Point-to-RBox Network for Oriented Object Detection via Single Point Supervision

Yucheng Wang[1]
yucheng.wang@whu.edu.cn
Chu He[2]
chuhe@whu.edu.cn
Xi Chen[1]
robertcx@whu.edu.cn

[1] School of Computer Science
Wuhan University
Hubei, CN

[2] School of Electronic Information
Wuhan University
Hubei, CN

Abstract

The Rotated Bounding Boxes used in Oriented object detection are labor-intensive and time-consuming to annotate manually. Unlike rotated boxes with fine granularity, point-level annotations only provide a single point for each object as supervision, greatly reducing the annotation burden. In this paper, we formalize the problem as using point annotations to generate high-quality pseudo rotated boxes that can be used to train existing detectors. To address the core challenge of generating pseudo rotated boxes, we propose the Point-to-RBox (P2RBox) network. First, we introduce a coarse-to-fine strategy to generate precise pseudo rotated boxes. Second, to account for objects with arbitrary orientation, we design a three-stream detection head guided by orientation-sensitive features in P2RBox to select the best pseudo rotated box. The extensive experiments on the DOTA and DIOR-R datasets indicate that the pseudo rotated boxes generated by P2RBox are viable substitutes for manually annotated rotated boxes. Using pseudo rotated boxes, a fully-supervised object detector can attain more than 90% of the performance achieved by the same detector trained with manually annotations. In addition, our method not only outperforms image-level weakly supervised detectors but also exhibits competitive performance compared to the fully supervised detectors.

1 Introduction

Oriented object detection has been extensively explored in complex scenes that require fine-grained bounding boxes, such as in retail scenes [4, 21], scene text [18], and the aerial images [20, 37]. In contrast to horizontal bounding boxes (HBox), the rotated bounding box (RBox) accurately depicts objects in various intricate situations [27].

Researchers [6, 7, 12] have dedicated their efforts in fully-supervised oriented object detection to improve the performance of detectors. However, they rely on large-scale datasets [5, 29, 38] with fine-grained RBox annotations. Regrettably, in order to correctly label the RBox, the position of the four points must be precisely adjusted so that all four edges align with the object. This process is not only time-consuming but also expensive. To address the

© 2023. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

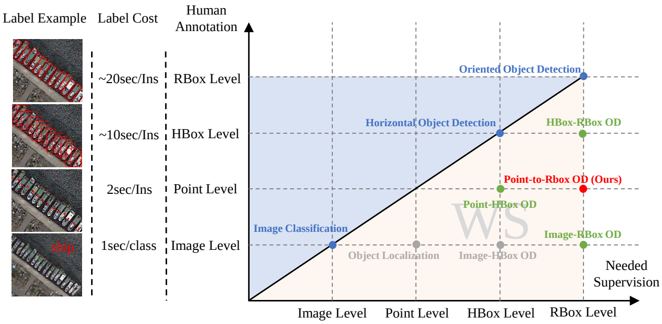


Figure 1: The vertical axis on the figure represents the form of manual annotation with increasing annotating cost, while the horizontal axis represents the form of expected performance results for visual tasks. When low-cost manual annotations are used to obtain high-cost results, it serves as the weakly supervised tasks shown in the red area. The red dot represents the weakly supervised object detection (OD) task where RBox are obtained from point annotation, which has not previously been investigated. The green dots represent weakly supervised tasks related to the task discussed in this paper.

imbalance between the demands of annotation and its associated costs, coarse-grained annotations used in weakly supervised oriented object detection are an effective mean of reducing annotation costs.

