PointOBB: Learning Oriented Object Detection via Single Point Supervision

Junwei Luo[1], Xue Yang[2]*, Yi Yu[3], Qingyun Li[4], Junchi Yan[5,2], Yansheng Li[1]†

[1]Wuhan University   [2]Shanghai AI Laboratory
[3]Southeast University   [4]Harbin Institute of Technology   [5]Shanghai Jiao Tong University

luojunwei@whu.edu.cn   yangxue@pjlab.org.cn   yuyi@seu.edu.cn
21b905003@stu.hit.edu.cn   yanjunchi@sjtu.edu.cn   yansheng.li@whu.edu.cn
https://github.com/Luo-Z13/pointobb

Abstract

*Single point-supervised object detection is gaining attention due to its cost-effectiveness. However, existing approaches focus on generating horizontal bounding boxes (HBBs) while ignoring oriented bounding boxes (OBBs) commonly used for objects in aerial images. This paper proposes **PointOBB**, the first single **Point**-based **OBB** generation method, for oriented object detection. PointOBB operates through the collaborative utilization of three distinctive views: an original view, a resized view, and a rotated/flipped (rot/flp) view. Upon the original view, we leverage the resized and rot/flp views to build a scale augmentation module and an angle acquisition module, respectively. In the former module, a Scale-Sensitive Consistency (SSC) loss is designed to enhance the deep network's ability to perceive the object scale. For accurate object angle predictions, the latter module incorporates self-supervised learning to predict angles, which is associated with a scale-guided Dense-to-Sparse (DS) matching strategy for aggregating dense angles corresponding to sparse objects. The resized and rot/flp views are switched using a progressive multi-view switching strategy during training to achieve coupled optimization of scale and angle. Experimental results on the DIOR-R and DOTA-v1.0 datasets demonstrate that PointOBB achieves promising performance, and significantly outperforms potential point-supervised baselines.*

1. Introduction

Oriented object detection in aerial images aims to locate the objects of interest using oriented bounding boxes (OBBs) and identify the categories. There have been numerous outstanding research studies in this field [10, 17, 27, 28, 43, 47–49]. However, manually annotating the fine-grained OBBs is time-consuming and expensive.

As to reducing the annotation cost, weakly-supervised horizontal object detection using image-level annotations has been well-developed [2, 6, 13, 14, 26, 37, 38, 41], but these methods have limited performance in complex aerial scenes and cannot predict the orientation of objects. In recent years, there has been growing attention to weakly supervised oriented object detection. As illus-
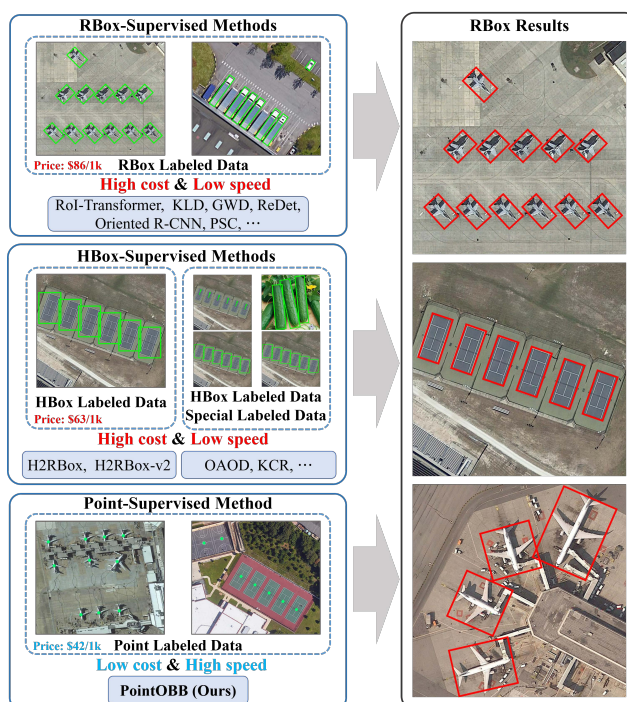

Figure 1. The main paradigmatic types of existing oriented object detection. Existing methods are primarily divided into rotated boxes (RBox)-supervised, horizontal boxes (HBox)-supervised, and Point-supervised approaches. Compared to RBox and HBox labels, point labels have lower costs and higher efficiency.

trated in Fig. 1, existing weakly-supervised methods employ coarser-grained annotations as weakly supervised signals to predict OBBs, *e.g.*, horizontal box [35, 50, 55, 57] annotations. However, box-based annotations are still inefficient and labor-intensive. Therefore, it is necessary to explore more cost-effective and efficient annotation forms.

Point-based annotation has recently gained attention in many tasks [7, 12, 25, 34]. In the field of object detection, the cost of point annotations is about 50.0% lower than HBB and 104.8% lower than OBB, and the efficiency of labeling point is significantly higher than both HBB and OBB[1]. Therefore, single point-supervised oriented object detection seems more meaningful, and as far as we know, the relevant research is still blank.

From the bird's-eye view, objects in aerial images showcase two specific characteristics: various spatial scales and arbitrary orientations. Considering plenty of tiny objects existing in aerial images [42], employing single point labels is undoubtedly more appropriate than multi point labels like [7]. Existing single point-supervised object detection methods [5, 33] follow the Multiple Instance Learning (MIL) fashion. This fashion optimizes through category labels, selecting the proposals with the highest confidence from proposal bags as the predicted boxes, thereby achieving perception of the object scale. However, the MIL fashion faces inherent instability in perceiving object boundaries, as it tends to focus on the most discriminative part of an object instead of its exact scale and boundary [37, 38]. We identify two critical issues when extending this fashion to oriented object detection in aerial images: i) How to address the inconsistency in MIL to obtain more accurate scale representations? ii) How can we learn the object's orientation under single point supervision?

In this paper, we introduce **PointOBB**, the first single **Point**-based **OBB** generation framework. It achieves collaborative learning of both angle and scale by incorporating three unique views. Specifically, we construct a resized view and a rotated/flipped (rot/flp) view as enhanced views. To address the abovementioned two issues, we propose two modules: a scale enhancement module via the original and resized views, and an angle acquisition module relying on the original and rot/flp views. The core of the former module is a Scale-Sensitive Consistency (SSC) loss, which aims to address the aforementioned inconsistency between the proposal's confidence score and its scale accuracy. To obtain accurate object angle prediction, we propose a Dense-to-Sparse (DS) matching strategy associated with the self-supervised angle learning branch in the latter module. Additionally, the enhanced views are switched during training through a progressive multi-view switching strategy. The

---

[1]According to https://cloud.google.com/ai-platform/data-labeling/pricing, and point annotations are just 1.2x more time consuming than obtaining image-level annotations [1, 11].

main contributions are as follows:

- As far as we know, this paper proposes the first method, named PointOBB, to achieve oriented object detection under single point supervision.
- The proposed method couples learning the object's scale and orientation through three distinctive views, guided via a progressive multi-view switching strategy.
- A SSC loss is designed to enhance the network's ability to perceive the object scale, and a scale-guided DS matching strategy is introduced to improve the accuracy of object angle prediction.
- Our method significantly outperforms existing competitive point-supervised alternatives on DIOR-R and DOTA-v1.0 datasets.

