 epochs and again at 90 epochs. Other settings are the same as those in [49].

5.2 Ablation Study

Comparison on different parameters: In Eq. 9, k is an adjustable factor in our kernel function to control the sensitivity of each pixel. In order to evaluate its influence as well as to find a proper value for the remaining experiments, we conduct a set of experiments by varying k values based on DOTA [39] dataset with the proposed anchor-based framework. To simplify discussions, results of $k = 5, 10, 15$ are detailed in Table 2 while their distributions can be visualized in Fig. 3(b). We finally select $k = 10$ for the rest of the experiments since it achieves the best accuracy.

Comparison for oriented bounding box: Based on DOTA [39] dataset, we compare the proposed PIoU loss with the commonly used L1 loss, SmoothL1 loss as well as L2 loss. For fair comparisons, we fix the backbone to VGGNet [36] and build the network based on FPN [20]. Table 3 details the comparisons and we can clearly see that the proposed PIoU Loss improves the detection performance by around 3.5%. HPIoU (Hard PIoU) loss is the simplified PIoU loss using Eq. 12. Its performance is slightly reduced but still comparable to PIoU loss. Thus, HPIoU loss can be a viable option in practise as it has lower computational complexity. We also observe that the proposed PIoU costs 15-20% more time than other three loss functions, which shows that it is still acceptable in

Table 2. Comparison between different sensitivity factor k in Eq. 9 for PIoU loss on DOTA dataset. RefineDet [46] is used as the detection model.

k	AP	AP_{50}	AP_{75}
5	46.88	59.03	34.73
10	54.24	67.89	40.59
15	53.41	65.97	40.84

Table 3. Comparison between different losses for oriented bounding box on DOTA dataset. RefineDet [46] is used as the detection model. HPIoU (Hard PIoU) loss refers to the PIoU loss simplified by Eq. 12. Training time is estimated in hours.

Loss	AP	AP_{50}	AP_{75}	Training Time
L1 Loss	50.66	64.14	37.18	20
L2 Loss	49.70	62.74	36.65	20
SmoothL1 Loss	51.46	65.68	37.25	21.5
PIoU Loss	**54.24**	**67.89**	**40.59**	**25.7**
HPIoU Loss	**53.37**	**66.38**	**40.36**	**24.8**

Table 4. Comparison between different losses for horizontal bounding box on PASCAL VOC2007 dataset. SSD [25] is used as the detection model.

Loss	AP	AP_{50}	AP_{60}	AP_{70}	AP_{80}	AP_{90}
SmoothL1 Loss	48.8	79.8	72.9	60.6	40.3	10.2
GIoU Loss [35]	49.9	79.8	74.1	63.2	41.9	12.4
PIoU Loss	**50.3**	**80.1**	**74.9**	**63.0**	**42.5**	**12.2**

practice. We also observed that HPIoU costs less training time than PIoU. Such observation verifies the theoretical analysis and usability of Eq. 12.

Comparison for horizontal bounding box: Besides, we also compare the PIoU loss with SmoothL1 loss and GIoU loss [35] for horizontal bounding box on PASCAL VOC dataset [7]. In Table 4, we observe that the proposed PIoU loss is still better than SmoothL1 loss and GIoU loss for horizontal bounding box regression, particularly at those AP metrics with high IoU threshold. Note that the GIoU loss is designed only for horizontal bounding box while the proposed PIoU loss is more robust and well suited for both horizontal and oriented bounding box. Together with the results in Table 3, we observe the strong generalization ability and effectiveness of the proposed PIoU loss.

5.3 Benchmark Results

Retail50K: We evaluate our PIoU loss with two OBB-detectors (*i.e.* the OBB versions of RefineDet [46] and CenterNet [49]) on Retail50K dataset. The experimental results are shown in Table 5. We observe that, both detectors achieve significant improvements with the proposed PIoU loss ($\sim$ 7% improvement for RefineDet-OBB and $\sim$ 6% improvement for CenterNet-OBB). One reason for obtaining such notable improvements is that the proposed PIoU loss is much better suited for oriented objects than the traditional regression loss. Moreover, the improvements from PIoU loss in Retail50K are more obvious than those

Table 5. Detection results on Retail50K dataset. The PIoU loss is evaluated on RefineDet [46] and CenterNet [49] with different backbone models.