Weakly supervised object detection aims to train detectors using only coarse-grained annotations as ground truth to obtain fine-grained detection results required for detection tasks, as Figure 1 depicts. The illustration shows that there are three types of coarse-grained annotation that can generate RBox including image-level annotation, point-level annotation, and HBox annotation. As image-level annotation is the easiest to obtain, the most explored subtask in weakly supervised object detection research is the transformation of image-level annotations into detection results at the bounding box level. The majority of the approaches are built on the foundation of WSDDN [1]. The absence of prior positional information in image-level annotations presents these methods with challenges related to discriminative regions and multiple instances. This ultimately leads to significantly inferior performance when compared to fully supervised detectors. To utilize datasets that already contains HBox annotations without requiring re-labeling, Yang *et al.* [33] utilize weakly supervised learning and symmetry learning to obtain RBox detection results from HBox annotations. However, this method demands manual HBox box annotation when facing unmarked images, thus limiting its applicability. In summary, We believe that image-level and HBox-level annotation are not the best coarse-grained annotation for weakly supervised oriented object detection. Therefore, exploring a more suitable coarse-grained annotation for RBox is of great practical importance.

Point-level annotation has been extensively used in a range of computer vision tasks [2, 3, 8, 13, 17, 23, 36]. Recall that the core of weakly supervised learning is to obtain a model with good performance while reducing the cost of annotation. In comparison to image-level annotation, point-level annotation can significantly enhance the performance of the model at a much lower labeling acquisition cost than bounding boxes [3]. Furthermore, point-level annotation is appropriate for annotating densely distributed or small objects. Therefore, an attractive question arises: **Is it possible to obtain RBox through point-level annotation to achieve weakly supervised oriented object detection?**

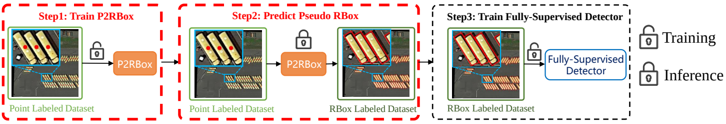


Figure 2: The overall pipeline. Only manually annotated point annotations are available on the original dataset. After completing Step 1 and Step 2, the P2RBox method can generate a pseudo RBox for each point annotation. These pseudo RBoxes can then be used to train arbitrary fully supervised models, as shown in Step 3.

In this paper, we propose P2RBox, a method that can effectively predict high-quality pseudo RBox from single-point annotation. The core idea of this paper is to reduce the cost of annotating RBoxes by using coarse-grained point annotations labeled by humans to obtain corresponding fine-grained RBoxes. The overall process is shown in Figure 2. As illustrated in step 1 and 2 in the figure, this paper trains P2RBox with the dataset that only contains point annotations, and then applies the trained P2RBox to infer corresponding pseudo RBoxes from the point annotations in the original dataset. Through the above process, P2RBox can obtain fine-grained RBox annotations with coarse-grained point annotations labeled by humans, which greatly reduces the time and cost of annotating rotation boxes. Another key advantage of this method is that the pseudo RBoxes generated by P2RBox can be used to train fully supervised oriented object detectors without changing their pipeline, as illustrated in step 3. To address the core challenge of generating high-quality pseudo RBoxes based on single-point annotations, P2RBox utilizes a coarse-to-fine approach, consisting of a Coarse Pseudo RBox Generation stage and multiple Accurate Pseudo RBox Refinement stages, to generate pseudo RBox while balancing accuracy and computing resources. In order to choose the highest quality pseudo RBox when the object's orientation is arbitrary, P2RBox utilizes the ARF [39] produce orientation-sensitive features by encoding orientation information. The Orientation Feature Fusion Stream selects the most informative orientation of each object to guide the prediction of the pseudo RBox.

Our main contributions are as follows:**(1)** To the best of our knowledge, this paper is the first to introduce point-level annotations to weakly supervised oriented object detection. By generating high-quality pseudo RBox for each object, we significantly reduce the manual annotation cost. **(2)** The coarse-to-fine pseudo RBox generation method gradually generates high-quality pseudo RBox while reducing computational resource consumption. Additionally, the three-stream detection head guided by orientation-sensitive features can accurately select the pseudo RBox that best fits the object. **(3)** The experiments on DOTA and DIOR-R show that the pseudo RBox generated by P2RBox can effectively replace manual annotation for training fully-supervised detectors. In addition, our method achieves performance far superior to image-level weakly supervised detectors [24], and demonstrates competitive performance with fully supervised detectors [7, 19, 31].