2. Related works

2.1. Fully-Supervised Oriented Object Detection

Oriented object detection algorithms primarily focus on aerial objects [44], multi-oriented scene texts [29, 32], 3D-objects [51], etc. Representative approaches include anchor-free detectors like Rotated FCOS [39], and anchor-based detectors such as Rotated RetinaNet [31], RoI Transformer [10], Oriented R-CNN [43], and ReDet [17]. Oriented RepPoints [23] proposes a method for sample quality assessment and allocation based on adaptive points. Some methods like R^3Det [47] and S^2A-Net [16] enhance the performance of the detectors through exploring feature alignment modules. To address the boundary discontinuity in angle regression, angle coders [45, 46, 54] transform the angle into boundary-free formats. Moreover, Gaussian-based losses like GWD [48] and KLD [49] analyze the nature of rotation representation and propose a Gaussian-based representation to improve performance.

2.2. Weakly-Supervised Oriented Object Detection

Existing weakly-supervised oriented object detection approaches can be divided into Image-supervised and HBox-supervised methods. Furthermore, we explore the feasibility of Point-supervised methods.

Image-Supervised. For methods using image-level annotation, WSODet [36] enhances the OICR [37] framework to predict HBBs. Then, it generates pseudo OBBs using contour features and predicted HBBs, followed by an Oriented RepPoints branch. However, image-supervised methods yield limited performance, and the generated OBBs are overly reliant on the quality of predicted HBBs.

HBox-Supervised. For methods using HBox-level annotation, although some methods (*e.g.*, BoxInst-RBox [40] and BoxLevelSet-RBox [24]) use the HBox-Mask-RBox style to generate OBBs, involving segmentation is often more computationally costive, and the whole procedure can be time-consuming [50]. H2RBox [50] pioneeringly pre-
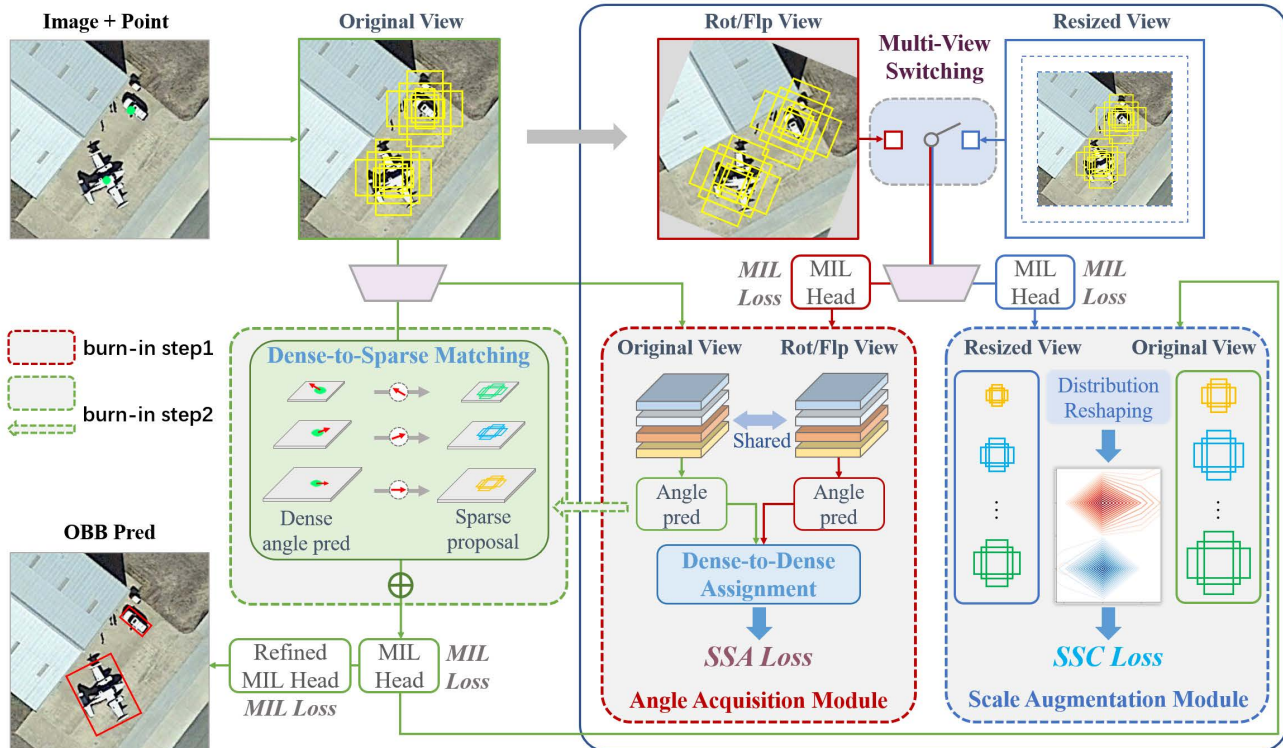
Figure 2. The pipeline of the PointOBB. PointOBB contains three views in total. From these views, an angle acquisition module and a scale augmentation module are constructed. The scale augmentation module incorporates a Scale-Sensitive Consistency (SSC) loss to enhance scale perception ability. Within the angle acquisition module, a Dense-to-Dense sample assignment is designed for angle learning, optimized using the Self-Supervised Angle (SSA) loss. Additionally, a Dense-to-Sparse (DS) matching strategy is proposed to obtain more precise object angles. During the training, a progressive multi-view switching strategy is designed to switch the resized and rot/flp views, along with their corresponding modules.

dicts the RBox directly from HBox annotations without redundant representations. It learns the angle from the geometry of circumscribed boxes and achieves remarkable performance. As a new version, H2RBox-v2 [55] exploits the inherent symmetry of objects. However, such methods still require collecting a large number of bounding box annotations. Additionally, some studies utilize HBox and special forms of annotation. OAOD [20] uses extra object angle, KCR [57] employs RBox-annotated source datasets with HBox-annotated target datasets. Sun et al. [35] combine HBox annotation and image rotation to align oriented objects horizontally or vertically. However, these specialized annotation forms lack universality.