Method	Backbone	AP	AP_{50}	AP_{75}	Time (ms)	FPS
RefineDet-OBB [46]	ResNet-50	53.96	74.15	33.77	142	7
RefineDet-OBB+PIoU	ResNet-50	**61.78**	**80.17**	**43.39**	**142**	**7**
RefineDet-OBB [46]	ResNet-101	55.46	77.05	33.87	167	6
RefineDet-OBB+PIoU	ResNet-101	**63.00**	**79.08**	**46.01**	**167**	**6**
CenterNet-OBB [49]	ResNet18	54.44	76.58	32.29	7	140
CenterNet-OBB+PIoU	ResNet18	**61.02**	**87.19**	**34.85**	**7**	**140**
CenterNet-OBB [49]	DLA-34	56.13	78.29	33.97	18.18	55
CenterNet-OBB+PIoU	DLA-34	**61.64**	**88.47**	**34.80**	**18.18**	**55**

Table 6. Detection results on HRSC2016 dataset. *Aug.* indicates data augmentation. *Size* means the image size that used for training and testing.

Method	Backbone	Size	Aug.	mAP	FPS
R^2CNN [14]	ResNet101	800 × 800	×	73.03	2
RC1 & RC2 [27]	VGG-16	-	-	75.7	<1fps
RRPN [28]	ResNet101	800 × 800	×	79.08	3.5
R^2PN [47]	VGG-16	-	√	79.6	<1fps
RetinaNet-H [41]	ResNet101	800 × 800	√	82.89	14
RetinaNet-R [41]	ResNet101	800 × 800	√	89.18	10
RoI-Transformer [6]	ResNet101	512 × 800	×	86.20	-
R^3Det [41]	ResNet101	300 × 300	√	87.14	18
	ResNet101	600 × 600	√	88.97	15
	ResNet101	800 × 800	√	89.26	12
CenterNet-OBB [49]	ResNet18	512 × 512	√	67.73	140
CenterNet-OBB+PIoU	**ResNet18**	**512 × 512**	√	**78.54**	**140**
CenterNet-OBB [49]	ResNet101	512 × 512	√	77.43	45
CenterNet-OBB+PIoU	**ResNet101**	**512 × 512**	√	**80.32**	**45**
CenterNet-OBB [49]	DLA-34	512 × 512	√	87.98	55
CenterNet-OBB+PIoU	**DLA-34**	**512 × 512**	√	**89.20**	**55**

in DOTA (*c.f.* Table 3), which could mean that the proposed PIoU loss is extremely useful for objects with high aspect ratios and complex environments. This verifies the effectiveness of the proposed method.

HRSC2016: The HRSC2016 dataset [26] contains 1070 images from two scenarios including ships on sea and ships close inshore. We evaluate the proposed PIoU with CenterNet [49] on different backbones, and compare them with several state-of-the-art detectors. The experimental results are shown in Table 6. It can be seen that the CenterNet-OBB+PIoU outperforms all other methods except R^3Det-800. This is because we use a smaller image size (512×512) than R^3Det-800 (800×800). Thus, our detector preserves a reasonably competitive detection performance, but with far better efficiency (55 fps *v.s* 12 fps). This exemplifies the strength of the proposed PIoU loss on OBB detectors.

DOTA: The DOTA dataset [39] contains 2806 aerial images from different sensors and platforms with crowd-sourcing. Each image is of size about 4000×4000

Table 7. Detection results on DOTA dataset. We report the detection results for each category to better demonstrate where the performance gains come from.