2 Related Work

Fully Supervised Oriented Object Detection Oriented object detection has recently gained attention in complex scenes such as aerial images due to the compactness of the oriented bounding boxes compared to the horizontal ones. These detectors can be mainly divided into anchor-based and anchor-free categories. Anchor-based detectors [7, 12] relocate preset anchor boxes to locate objects. In order to solve the problem of preset anchor boxes being difficult to match with objects of different aspect ratios and orientations, anchor-free detectors [15] use keypoints to locate objects.

Weakly Supervised Object Detection Weakly supervised object detection aims to train models using only low-cost coarse-grained labels as the ground truth to predict fine-grained detection results, thus accurately localizing and classifying object instances in images. Bilen *et al.* [1] first implemented Multiple Instance Learning (MIL) into weakly supervised object detection. This approach involves generating a bag of proposals for each image and subsequently classifying each proposal using a two-stream detection head. However, it is plagued by two problems. The first one is the discriminative region problem, where the model only focuses on the part of the object that is most discriminative and ignores the rest of the object. The second one is the multiple-instance problem, where the model can easily overlook objects with low scores within the same category. The following research endeavors to resolve the discriminative region problem by employing three methods: utilizing object context area information [22, 28], refining proposals via cascaded classifiers [25, 26], as well as through segmentation-detection collaboration [9, 16].

Point-Level Labels in Visual Tasks Point-level annotation provides stronger prior information regarding the object location compared to image-level annotation, while adding relatively minimal annotation costs [2, 3]. Additionally, the cost of point-level annotation is relatively low when compared to other instance-level annotations such as horizontal or rotated boxes [3]. Recently, point-level annotation has been widely studied in various visual tasks [8, 13, 17, 34]. However, point-level annotation is a relatively new innovation in object detection. Chen *et al.* [2] and Zhang *et al.* [36] used point-level annotation in weakly semi-supervised object detection tasks, and Chen *et al.* [3] designed a network called P2BNet, exclusively for point-level annotation. The network generated proposal boxes for each annotated point, rather than for the entire image, and achieved good performance. The above weakly supervised detectors demonstrate the huge potential of point-level annotation in object detection. Simultaneously, we believe that point-level annotation is well-suited for scenarios where the objects have arbitrary orientations or are densely arranged.

3 Method

The main objective of this paper is to generate high-quality pseudo RBoxes based on point annotations, as shown in Steps 1 and 2 in Figure 2. Therefore, this section gives a comprehensive presentation of the P2RBox method that obtains pseudo RBoxes based on point annotations, and its structure is shown in Figure 3.

3.1 Coarse-to-Fine Pseudo-Rotated Box Generation

Point-level annotation can only provide a prior information about the location and category of the object, which prevents us from adjusting the angle and size of the region proposal during training to obtain pseudo RBox. Therefore, we must directly generate a large number of rotated region proposals with different angles, scales, and aspect ratios to ensure precise detection of all objects in the image. Due to the angle symmetry, the angle range for generating pseudo RBox is 180 degrees, and the angular interval of the pseudo RBox needs to be 1° to minimize the accuracy loss caused by converting a continuous problem into a discrete

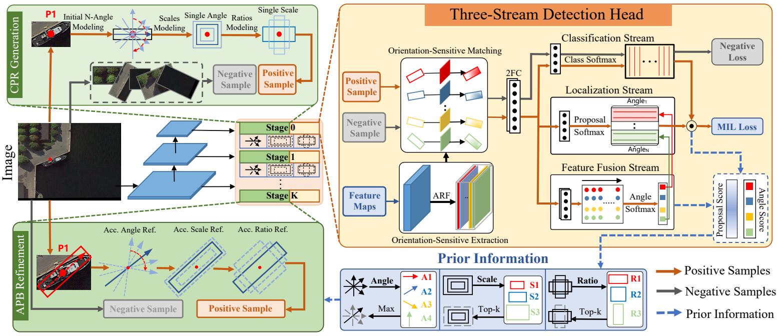


Figure 3: P2RBox's overall structure includes two main components: the coarse-to-fine pseudo RBox generation module, illustrated in green, and three-stream detection head guided by orientation-sensitive features, illustrated in yellow.

one [30]. However, generating rotation region proposals with an angle interval of 1° for all angles in the range of $[-90^\circ, 90^\circ)$ at once is not feasible. Therefore, in this paper, the angles of the generated rotated region proposals are gradually refined by cascading K stages, and the number of angles generated at each stage is N. The detailed structure is shown in the green part of Figure 3.