Point-Supervised. Point-based annotations have been widely used in various tasks like object detection [4, 5, 19, 33, 34, 52], panoptic segmentation [12, 25], instance segmentation [1, 7], and so on [21, 53]. Due to its cost-effectiveness and efficiency, single-point supervised object detection has garnered attention. Click [33] makes an early exploration of point-supervised object detection. It proposes center-click annotations and incorporates them into MIL fashion. P2BNet [5] employs a coarse-to-fine strategy

and adds negative samples to improve the quality of predictions. However, these methods only obtain horizontal boxes and disregard the inherent instability of MIL fashion.

Based on the above, to finally obtain RBox through the point label, one potential method is the Point-to-Mask [25] approach, which entails finding the minimum circumscribed rectangle of the mask. Another possible method involves simply combining the Point-to-HBox and HBox-to-RBox approaches. In our experiments, these potential approaches are employed for comparison. Overall, no existing methods can directly achieve oriented object detection via single point supervision. This paper aims to fill this blank and provide a valuable starting point.

3. Method

3.1. Overall Framework

In existing image-supervised and point-supervised object detection methods, MIL-based paradigms demonstrate fundamental perceptual capabilities for object scale. Therefore, we employ the classic MIL fashion as the underlying network. The overall framework of our approach is illustrated

in Fig. 2. From the original view, initial proposal bags are generated via the point labels. Then angle predictions from the angle acquisition module are selected to align with these proposals through the DS matching strategy, equipping the horizontal proposals with orientation. From the generated rotated proposals, reliable ones are selected and refined via a MIL head and a refined MIL head to obtain the final OBB predictions. The acquired OBBs serve as pseudo labels for the final training of oriented object detectors.

The above pipeline employs three unique views in total. Based on the original view, the resized view is created by random scaling, while the rot/flp view is created by random rotating or vertical flipping. Through these three views, we construct a scale augmentation module and an angle acquisition module. The former module leverages original and resized views to enhance the network in perceiving object scales, while the latter module is designed to learn object angles. Two extra views serve as enhanced views and are switched during the training process by the proposed progressive multi-view switching strategy.

3.2. Progressive Multi-View Switching Strategy

To enable the network to gradually acquire discriminative capabilities for object scale and predictive abilities for object orientation, we design a progressive multi-view switching strategy to optimize our framework.

This strategy consists of three stages: i) In the **first stage**, we begin by constructing a resized view from the original view using a scale factor σ. Leveraging the scale-equivalence constraints between these two views, we design a scale augmentation module to enhance the network's accuracy in perceiving object scales, which will be elaborated in subsequent sections. ii) In the **second stage** (*i.e.*, "burn-in step1" in Fig. 2), the network has acquired fundamental perceptual abilities to the scale and boundary of objects. However, the orientation information is still lacking. Therefore, we switch the resized view to a rot/flp one to construct an angle acquisition module with the original view. This module utilizes the dense-to-dense sample assignment for self-supervised angle learning. iii) In the **third stage** (*i.e.*, "burn-in step2" in Fig. 2), the network has been able to obtain accurate angles from the dense feature. Utilizing the proposed DS matching strategy, we align the dense angle predictions with the sparse proposals by leveraging neighboring receptive fields to obtain the object orientation.

3.3. Scale Augmentation Module

Objects in aerial images exhibit significant scale variation. Therefore, under the MIL paradigm, the inconsistency between the confidence scores of the predicted boxes and the actual positional accuracy will be further exacerbated. To address this issue, we introduce the scale augmentation module, which is centered around the design of Scale-
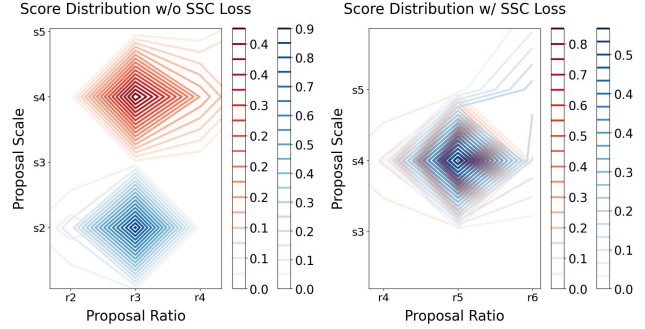

Figure 3. Distribution of the same proposal group's instance scores before and after applying the proposed SSC loss. We apply a local magnification of the coordinates for better visibility, and the scores are depicted on the right-hand side. The red lines and the blue lines represent scores from the original view and resized view, respectively. It can be seen that the SSC loss effectively narrows the gap between these two score distributions.

Sensitive Consistency (SSC) loss.

In an ideal situation, the predicted size for the same object — the scale of the proposal with the highest confidence score — should be consistent across views with varying resolutions. Guided by this criterion, the SSC loss aims to minimize the disparity in distributions of predicted scores between the original and the resized views.

Given the original view's proposal bags B_o, we get the corresponding resized proposal bags B_d from the resized view. The output scores (*i.e.*, class scores and instance scores) of B_o and B_d are obtained by the dual-stream branches within the classical MIL head. To the i-th proposal bag, the scores derived from the original and resized views are denoted as $S_{i_o}^{cls}, S_{i_o}^{ins}$ and $S_{i_d}^{cls}, S_{i_d}^{ins}$, respectively. These sets of scores have been processed through an activation function (*e.g.*, softmax), and their dimensions are $\mathbb{R}^{N \times C}$, where N indicates the number of proposals in this bag and C indicates the number of categories. To ensure scale equivalence among the outputs of different views, we first reshape the proposals' score distributions by the scale. Specifically, we define a set of basic scales $\{s_1, s_2, ..., s_G\}$, where G is the total number of the basic scales. The dimension of the output scores is reshaped from $\mathbb{R}^{N \times C}$ to $\mathbb{R}^{G \times K}$, where K represents the number of scale-independent variables, such as aspect ratio and category. After reshaping the score distributions scale-wise from different views, we employ cosine similarity to measure their consistency:

$$sim_{m,g}^{ins} = 1 - \frac{[S_{i_o}^{ins}]_{m,g} \cdot [S_{i_d}^{ins}]_{m,g}}{\|[S_{i_o}^{ins}]_{m,g}\| \cdot \|[S_{i_d}^{ins}]_{m,g}\|}, \quad (1)$$

$$sim_{m,g}^{cls} = 1 - \frac{[S_{i_o}^{cls}]_{m,g} \cdot [S_{i_d}^{cls}]_{m,g}}{\|[S_{i_o}^{cls}]_{m,g}\| \cdot \|[S_{i_d}^{cls}]_{m,g}\|}, \quad (2)$$

where m represents the m-th point label, and g represents the g-th group of proposals after being grouped based on the

basic scales. With the similarity measurements, the overall SSC loss is formulated as:

$$\mathcal{L}_{SSC} = \sum_{m=1}^{M} \sum_{g=1}^{G} \left\{ \omega_1 \ell_s(sim_{m,g}^{ins}, 0) + \omega_2 \ell_s(sim_{m,g}^{cls}, 0) \right\},$$