Method	Backbone	Size	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
SSD [25]	VGG16	512	39.8	9.1	0.6	13.2	0.3	0.4	1.1	16.2	27.6	9.2	27.2	9.1	3.0	1.1	1.0	10.6
YOLOV2 [33]	DarkNet19	416	39.6	20.3	36.6	23.4	8.9	2.1	4.8	44.3	38.4	34.7	16.0	37.6	47.2	25.5	7.5	21.4
R-FCN [4]	ResNet101	800	37.8	38.2	3.6	37.3	6.7	2.6	5.6	22.9	46.9	66.0	33.4	47.2	10.6	25.2	18.0	26.8
FR-H [34]	ResNet101	800	47.2	61.0	9.8	51.7	14.9	12.8	6.9	56.3	60.0	57.3	47.8	48.7	8.2	37.3	23.1	32.3
FR-O [39]	ResNet101	800	79.1	69.1	17.2	63.5	34.2	37.2	36.2	89.2	69.6	59.0	49.	52.5	46.7	44.8	46.3	52.9
R-DFPN [42]	ResNet101	800	80.9	65.8	33.8	58.9	55.8	50.9	54.8	90.3	66.3	68.7	48.7	51.8	55.1	51.3	35.9	57.9
R^2CNN [14]	ResNet101	800	80.9	65.7	35.3	67.4	59.9	50.9	55.8	90.7	66.9	72.4	55.1	52.2	55.1	53.4	48.2	60.7
RRPN [28]	ResNet101	800	88.5	71.2	31.7	59.3	51.9	56.2	57.3	90.8	72.8	67.4	56.7	52.8	53.1	51.9	53.6	61.0
RefineDet [46]	VGG16	512	80.5	26.3	33.2	28.5	63.5	75.1	78.8	90.8	61.1	65.9	12.1	23.0	50.9	50.9	22.6	50.9
RefineDet+PIoU	VGG16	512	**80.5**	**33.3**	**34.9**	**28.1**	**64.9**	**74.3**	**78.7**	**90.9**	**65.8**	**66.6**	**19.5**	**24.6**	**51.1**	**50.8**	**23.6**	**52.5**
RefineDet [46]	ResNet101	512	80.7	44.2	27.5	32.8	61.2	76.1	78.8	90.7	69.9	73.9	24.9	31.9	55.8	51.4	26.8	55.1
RefineDet+PIoU	ResNet101	512	**80.7**	**48.8**	**26.1**	**38.7**	**65.2**	**75.5**	**78.6**	**90.8**	**70.4**	**75.0**	**32.0**	**28.0**	**54.3**	**53.7**	**29.6**	**56.5**
CenterNet [49]	DLA-34	512	81.0	64.0	22.6	56.6	38.6	64.0	64.9	90.8	78.0	72.5	44.0	41.1	55.5	55.0	57.4	59.1
CenterNet+PIoU	DLA-34	512	**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**

pixels and contains objects of different scales, orientations and shapes. Note that image in DOTA is too large to be directly sent to CNN-based detectors. Thus, similar to the strategy in [39], we crop a series of 512×512 patches from the original image with the stride set to 256. For testing, the detection results are obtained from the DOTA evaluation server. The detailed performances for each category are reported so that deeper observations could be made. We use the same short names, benchmarks and forms as those existing methods in [41] to evaluate the effectiveness of PIoU loss on this dataset. The final results are shown in Table 7. We find that the performance improvements vary among different categories. However, it is interesting to find that the improvement is more plausible for some categories with high aspect ratios. For example, harbour (HA), ground track field (GTF), soccer-ball field (SBF) and basketball court (BC) all naturally have large aspect ratios, and they appear to benefit from the inclusion of PIoU. Such observations confirm that the PIoU can effectively improve the performance of OBB detectors, particularly on objects with high-aspect ratios. These verify again the effectiveness of the proposed PIoU loss on OBB detectors. We also find that our baselines are relatively low than some state-of-the-art performances. We conjecture the main reason is that we use much smaller input size than other methods (512 vs 1024 on DOTA). However, note that the existing result (89.2 mAP) for HRSC2016 in Table 6 already achieves the state-of-the-art level performance with only 512×512 image size. Thus, the proposed loss function can bring gain in this strong baseline.

In order to visually verify these performance improvements, we employ the anchor-based model RefineDet [46] and conduct two independent experiments using PIoU and SmoothL1 losses. The experiments are applied on all three datasets (*i.e.* Retail50K, DOTA [39], HRSC2016 [26]) and selected visual results are presented in Figure 6. We can observe that the OBB detector with PIoU loss (in red boxes) has more robust and accurate detection results than the one with SmoothL1 loss (in yellow boxes) on all three datasets, particularly on Retail50K, which demonstrates its strength in improving the performance for high aspect ratio oriented objects. Here, we also evaluate the proposed HPIoU loss with the same configuration of PIoU. In our experiments, the performances of HPIoU loss are slightly lower than those of PIoU loss (0.87, 1.41 and 0.18 mAP on DOTA, Retail50K and HRSC2016 respectively), but still better than