Coarse Pseudo RBox Generation This stage aims to generate a large number of coarse rotated region proposals based on each point annotation through a combination of preset scales, aspect ratios, and angles. Formally, given an input image $\mathcal{I}$, it comprises $\mathcal{R}$ point annotations $P = \{p_u\}_{u=1}^{R}$, with each point annotation p_u in the form of (x, y, c). Here, (x, y) indicates the point position, and c denotes the corresponding object category. The output of this stage comprises of $\mathcal{R}$ rotated region proposal bags $B = \{B_u\}_{u=1}^{\mathcal{R}}$, where $B_u = \{b_i\}_{i=1}^{M}$. The parameter M represents the number of rotated region proposals generated for each point annotation. For the Coarse Pseudo RBox Generation stage, the angles for each point annotation generated region proposals are given by $\Theta = \{\theta_i \mid \theta_i = -90 + \frac{180}{N} \times i\}$, while the prior scales $S = \{S_v\}_{v=1}$ and aspect ratios $R = \{R_v\}_{v=1}$ are predefined due to the lack of object scale and aspect ratio information. The process of generating the corresponding rotation region proposal bag B_u^0 for each p_u is shown Eq. 1 and 2.

$$B_u = b_1, \ldots, b_M \mid b_i = (p_i.x, p_i.y, h_i, w_i, \theta_i), B_u \in \mathbb{R}^{M \times 5}, M = |S| \times |R| \times |\Theta| \tag{1}$$

$$h_i = S_v \times R_v;\ w_i = S_v \times \frac{1}{R_v} \tag{2}$$

Accurate Pseudo RBox Refinement The purpose of each refinement stage is to refine the angle, scale, and aspect ratio based on prior information obtained from the previous stage, in order to generate more accurate rotated region proposals. Refinement stage k, for instance, acquires prior information for each object from stage k-1. This information includes midpoint coordinates $(p_{u.x}^*, p_{u.y}^*)$, Top-K scales S^* and aspect ratios R^*, and the angle θ^* with the highest score. The current stage generates N angles uniformly to refine the angle in the interval $[\theta^* - G_k, \theta^* + G_k)$, where each angle is determined by Eq. 3. Furthermore, scale and aspect ratio refinement are achieved using Eq. 4. The α denotes the refinement hyperparameter, which is set to 1.2 in this paper. To account for deviation between the point annotation and the object center, this paper uses the method described in [3] to jitter the center point. Using Θ^k, S^k, and R^k obtained earlier in the process, we generate a bag of rotated region proposal $\mathcal{B}_u^k$ in the k-th stage by applying the method in Eq. 1 and 2.

$$\theta_i^k = (\theta^* - G_k) + \frac{2 \times G_k}{N} \times i, i \in N, G_k = \lceil \frac{G_{k-1} \times 2}{N} \rceil \tag{3}$$

$$S^k = \{s^* \times \alpha, s^* \times \frac{1}{\alpha} \mid s^* \in S^*\};\ R^k = \{r^* \times \alpha \mid r^* \in R^*\} \tag{4}$$

3.2 Three-Stream Detection Head guided by Orientation-Sensitive Feature

Once the rotated region proposal bag B_u^k is generated, a three-branch detection head guided by orientation-sensitive features matches each proposal b_i^k with the orientation-sensitive feature $\hat{\mathcal{Y}}$ based on angle. After the matching process is completed, we applied RRoI Align [7] to obtain the corresponding feature F_u^k for B_u^k. The best-quality rotated region proposal is then selected by the three-stream structure to serve as prior information for the next stage.