(3)

where M is the number of point labels, ℓ_s is SmoothL1 loss, ω_1 and ω_2 are weights and set to 2.0 and 1.0, respectively. Utilizing the SSC loss, the MIL network aligns the distribution of scores for proposals from the same label across different views, as depicted in Fig. 3. This alignment helps mitigate inconsistencies between confidence scores and positional precision, enhancing the accuracy of perceiving object scale.

3.4. Angle Acquisition Module

To learn the orientation under the absence of angle supervision, we begin by considering the inherent symmetry of objects. Previous exploration has studied symmetry under HBox supervision [55]. We discover that symmetry-based self-supervised learning has robustness to annotation noise. This indicates that even from a single point, accurate angle prediction is potential.

However, the absence of scale information makes it challenging to apply common sample assignment strategies in angle learning. Therefore, it is crucial to select appropriate samples for angle learning and match the angle prediction with corresponding objects to obtain final OBBs. To achieve this, we construct the angle acquisition module via the rot/flp view, with the proportion between rotate and flip set to 95:5 as [55]. Angle learning is accomplished through dense-to-dense sample assignment, and the dense-to-sparse matching strategy realizes angle matching.

Dense-to-Dense Assignment. As described in Sec. 3.1, in the second stage, an angle acquisition module is built upon the rot/flp view, and it contains a self-supervised angle branch for angle learning. Both the views are sent into parameter-shared feature extractors (*e.g.*, ResNet50 [18] and FPN [30]) to obtain dense feature pyramid features. As scale information is lacking, we select grid points on all feature levels on the center area around the ground-truth points as positive samples. For positive samples corresponding to the same point label in a level, we obtain the average of their predicted angles as the prediction value.

Dense-to-Sparse Matching. To match the dense feature-based angle predictions and the sparse feature-based proposals, simply searching for the nearest grid points to the center of a proposal is inappropriate. Due to the potential disparity between the grid points' receptive field and the object proposals' scale, the angle predictions may not be based on the actual object region. We perform a hierarchical pairing to consistently match the receptive fields involved in angle prediction with the scale of proposals. Assuming the

feature pyramid is $lvls_{fea} = \{p_1, p_2, ...p_P\}$ with P levels, we also classify the proposal into several levels based on the scale, $lvls_{prop} = \log_2 \left(\frac{\sqrt{w \cdot h}}{basescale} + 1e - 6 \right)$, where $basescale$ is a pre-set scale parameter, w and h represent the width and height of the proposal, respectively. For each corresponding pair of levels $lvls_{fea}$ and $lvls_{prop}$, we utilize the average angle predictions from the central region as the orientation of the proposal, as shown in Fig. 2, thereby aggregating the dense angle predictions for the sparse objects.

SSA Loss. According to the affine transformation relationship between the original view and the rot/flp view, object angles can be learned through the Self-Supervised Angle (SSA) loss. Assuming that the enhanced view is generated via random rotation with an angle θ', the angle predictions from both the original and rotated view are supposed to satisfy the same rotation relationship. If the enhanced view is generated via vertical flip, the angle predictions should satisfy the differences $k\pi$, where k is an integer to keep the results in the same cycle. The loss between the outputs of two views can be represented as:

$$\begin{cases} \mathcal{L}_{\text{rot}} = \min_{k \in \mathbb{Z}} \sum_{p=1}^{P} \ell_{\text{angle}}(\theta_{\text{rot}}^p - \theta^p, k\pi + \theta') \\ \mathcal{L}_{\text{flp}} = \min_{k \in \mathbb{Z}} \sum_{p=1}^{P} \ell_{\text{angle}}(\theta_{\text{flp}}^p + \theta^p, k\pi) \end{cases},$$

(4)

where ℓ_{angle} is SmoothL1 loss, θ^p, θ_{flp}^p, and θ_{rot}^p represent the angle predictions from the p-th level of the feature pyramids from the original view, rot/flp view by flip, and rot/flp view by rotation, respectively. The SSA loss is defined as:

$$\mathcal{L}_{SSA} = \mathcal{L}_{\text{rot}} + \mathcal{L}_{\text{flp}}.$$

(5)

3.5. Optimization with Overall MIL Loss

In addition to the previously mentioned SSC and SSA loss, we introduce the MIL loss involved in our approach. As in Fig. 2, each constructed view undergoes the MIL head, and the output of the original view passes through an extra refined MIL head to obtain the final predictions. The corresponding loss can be summarized as follows.

Specifically, each point label is associated with a corresponding proposal bag. Through the MIL head, the output instance scores and class scores of the n-th proposal in the i-th proposal bag is $\left(S_{i,n}^{ins}, S_{i,n}^{cls} \right)$. The output of the whole proposal bag can be denoted as $S_i = \sum_{n=1}^{N} S_{i,n}^{ins} \odot S_{i,n}^{cls}$, where N is the number of proposals in this bag. The MIL losses for the original view, rot/flp view, and resized view are denoted as $\mathcal{L}_{MIL_{ori}}$, $\mathcal{L}_{MIL_{rfv}}$, and $\mathcal{L}_{MIL_{res}}$, respectively. These three losses follow the common form of the general MIL loss $\mathcal{L}_{MIL_{init}}$, which is defined as:

$$\mathcal{L}_{MIL_{init}} = -\frac{1}{I} \sum_{i=1}^{I} \sum_{c=1}^{C} \Big\{ Q_{i,c} \log(S_{i,c})$$

$$+ (1 - Q_{i,c})\big(1 - \log(1 - S_{i,c})\big) \Big\},$$

(6)