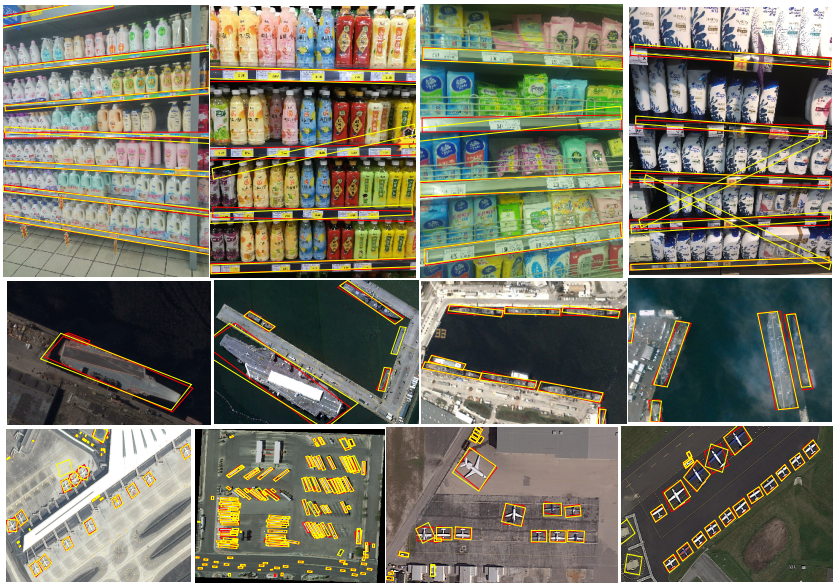


Fig. 6. Samples results using PIoU (red boxes) and SmoothL1 (yellow boxes) losses on Retail50K (first row), HRSC2016 (second row) and DOTA (last row) datasets.

smooth-L1 loss while having higher training speed than PIoU loss. Overall, the performances of HPIoU are consistent on all three datasets.

6 Conclusion

We introduce a simple but effective loss function, PIoU, to exploit both the angle and IoU for accurate OBB regression. The PIoU loss is derived from IoU metric with a pixel-wise form, which is simple and suitable for both horizontal and oriented bounding box. To demonstrate its effectiveness, we evaluate the PIoU loss on both anchor-based and anchor-free frameworks. The experimental results show that PIoU loss can significantly improve the accuracy of OBB detectors, particularly on objects with high-aspect ratios. We also introduce a new challenging dataset, Retail50K, to explore the limitations of existing OBB detectors as well as to validate their performance after using the PIoU loss. In the future, we will extend PIoU to 3D rotated object detection. Our preliminary results show that PIoU can improve PointPillars [16] on KITTI val dataset [8] by 0.65, 0.64 and 2.0 AP for car, pedestrian and cyclist in moderate level, respectively.

Acknowledgements

The paper is supported in part by the following grants: China Major Project for New Generation of AI Grant (No.2018AAA0100400), National Natural Science Foundation of China (No. 61971277). The work is also supported by funding from Clobotics under the Joint Research Program of Smart Retail.

References

1. Benedek, C., Descombes, X., Zerubia, J.: Building development monitoring in multitemporal remotely sensed image pairs with stochastic birth-death dynamics. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(1), 33–50 (2012)
2. Cannon, A.J.: Quantile regression neural networks: implementation in r and application to precipitation downscaling. Computers & Geosciences **37**, 1277–1284 (2011)
3. Chen, B., Tsotsos, J.K.: Fast visual object tracking with rotated bounding boxes. In: IEEE International Conference on Computer Vision. pp. 1–9 (2019)
4. Dai, J., Li, Y., He, K., Sun, J.: R-fcn: Object detection via region-based fully convolutional networks. In: Advances in Neural Information Processing Systems. pp. 379–387 (2016)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
6. Ding, J., Xue, N., Long, Y., Xia, G.S., Lu, Q.: Learning roi transformer for detecting oriented objects in aerial images. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9 (2019)
7. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision **111**(1), 98–136 (2015)
8. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013)
9. Goldman, E., Herzig, R., Eisenschtat, A., Goldberger, J., Hassner, T.: Precise detection in densely packed scenes. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9 (2019)
10. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: IEEE International Conference on Computer Vision. pp. 2980–2988 (2017)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
12. He, S., Lau, R.W.: Oriented object proposals. In: IEEE International Conference on Computer Vision. pp. 280–288 (2015)
13. Huber, P.J.: Robust estimation of a location parameter. Annals of Statistics **53**, 73101 (1964)
14. Jiang, Y., Zhu, X., Wang, X., Yang, S., Li, W., Wang, H., Fu, P., Luo, Z.: R2cnn: rotational region cnn for orientation robust scene text detection. arXiv:1706.09579 (2017)
15. Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V.R., Lu, S., et al.: Icdar competition on robust reading. In: International Conference on Document Analysis and Recognition. pp. 1156–1160 (2015)
16. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 12697–12705 (2019)
17. Law, H., Deng, J.: Cornernet: Detecting objects as paired keypoints. In: European Conference on Computer Vision. pp. 734–750 (2018)