Orientation-Sensitive Feature Extraction and Matching Inspired by the beneficial effects of orientation-sensitive features on the angle prediction task [11], this paper used ARF [39] to encode the orientation information to produce the orientation-sensitive features. Subsequently, the proposed rotation region was aligned with the orientation-sensitive feature to ensure that its features better represented the orientation of the object. Specifically, ARF F is a $k \times k \times A$ filter that actively rotates $A-1$ times during the convolution process, producing a feature map with A orientation channels (A is 8 by default). For the input feature map $\mathcal{Y}$, the output of the j-th orientation of $\hat{\mathcal{Y}}$ produced by ARF F is shown in Eq. 5. For each rotated region proposal b_i in B_u^k, the orientation channels in $\hat{\mathcal{Y}}$ with the closest rotation angle are matched. From the results of that matching process, corresponding depth feature F_u^k is extracted for B_u^k through RRoI Align. This entire process is demonstrated in Figure 3.

$$\hat{\mathcal{Y}}^j = \sum_{a=0}^{A-1} F_{\theta_j}^a \cdot Y^a, \theta_j = j\frac{\pi}{A}, j = 0, \ldots, A-1 \tag{5}$$

Orientation Feature Fusion Stream This paper first provides a brief introduction to the classification and localization streams that are similar to WSDDN. The classification stream predicts class scores for every proposal, and the localization stream predicts every proposal's associated probability score for each category. Specifically, after the orientation-sensitive feature F_u^k of the rotated region proposal is processed by the classification stream f_{cls} and the localization stream f_{ins} composed of fully connected layers, the corresponding category probability $Score_u^{cls} \in R^{M \times C}$ and localization probability $Score_u^{ins} \in R^{M \times C}$ can be obtained using the Softmax function, where C represents the number of categories. However, when dealing with arbitrary orientation, relying solely on the classification stream and localization stream cannot achieve accurate object orientation prediction.

Inaccurate orientation predictions for rotated region proposals can not only cause a significant reduction in the coincidence degree between the pseudo RBox and the corresponding object, but also seriously affect the orientation prediction of subsequent stages. Therefore, this paper designs an orientation feature fusion stream to select the orientation with the richest feature information for angle prediction of each object. In particular, for the the orientation-sensitive feature F_u^k of the u-th object, after processed by the orientation feature fusion stream f_{ang} consisting of fully connected layers, the score values of each region proposal for every orientation is summed. Then, the Softmax function is applied to obtain the probability of the object facing different orientations, as shown in Eq. 6, where $[\cdot]_i$ denotes the i-th orientation in F_u^k.

$$[Ang_u^k]_i = \sum_{i=0}^{M} [f_{ang}\left(F_u^k\right)]_i;\ [Score_u^{angle}]_j = e(Ang_u^k)_j / \sum_{j=0}^{N} e(Ang_u^k)_j \tag{6}$$

The Hadamard product is computed between $Score_u^{cls}$ and $Score_u^{ins}$, and then every region proposal in the resulting score map is multiplied by its probability at the corresponding angle in $Score_u^{angle}$ to obtain the final score of each region proposal for different categories. Finally, the scores of all region proposals $S_u^k \in R^{M \times C}$ are summed to obtain the final score $\widehat{S_u^k} \in R^C$ of the rotated region proposal bag corresponding to point annotation u. The Eq. 7 illustrate this process, where $[\cdot]_j$ denotes the j-th proposal in B_u^k. The $\widehat{S_u^k}$ can be regarded as the weighted sum of all region proposals in B_u^k, in terms of angle information, classification information and location information. The cross-entropy loss is calculated by using $\widehat{S_u^k}$ with the corresponding point annotation, as represented in Eq. 8. Here, $d_u \in \{0,1\}^C$ refers to the one-hot label representing the category. Meanwhile, we also adopts the same method as Chen *et al.* [3] to select negative samples and calculate negative sample loss.

$$\left[S_u^k\right]_j = \left[Score_u^{cls} \odot Score_u^{cls}\right]_j \times \left[Score_u^{angle}\right]_j;\ \widehat{S_u^k} = \sum_{j=1}^{M} \left[S_u^k\right]_j \tag{7}$$

$$\mathcal{L}_{MIL} = -\frac{1}{R} \sum_{u=1}^{R} \sum_{c=1}^{C} [d_u]_c \, log\left(\left[\widehat{S_u^k}\right]_c\right) + (1 - [d_u]_c) \, log\left(1 - \left[\widehat{S_u^k}\right]_c\right) \tag{8}$$