Method	APL	APO	BF	BC	BR	CH	ESA	ETS	DAM	GF	GTF	HA	OP	SH	STA	STO	TC	TS	VE	WM	8-mAP$_{50}$	mAP$_{50}$
RBox-supervised:																						
Rotated RetinaNet [31]	58.9	19.8	73.1	81.3	17.0	72.6	68	47.3	20.7	74.0	73.9	32.5	32.4	75.1	67.2	58.9	81.0	44.5	38.3	62.6	64.26	54.96
Rotated FCOS [39]	61.4	38.7	74.3	81.1	30.9	72.0	74.1	62.0	25.3	69.7	79.0	32.8	48.5	80.0	63.9	68.2	81.4	46.4	42.7	64.4	66.59	59.83
Oriented R-CNN [43]	63.1	34.0	79.1	87.6	41.2	72.6	76.6	65.0	26.9	69.4	82.8	40.7	55.9	81.1	72.9	62.7	81.4	53.6	43.2	65.6	**69.88**	**62.80**
Image-supervised:																						
WSODet [36]	20.7	29.0	63.2	67.3	0.2	65.5	0.4	0.1	0.3	49.0	28.9	0.3	1.5	1.2	53.4	16.4	40.0	0.1	6.1	0.1	28.46	22.20
HBox-supervised:																						
H2RBox [50]	68.1	13.0	75.0	85.4	19.4	72.1	64.4	60.0	23.6	68.9	78.4	34.7	44.2	79.3	65.2	69.1	81.5	53.0	40.0	61.5	**67.80**	**57.80**
H2RBox-v2 [55]	67.2	37.7	55.6	80.8	29.3	66.8	76.1	58.4	26.4	53.9	80.3	25.3	48.9	78.8	67.6	62.4	82.5	49.7	42.0	63.1	64.06	57.64
Point-supervised:																						
Point2Mask-RBox [25]	15.6	0.1	50.6	25.4	4.2	50.9	23.8	17.5	8.1	0.8	9.6	1.4	15.3	1.6	5.8	6.3	17.9	7.1	4.5	9.2	15.25	13.77
P2BNet [5] + H2RBox [50]	52.7	0.1	60.6	80.0	0.1	22.6	11.5	5.2	0.7	0.2	42.8	2.8	0.2	25.1	8.6	29.1	69.8	9.6	7.4	22.6	42.65	22.59
P2BNet [5] + H2RBox-v2 [55]	51.6	3.0	65.2	78.3	0.1	8.1	7.6	6.3	0.8	0.3	44.9	2.3	0.1	35.9	9.3	39.2	79.0	8.8	10.3	21.3	45.94	23.61
Ours (Rotated FCOS)	58.4	17.1	70.7	77.7	0.1	70.3	64.7	4.5	7.2	0.8	74.2	9.9	9.1	69.0	38.2	49.8	46.1	16.8	32.4	29.6	54.80	37.31
Ours (Oriented R-CNN)	58.2	15.3	70.5	78.6	0.1	72.2	69.6	1.8	3.7	0.3	77.3	16.7	4.0	79.2	39.6	51.7	44.9	16.8	33.6	27.7	**57.38**	**38.08**

Table 1. Accuracy on the DIOR-R testing set. The categories in DIOR-R include Airplane (APL), Airport (APO), Baseball Field (BF), Basketball Court (BC), Bridge (BR), Chimney (CH), Expressway Service Area (ESA), Expressway Toll Station (ETS), Dam (DAM), Golf Field (GF), Ground Track Field (GTF), Harbor (HA), Overpass (OP), Ship (SH), Stadium (STA), Storage Tank (STO), Tennis Court (TC), Train Station (TS), Vehicle (VE) and Windmill (WM). **8-mAP$_{50}$** means the mAP of APL, BF, BC, GTF, HA, SH, TC, and VE, which are representative categories in remote sensing. "-RBox" means the minimum rectangle operation is performed on the Mask to obtain RBox.

Annotation	Method	7-mAP$_{50}$	mAP$_{50}$
RBox	Rotated RetinaNet [31]	73.19	68.69
	Rotated FCOS [39]	**76.96**	**71.28**
HBox	Sun et al. [35]	32.14	38.60
	H2RBox [50]	70.29	67.21
	H2RBox-v2 [55]	**75.88**	**72.52**
Point	P2BNet [5] + H2RBox [50]	19.02	19.63
	P2BNet [5] + H2RBox-v2 [55]	23.58	21.87
	Ours (Rotated FCOS)	43.35	30.08
	Ours (Oriented R-CNN)	**49.01**	**33.31**

Table 2. Accuracy on the DOTA-v1.0 testing set. **7-mAP$_{50}$** means the mAP of 7 representative categories: Plane (PL), Baseball Diamond (BD), Ground Track Field (GTF), Small Vehicle (SV), Large Vehicle (LV), Ship (SH), and Harbor (HA).

Module		DIOR-R [8]			DOTA-v1.0 [42]		
SSC	DS	mIoU	8-mAP$_{50}$	mAP$_{50}$	mIoU	7-mAP$_{50}$	mAP$_{50}$
		47.95	47.40	30.16	40.26	43.65	28.44
✓		53.17	52.19	36.39	43.92	47.03	32.98
	✓	50.78	49.89	31.96	42.54	44.34	30.63
✓	✓	**56.08**	**54.80**	**37.31**	**45.35**	**49.01**	**33.31**

Table 3. The effects of Scale-Sensitive Consistent (SSC) loss and Dense-to-Sparse (DS) matching strategy in DIOR-R and DOTA-v1.0 datasets. We evaluate Rotated FCOS on the DIOR-R testing set and Oriented R-CNN on the DOTA-v1.0 testing set.

loss of our framework is expressed as:

$$\mathcal{L} = \mathcal{L}_{MIL} + \alpha\mathcal{L}_{SSC} + \beta\mathcal{L}_{SSA}, \qquad (8)$$

where α and β keep the same setting with Eq. 7.

4. Experiments

4.1. Datasets and Implementation Details

Dataset. DIOR-R [8] is an aerial image dataset with OBB annotation based on its previous version DIOR [22], which consists of 23,463 images, 20 categories, and 190,288 instances. DOTA-v1.0 [42] is a large-scale aerial images dataset for object detection, which contains 2,806 images, 15 categories, and 188,282 instances with both OBB and HBB annotations. Algorithms are trained on both the training and the validation set, and evaluated on the testing set.