18. Li, S., Zhang, Z., Li, B., Li, C.: Multiscale rotated bounding box-based deep learning method for detecting ship targets in remote sensing images. Sensors **18**(8), 1–14 (2018)
19. Liao, M., Zhu, Z., Shi, B., Xia, G.s., Bai, X.: Rotation-sensitive regression for oriented scene text detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 5909–5918 (2018)
20. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2117–2125 (2017)
21. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: IEEE International Conference on Computer Vision. pp. 2980–2988 (2017)
22. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision. pp. 740–755 (2014)
23. Liu, K., Mattyus, G.: Fast multiclass vehicle detection on aerial images. IEEE Geoscience and Remote Sensing Letters **12**(9), 1938–1942 (2015)
24. Liu, L., Pan, Z., Lei, B.: Learning a rotation invariant detector with rotatable bounding box. arXiv:1711.09405 (2017)
25. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European Conference on Computer Vision. pp. 21–37 (2016)
26. Liu, Z., Wang, H., Weng, L., Yang, Y.: Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds. IEEE Geoscience and Remote Sensing Letters **13**(8), 1074–1078 (2016)
27. Liu, Z., Yuan, L., Weng, L., Yang, Y.: A high resolution optical satellite image dataset for ship recognition and some new baselines. In: International Conference on Pattern Recognition Applications and Methods. vol. 2, pp. 324–331 (2017)
28. Ma, J., Shao, W., Ye, H., Wang, L., Wang, H., Zheng, Y., Xue, X.: Arbitrary-oriented scene text detection via rotation proposals. IEEE Transactions on Multimedia **20**, 3111–3122 (2018)
29. Mood, A.M.: Introduction to the theory of statistics. pp. 229–229. McGraw-Hill (1974)
30. Muller, R.R., Gerstacker, W.H.: On the capacity loss due to separation of detection and decoding. IEEE Transactions on Information Theory **50**, 1769–1778 (2004)
31. Razakarivony, S., Jurie, F.: Vehicle detection in aerial imagery : A small target detection benchmark. Journal of Visual Communication and Image Representation **34**(1), 187–203 (2016)
32. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. pp. 779–788 (2016)
33. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. arXiv:1612.08242 (2016)
34. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems. pp. 91–99 (2015)
35. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 658–666 (2019)
36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)

37. Tychsen-Smith, L., Petersson, L.: Lars petersson: Improving object localization with fitness nms and bounded iou loss. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 6877–6887 (2018)
38. Willmott, C.J., Matsuura, K.: Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. Climate Research **30**, 7982 (2005)
39. Xia, G.S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L.: Dota: A large-scale dataset for object detection in aerial images. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3974–3983 (2018)
40. Xie, T., Liu, F., Feng, D.: Fast collision attack on md5. IACR Cryptology ePrint Archive **2013**, 170 (2013)
41. Yang, X., Liu, Q., Yan, J., Li, A.: R3det: Refined single-stage detector with feature refinement for rotating object. arXiv:1908.05612 (2019)
42. Yang, X., Sun, H., Fu, K., Yang, J., Sun, X., Yan, M., Guo, Z.: 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)
43. Yang, X., Yang, J., Yan, J., Zhang, Y., Zhang, T., Guo, Z., Sun, X., Fu, K.: Scrdet: Towards more robust detection for small, cluttered and rotated objects. In: IEEE International Conference on Computer Vision. pp. 1–9 (2019)
44. Yu, F., Wang, D., Shelhamer, E., Darrell, T.: Deep layer aggregation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2403–2412 (2018)
45. Yu, J., Jiang, Y., Wang, Z., Cao, Z., Huang, T.: Unitbox: An advanced object detection network. In: ACM International Conference on Multimedia. pp. 516–520 (2016)
46. Zhang, S., Wen, L., Bian, X., Lei, Z., Li, S.Z.: Single-shot refinement neural network for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4203–4212 (2018)
47. Zhang, Z., Guo, W., Zhu, S., Yu, W.: Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks. IEEE Geoscience and Remote Sensing Letters **15**(11), 1745–1749 (2018)
48. Zhou, D., Fang, J., Song, X., Guan, C., Yin, J., Dai, Y., Yang, R.: Iou loss for 2d/3d object detection. In: IEEE International Conference on 3D Vision. pp. 1–10 (2019)
49. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv:1904.07850 (2019)
50. Zhu, H., Chen, X., Dai, W., Fu, K., Ye, Q., Jiao, J.: Orientation robust object detection in aerial images using deep convolutional neural network. In: IEEE International Conference on Image Processing. pp. 3735–3739 (2015)