4 Experiment

4.1 Datasets and Evaluation Method

Datasets **DOTA-v1.0** is presently among the most widely employed datasets for oriented object detection in aerial images. It comprises 2806 images, 188,282 instances annotated with RBoxes, and is classified into 15 categories. For training and testing, we follow a standard protocol by cropping images into 1024 × 1024 patches with a stride of 824. **DIOR-R** is an aerial image dataset annotated by RBoxes based on its horizontal annotation version DIOR [14]. The dataset consists of 23,463 images, 190,288 instances, and is classified into 20 categories.

Evaluation Method The proposed P2RBox aims to generate pseudo RBoxes as similar as possible to manually annotated ones for each object in datasets with point annotations. We used the method presented in reference [3] to obtain quasi-center point annotations for the training sets of both DOTA and DIOR-R. Then, P2RBox was applied to generate pseudo RBoxes based on the point annotations. The effectiveness of P2RBox is evaluated using the following three approaches: **(1)** The ability of P2RBox to convert point annotations into pseudo RBoxes was examined by computing the rotational mIOU between the pseudo RBoxes and the manually annotated RBoxes. **(2)** To verify if the pseudo RBoxes generated by P2RBox are competitive substitutes for manual annotations in training fully-supervised oriented object detectors, we trains various representative fully-supervised detectors [6, 7, 10, 12, 15, 19, 29] with pseudo RBoxes and compares their performance with the versions trained using manually annotated annotations. **(3)** The P2RBox-RFR framework based on point annotations, which consists of the P2RBox and the Rotated Faster RCNN(RFR), is

Method	Label	AP_{DOTA}	AP_{DIOR}
Two-Stages:			
R-FR[29]*	PB	0.656(96%)	0.568(95%)
	GT	0.681	0.595
RoI-T[7]*	PB	0.652(93%)	0.602(94%)
	GT	0.696	0.639
One-Stages:			
R-RN [19]*	PB	0.658(98%)	0.535(98%)
	GT	0.667	0.546
CFA [10]†	PB	0.695(97%)	0.57(98%)
	GT	0.712	0.578
ORep [15]†	PB	0.715(97%)	0.635(98%)
	GT	0.739	0.654
Transformer-Based:			
Ao2-D [6]	PB	0.746(97%)	0.664(94%)
	GT	0.773	0.702

(a)

Method	AP_{DOTA}	AP_{DIOR}
RBox-Supervised		
R-RN	0.667	0.546
RoI-T	0.696	0.639
GWD[31]	0.717	0.578
KLD[32]	0.725	0.58
HBox-Supervised		
H2RB[33]	0.678	0.57
H2RB‡	0.744	–
H2RB2[35]	0.723	0.612
H2RB2‡	0.779	–
Image-Supervised		
WSODet	0.284	0.222
Point-Supervised		
Ours	0.656	0.568
Ours‡	0.713	–

(b)

Table 1: P2RBox is implemented based on the MMRotate framework [40] with 12 epochs. The fully-supervised detectors employed in this paper were implemented following the standard settings for both training and testing. The percentages represent the difference in performance between models trained with PB and those trained with GT. The $*$ and † in (a) respectively denote Anchor-Based and Anchor-Free detectors, while the ‡ in (b) represents Multi-Scale training. *AP* means the standard protocol AP_{50}.

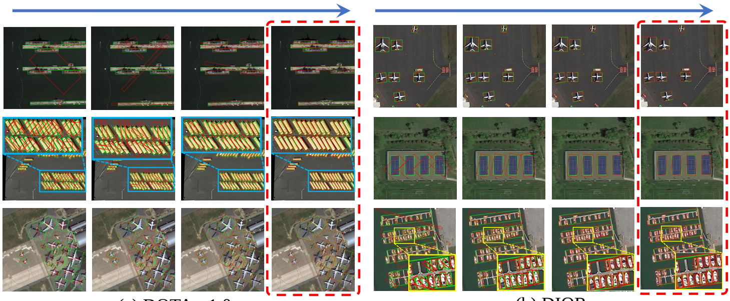
compared with other oriented object detectors using different annotation forms to validate the advantages of point-level annotations in oriented object detection.