Single Point Annotation. To accurately reproduce the biases during manual annotation, we generate point labels

where I is the number of proposal bags in the batch, C is the number of categories, $Q_{i,c}$ is the one-hot category label, and $S_{i,c}$ represents the score of c-th category in S_i. For the refined MIL head, the $\mathcal{L}_{MIL_{ref}}$ employs focal loss [31] to calculate the classification loss between $Q_{i,c}$ and $S_{i,c}$. Therefore, considering the proposed progressive multi-view switching strategy, the overall MIL loss is defined as:

$$\begin{aligned}\mathcal{L}_{MIL} = \mathcal{L}_{MIL_{ori}} &+ \mathcal{L}_{MIL_{ref}} \\ &+ \alpha\mathcal{L}_{MIL_{rfv}} + \beta\mathcal{L}_{MIL_{res}},\end{aligned} \qquad (7)$$

where α is 1, β is 0 in the first stage, and α is 0, β is 1 in the second and third stages as depicted in Sec. 3.2. The overall
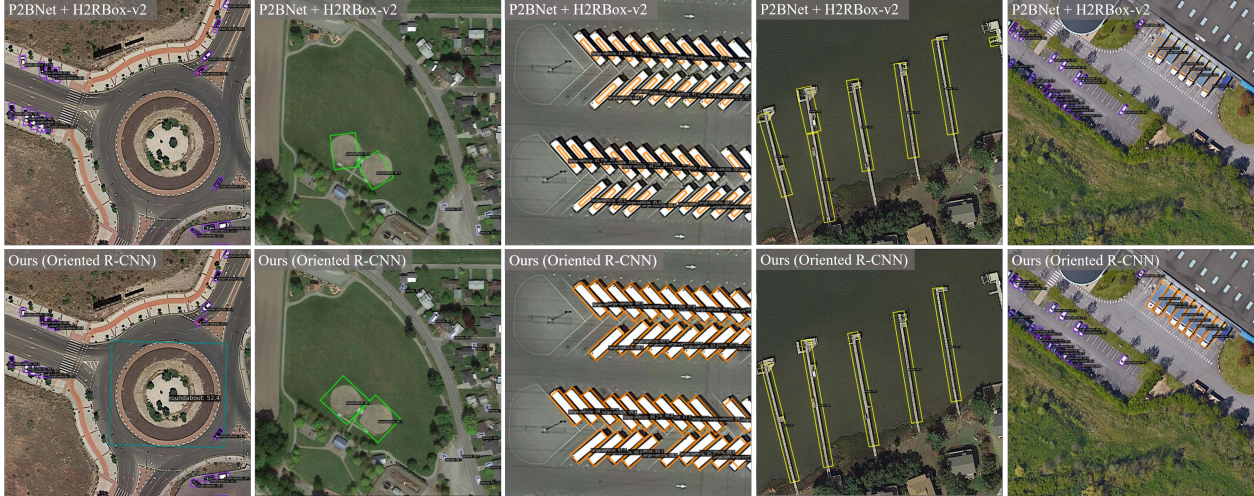
Figure 4. The visual detection results. The first row showcases the results of the P2BNet [5]-H2RBox-v2 [55] cascade pipeline. The second row showcases the results of the proposed PointOBB combined with Oriented R-CNN.

Original	Resized	Rot/Flp	mIoU	8-mAP$_{50}$	mAP$_{50}$
✓			52.90	53.45	34.94
✓	✓		54.44	54.73	36.26
✓		✓	53.49	54.92	35.80
✓	✓	✓	**56.08**	**57.38**	**38.08**

Table 4. Ablation studies of using MIL loss at different views on DIOR-R, using Oriented R-CNN.

based on the central region of the OBB labels. Specifically, we select a random point within a range comprising 10% in height and width relative to the OBB as the label. The effect of the range will be discussed in Sec. 4.3.

Experiment Settings. The algorithms employed in our experiments are from two open-source pytorch-based algorithm libraries, MMRotate [56] and MMDetection [3]. We follow the default setting in MMRotate, for the DOTA-v1.0 dataset, large-size images are cropped into $1,024 \times 1,024$ patches with a 200-pixel overlap. For the DIOR-R dataset, we keep the image size at the original size of 800×800.

Experiments are performed on a server with 2 Tesla V100 GPUs and 16GB memory. We adopt the "2×" schedule for training all methods and adopt the "1×" schedule when using our generated pseudo OBBs to train RBox-supervised algorithms. The SGD optimizer is employed with a learning rate of 0.005, momentum of 0.9, and weight decay of 0.0001. A linear warm-up strategy is applied for the initial 500 iterations with a rate of 0.001 and batch size is 2. Apart from WSODet [36] using VGG16 [15], all the listed models are configured based on the ResNet50 [18] backbone which is first pre-trained on ImageNet [9]. For the algorithms used for comparison, we follow their default settings. For our method, the *basescale* is set to 56 and the resized view's scale factor σ is randomly chosen between

0.5 and 1.5. The "burn-in step1" and "burn-in step2" are set to the 6th epoch and the 8th epoch, respectively. All angle definitions follow le90, and random flipping is employed as the only data augmentation technique in training.

Evaluation Metric. Mean Average Precision (mAP) is adopted as the primary metric to compare our methods with existing alternatives. To better assess the quality of pseudo-labels generated from points, we report the mean Intersection over Union (mIoU) between ground-truth boxes and predicted pseudo OBBs in the training set.

4.2. Main Results

The comparisons on the DIOR-R and DOTA-v1.0 datasets are shown in Tab. 1 and Tab. 2.

Results on DIOR-R. As shown in Tab. 1, PointOBB achieves AP$_{50}$ of 37.31% and 38.08% by training Rotated FCOS [39] and Oriented R-CNN [43]. For comparison, we employ state-of-the-art Point-to-HBox-to-RBox (*i.e.*, P2BNet [5] + H2RBox-v2 [55]) and Point-to-Mask-to-RBox (*i.e.*, Point2Mask-RBox [25]) as plain two-stage approaches. Our results significantly outperform these two-stage approaches, exhibiting a minimum improvement of 14.47% in the mAP$_{50}$ metric. What's more, it is worth noting that our method results in performances of over 85% comparable to fully-supervised methods in eight major categories (*i.e.*, the 8-mAP$_{50}$ metric), demonstrating the potential of point supervision.

Results on DOTA-v1.0. As shown in Tab. 2, our method achieves 33.31% in mAP$_{50}$ metric with Oriented R-CNN, surpassing the plain two-stage approach by 11.44%.

The visual result is shown in Fig. 4. The proposed PointOBB effectively addresses the issue of local focus MIL fashion and accurately predicts the object orientation.
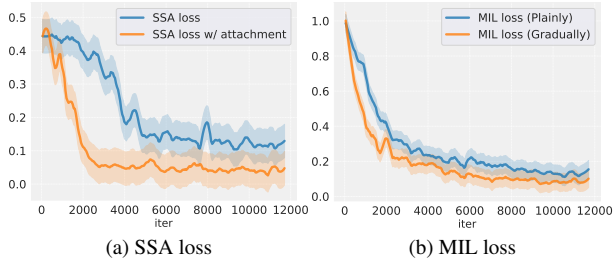
(a) SSA loss   (b) MIL loss

Figure 5. (a) indicates the SSA loss under different settings. In (b), "MIL loss (Gradually)" means training the network with the proposed gradual strategy (*i.e.*, in chronological order), and "MIL loss (Plainly)" means training the angle branch from the beginning directly.