4.2 Performance Comparisons

Comparison between Pseudo RBox and Manual Annotation The P2RBox generated pseudo RBoxes with mIOU values of 0.874 and 0.902 on the training sets of DOTA-v1.0 and DIOR datasets, respectively, when compared with the manually annotated RBoxes. Some visualization results are shown in the red dashed box of each dataset in Figure 4. The above results demonstrate that the pseudo RBoxes generated by P2RBox have a high degree of coincidence with manually annotated ones.

Performance Evaluation of Detectors Trained with Pseudo RBoxes The performance of multiple fully-supervised detectors trained on training datasets with pseudo RBoxes(PB) and manually annotated RBoxes(GT) is shown in Table 1(a). Various types of fully-supervised detectors trained with PB can achieve over 90% performance compared to their performance when trained with GT. Therefore, P2RBox can replace manual annotation with pseudo-rotational boxes generated from point annotation for training oriented object detectors, considerably lowering the time and cost associated with manual annotation.

Performance Evaluation of P2RBox-RFR Framework Due to the inability of P2RBox to test on the testing set without point-level annotations, the P2RBox-RFR framework trains RFR detector with pseudo RBoxes produced by P2RBox from the training set and then test



(a) DOTA-v1.0

(b) DIOR

Figure 4: The arrows in the figure indicate the pseudo RBoxes (Red) generated by P2RBox in stages 0 to 3 for each dataset along with the manually annotated RBoxes (Green). As the number of stages increases, P2RBox can generate more precise Pseudo RBoxes. The red dashed box represents the final generated Pseudo RBox. Note that manually annotated RBoxes **cannot be used** at any stage of training.

N	K	mIOU	AP
4	5	0.869	0.642
6	3	0.874	0.656
8	2	0.531	0.415
10	1	0.386	0.227

(a)

OSF Extraction	OSF Fusion	mIOU	AP
–	–	0.793	0.584
√	–	0.821	0.614
–	√	0.844	0.627
√	√	0.874	0.656

(b)

Table 2: mIOU represents the accuracy of the pseudo RBox generated by P2RBox, while AP represents the performance of P2RBox-RFR on DOTA.

RFR on the testing set. The performance comparison of this framework with other detectors is shown in Table 1(b). The comparison results reveal significant improvement in detection performance by the P2RBox-RFR framework compared to the image-level weakly supervised directed object detector WSODet[24], with almost no added annotation cost. Additionally, the small performance gap between our method and some fully supervised detectors suggests that the potential of point-level annotation in practical applications.

4.3 Ablation Study

Number of Refinement stages. P2RBox generates N angles evenly in K stages to create the pseudo RBox with the angle interval $G_k = 1^\circ$ on the initial range of 180°. Thus, the number of angles N determines the corresponding refinement stage number K, and the influence of its value is shown in Table 2 (a). The accuracy of the proposed model is highly dependent on the angle number, represented by N, and the corresponding refinement stage number, represented as K. As the number of stages K increases, the model generates and selects more precise pseudo-rotation boxes. However, the performance of the model will eventually saturate. When N is too large, the refinement degree will be insufficient. The pseudo RBoxes at different stages are shown in Figure 4.

Effectiveness of Orientation-sensitive Feature Extraction and Fusion. Table 2 (b) illustrates the influence of orientation-sensitive feature extraction and fusion on the proposed method's performance. The results indicate that the fusion stream can effectively utilize orientation-sensitive features to achieve precise angle prediction.

5 Conclusion

In this paper, we propose a point-based weakly supervised oriented object detector based on point-level annotation, called P2RBox. The core idea of P2RBox is to generate pseudo RBoxes based on point-level annotations. Experiments on the DOTA and DIOR-R datasets have shown that P2RBox can generated high-quality pseudo RBoxes.

References

[1] 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.

[2] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 8823–8832, 2021.

[3] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 51–67. Springer, 2022.

[4] Zhiming Chen, Kean Chen, Weiyao Lin, John See, Hui Yu, Yan Ke, and Cong Yang. Piou loss: Towards accurate oriented object detection in complex environments. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 195–211. Springer, 2020.

[5] 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] Linhui Dai, Hong Liu, Hao Tang, Zhiwei Wu, and Pinhao Song. Ao2-detr: Arbitrary-oriented object detection transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

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

[8] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Pointly-supervised panoptic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 319–336. Springer, 2022.

[9] Yan Gao, Boxiao Liu, Nan Guo, Xiaochun Ye, Fang Wan, Haihang You, and Dongrui Fan. C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9834–9843, 2019.

[10] Zonghao Guo, Chang Liu, Xiaosong Zhang, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. Beyond bounding-box: Convex-hull feature adaptation for oriented and densely packed object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8792–8801, 2021.

[11] 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.

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

[13] Issam H Laradji, Negar Rostamzadeh, Pedro O Pinheiro, David Vazquez, and Mark Schmidt. Proposal-based instance segmentation with point supervision. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2126–2130. IEEE, 2020.

[14] 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.

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

[16] Xiaoyan Li, Meina Kan, Shiguang Shan, and Xilin Chen. Weakly supervised object detection with segmentation collaboration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9735–9744, 2019.

[17] Youyou Li, Binbin He, Farid Melgani, and Teng Long. Point-based weakly supervised learning for object detection in high spatial resolution remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14: 5361–5371, 2021.

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

[19] 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.

[20] Wei Liu, Tao Zhang, Shengjun Huang, and Kaiwen Li. A hybrid optimization framework for uav reconnaissance mission planning. *Computers & Industrial Engineering*, 173:108653, 2022.

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

[22] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G Schwing, and Jan Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10598–10607, 2020.

[23] Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3365–3374, 2021.

[24] 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.

[25] 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.

[26] Fang Wan, Pengxu Wei, Jianbin Jiao, Zhenjun Han, and Qixiang Ye. Min-entropy latent model for weakly supervised object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1297–1306, 2018.

[27] Kun Wang, Zhang Li, Ang Su, and Zi Wang. Oriented object detection in optical remote sensing images: A survey. *arXiv preprint arXiv:2302.10473*, 2023.

[28] Yunchao Wei, Zhiqiang Shen, Bowen Cheng, Honghui Shi, Jinjun Xiong, Jiashi Feng, and Thomas Huang. Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 434–450, 2018.

[29] 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.

[30] Xue Yang and Junchi Yan. Arbitrary-oriented object detection with circular smooth label. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 677–694. Springer, 2020.

[31] 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.

[32] 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.

[33] Xue Yang, Gefan Zhang, Wentong Li, Xuehui Wang, Yue Zhou, and Junchi Yan. H2rbox: Horizonal box annotation is all you need for oriented object detection. *arXiv preprint arXiv:2210.06742*, 2022.

[34] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 4868–4877, 2022.

[35] 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. *arXiv preprint arXiv:2304.04403*, 2023.

[36] Shilong Zhang, Zhuoran Yu, Liyang Liu, Xinjiang Wang, Aojun Zhou, and Kai Chen. Group r-cnn for weakly semi-supervised object detection with points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9417–9426, 2022.

[37] Tianwen Zhang, Xiaoling Zhang, Chang Liu, Jun Shi, Shunjun Wei, Israr Ahmad, Xu Zhan, Yue Zhou, Dece Pan, Jianwei Li, et al. Balance learning for ship detection from synthetic aperture radar remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 182:190–207, 2021.

[38] Zhengning Zhang, Lin Zhang, Yue Wang, Pengming Feng, and Ran He. Shiprsimagenet: A large-scale fine-grained dataset for ship detection in high-resolution optical remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:8458–8472, 2021.

[39] Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Oriented response networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 519–528, 2017.

[40] 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.