Plainly	Gradually	Attachment	mIoU	8-mAP$_{50}$	mAP$_{50}$
✓			34.92	36.13	23.71
✓		✓	37.45	40.85	25.52
	✓		48.92	50.22	32.23
	✓	✓	**56.08**	**57.38**	**38.08**

Table 5. Ablation studies of the burn-in steps on DIOR-R. "Attachment" indicates attaching gradients between the angle branch and the other parts.

Parameter	Grouping Type			Point Range		
Setting	proposal	ratio	scale	0%	10%	20%
mIoU	43.95	54.38	**56.08**	49.47	**56.08**	54.21
8-mAP$_{50}$	45.43	55.45	**57.38**	48.45	**57.38**	53.65
mAP$_{50}$	30.42	36.71	**38.08**	32.11	**38.08**	35.78

Table 6. Ablation studies of the grouping type used in the SSC loss, and the generation range of point annotation on DIOR-R.

4.3. Ablation Studies

In this section, we conduct several ablation studies to validate key parts of PointOBB.

The effect of SSC and DS. Tab. 3 studies the effects of the proposed two key designs: the SSC loss and the DS matching strategy. As shown in Tab. 3, both contribute to performance improvement. It indicates that the scale-consistency-based SSC loss has a significant effect, and the DS matching strategy also brings a boost by mitigating misalignment arising from unknown scales.

The effect of different MIL losses. Tab. 4 illustrates the effects of applying MIL loss at different views. It can be observed that both enhanced views contribute to the improvement of accuracy, which may be because they represent different forms of data augmentation.

The effect of the angle branch's setting. We explore the impact of the gradient backpropagation on the angle branch within the angle acquisition module. As shown in Fig. 5 (a)

and Tab. 5, jointly optimizing the angle with the base network ("Attachment") leads to faster convergence and better accuracy. Emphasizing the necessity for collaborative optimization between the angle and scale.

The effect of the burn-in steps. Following the above "Attachment" setting, we explore the impact of burn-in steps through MIL loss. We set burn-in steps to 0 as a "Plainly" strategy, *i.e.*, introducing angle learning at the beginning, and set our defined burn-in steps as "Gradually" strategy. As shown in Fig. 5 (b) and Tab. 5, the "Plainly" strategy affects the initial optimization of the MIL network, resulting in performance degradation. In contrast, the "Gradually" strategy introduces the two burn-in steps when scale and angle learning essentially converge, effectively guiding the network to optimize in a progressive manner, thus yielding improvements.

The effect of point range. The right part of Tab. 6 displays the impact of different ranges during point label generation. The introduction of appropriate noise proves to be advantageous compared to the center point label (*i.e.*0%). We analyze that in the case of certain categories, the center point constrains the network's perceptual range, hindering the perception of object boundaries.

The effect of grouping type in SSC loss. In the left part of Tab. 6, we assess the impact of grouping scores based on scale, ratio, and proposal in the SSC loss. This demonstrates the effectiveness of scale-based grouping, aligning with the original intention of our design.

5. Conclusion

This paper introduces PointOBB, the first single point-based OBB generation framework for oriented object detection. PointOBB comprehensively learns the scale and orientation of objects through three distinctive views and is guided by a progressive multi-view switching strategy. Utilizing the three views, we construct a scale augmentation module and an angle acquisition module. The scale augmentation module enhances the network's capacity to perceive object scale by incorporating a Scale-Sensitive Consistency (SSC) loss. The angle acquisition module achieves self-supervised angle learning and further contributes to improving the accuracy of object angle prediction via Dense-to-Sparse (DS) matching. Our method outperforms the state-of-the-art alternatives on the DIOR-R and DOTA-V1.0 datasets. We hope this work can serve as a meaningful starting point for the development of single point-supervised oriented object detection.

Future work. PointOBB exhibits suboptimal accuracy for specific classes (e.g., BR, GF, and OP), possibly due to their relatively ambiguous definition of boundaries. This motivates further exploration and utilization of object features in aerial images, such as contextual features.

References

[1] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European Conference on Computer Vision*, pages 549–565. Springer, 2016. 2, 3

[2] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2846–2854, 2016. 1

[3] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 7

[4] Liangyu Chen, Tong Yang, Xiangyu Zhang, Wei Zhang, and Jian Sun. Points as queries: Weakly semi-supervised object detection by points. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8823–8832, 2021. 3

[5] Pengfei Chen, Xuehui Yu, Xumeng Han, Najmul Hassan, Kai Wang, Jiachen Li, Jian Zhao, Humphrey Shi, Zhenjun Han, and Qixiang Ye. Point-to-box network for accurate object detection via single point supervision. In *European Conference on Computer Vision*, pages 51–67. Springer, 2022. 2, 3, 6, 7

[6] Ze Chen, Zhihang Fu, Rongxin Jiang, Yaowu Chen, and Xian-Sheng Hua. Slv: Spatial likelihood voting for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12995–13004, 2020. 1

[7] Bowen Cheng, Omkar Parkhi, and Alexander Kirillov. Pointly-supervised instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2617–2626, 2022. 2, 3

[8] Gong Cheng, Jiabao Wang, Ke Li, Xingxing Xie, Chunbo Lang, Yanqing Yao, and Junwei Han. Anchor-free oriented proposal generator for object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022. 6

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 7

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

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2010. 2

[12] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Pointly-supervised panoptic segmentation. In *European Conference on Computer Vision*, pages 319–336. Springer, 2022. 2, 3

[13] Xiaoxu Feng, Xiwen Yao, Gong Cheng, and Junwei Han. Weakly supervised rotation-invariant aerial object detection

[14] Xiaoxu Feng, Xiwen Yao, Hui Shen, Gong Cheng, Bin Xiao, and Junwei Han. Learning an invariant and equivariant network for weakly supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[15] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. 7

[16] Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2021. 2

[17] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Redet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2786–2795, 2021. 1, 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5, 7

[19] Shitian He, Huanxin Zou, Yingqian Wang, Boyang Li, Xu Cao, and Ning Jing. Learning remote sensing object detection with single point supervision. *arXiv preprint arXiv:2305.14141*, 2023. 3

[20] Javed Iqbal, Muhammad Akhtar Munir, Arif Mahmood, Afsheen Rafaqat Ali, and Mohsen Ali. Leveraging orientation for weakly supervised object detection with application to firearm localization. *Neurocomputing*, 440:310–320, 2021. 3

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 3

[22] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:296–307, 2020. 6

[23] Wentong Li, Yijie Chen, Kaixuan Hu, and Jianke Zhu. Oriented reppoints for aerial object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1829–1838, 2022. 2

[24] Wentong Li, Wenyu Liu, Jianke Zhu, Miaomiao Cui, Xian-Sheng Hua, and Lei Zhang. Box-supervised instance segmentation with level set evolution. In *European Conference on Computer Vision*, pages 1–18. Springer, 2022. 2

[25] Wentong Li, Yuqian Yuan, Song Wang, Jianke Zhu, Jianshu Li, Jian Liu, and Lei Zhang. Point2mask: Point-supervised panoptic segmentation via optimal transport. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 572–581, 2023. 2, 3, 6, 7

[26] Yansheng Li, Yongjun Zhang, Xin Huang, and Alan L Yuille. Deep networks under scene-level supervision for multi-class

geospatial object detection from remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 146: 182–196, 2018. 1

[27] Yuxuan Li, Qibin Hou, Zhaohui Zheng, Ming-Ming Cheng, Jian Yang, and Xiang Li. Large selective kernel network for remote sensing object detection. *arXiv preprint arXiv:2303.09030*, 2023. 1

[28] Yansheng Li, Junwei Luo, Yongjun Zhang, Yihua Tan, Jin-Gang Yu, and Song Bai. Learning to holistically detect bridges from large-size vhr remote sensing imagery. *arXiv preprint arXiv:2312.02481*, 2023. 1

[29] Minghui Liao, Zhen Zhu, Baoguang Shi, Gui-song Xia, and Xiang Bai. Rotation-sensitive regression for oriented scene text detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5909–5918, 2018. 2

[30] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 5

[31] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017. 2, 6

[32] Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie 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. 2

[33] Dim P Papadopoulos, Jasper RR Uijlings, Frank Keller, and Vittorio Ferrari. Training object class detectors with click supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6374–6383, 2017. 2, 3

[34] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Alexander G Schwing, and Jan Kautz. Ufo 2: A unified framework towards omni-supervised object detection. In *European Conference on Computer Vision*, pages 288–313. Springer, 2020. 2, 3

[35] Yongqing Sun, Jie Ran, Feng Yang, Chenqiang Gao, Takayuki Kurozumi, Hideaki Kimata, and Ziqi Ye. Oriented object detection for remote sensing images based on weakly supervised learning. In *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2021. 2, 3, 6

[36] Zhiwen Tan, Zhiguo Jiang, Chen Guo, and Haopeng Zhang. Wsodet: A weakly supervised oriented detector for aerial object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12, 2023. 2, 6, 7

[37] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2843–2851, 2017. 1, 2

[38] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(1):176–191, 2018. 1, 2

[39] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9627–9636, 2019. 2, 6, 7

[40] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. Boxinst: High-performance instance segmentation with box annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5443–5452, 2021. 2

[41] F Wan, P Wei, Z Han, J Jiao, and Q Ye. Min-entropy latent model for weakly supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(10):2395–2409, 2019. 1

[42] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2018. 2, 6

[43] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented r-cnn for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3520–3529, 2021. 1, 2, 6, 7

[44] Yongchao Xu, Mingtao Fu, Qimeng Wang, Yukang Wang, Kai Chen, Gui-Song Xia, and Xiang Bai. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(4):1452–1459, 2020. 2

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

[46] Xue Yang, Liping Hou, Yue Zhou, Wentao Wang, and Junchi Yan. Dense label encoding for boundary discontinuity free rotation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 15819–15829, 2021. 2

[47] Xue Yang, Junchi Yan, Ziming Feng, and Tao He. R3det: Refined single-stage detector with feature refinement for rotating object. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3163–3171, 2021. 1, 2

[48] Xue Yang, Junchi Yan, Qi Ming, Wentao Wang, Xiaopeng Zhang, and Qi Tian. Rethinking rotated object detection with gaussian wasserstein distance loss. In *International Conference on Machine Learning*, pages 11830–11841. PMLR, 2021. 2

[49] Xue Yang, Xiaojiang Yang, Jirui Yang, Qi Ming, Wentao Wang, Qi Tian, and Junchi Yan. Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. *Advances in Neural Information Processing Systems*, 34:18381–18394, 2021. 1, 2

[50] Xue Yang, Gefan Zhang, Wentong Li, Yue Zhou, Xuehui Wang, and Junchi Yan. H2rbox: Horizontal box annotation is all you need for oriented object detection. In *International Conference on Learning Representations*, 2022. 2, 6

[51] Xue Yang, Gefan Zhang, Xiaojiang Yang, Yue Zhou, Wentao Wang, Jin Tang, Tao He, and Junchi Yan. Detecting rotated objects as gaussian distributions and its 3-d generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4335–4354, 2022. 2

[52] Xinyi Ying, Li Liu, Yingqian Wang, Ruojing Li, Nuo Chen, Zaiping Lin, Weidong Sheng, and Shilin Zhou. Mapping degeneration meets label evolution: Learning infrared small target detection with single point supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 15528–15538, 2023. 3

[53] Xuehui Yu, Pengfei Chen, Di Wu, Najmul Hassan, Guorong Li, Junchi Yan, Humphrey Shi, Qixiang Ye, and Zhenjun Han. Object localization under single coarse point supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4868–4877, 2022. 3

[54] Yi Yu and Feipeng Da. Phase-shifting coder: Predicting accurate orientation in oriented object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13354–13363, 2023. 2

[55] Yi Yu, Xue Yang, Qingyun Li, Yue Zhou, Gefan Zhang, Junchi Yan, and Feipeng Da. H2rbox-v2: Boosting hbox-supervised oriented object detection via symmetric learning. *Advances in Neural Information Processing Systems*, 2023. 2, 3, 5, 6, 7

[56] Yue Zhou, Xue Yang, Gefan Zhang, Jiabao Wang, Yanyi Liu, Liping Hou, Xue Jiang, Xingzhao Liu, Junchi Yan, Chengqi Lyu, et al. Mmrotate: A rotated object detection benchmark using pytorch. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7331–7334, 2022. 7

[57] Tianyu Zhu, Bryce Ferenczi, Pulak Purkait, Tom Drummond, Hamid Rezatofighi, and Anton van den Hengel. Knowledge combination to learn rotated detection without rotated annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2, 3