BGHR: Bridging the Gap Between HBox-Supervised and RBox-Supervised Oriented Object Detection via Adaptive Fine-Grained Sample Mining

Chenlin Fu, Yingying Zhu*

College of Computer Science and Software Engineering, Shenzhen University, China
fuchenlin2021@email.szu.edu.cn, zhuyy@szu.edu.cn

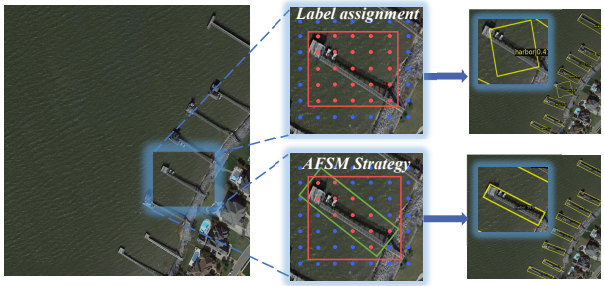
Abstract

Oriented object detection is crucial for complex scenes such as aerial images and industrial inspection, providing precise delineation by minimizing background interference. Recently, the weakly-supervised detector paradigm H2RBox has demonstrated promise in learning rotated bounding box (RBox) from the more readily available horizontal bounding box (HBox), alleviating the scarcity and high cost of RBox annotations. However, these H2RBox-based methods have primarily focused on the gap in orientation information between HBox- and RBox-supervised approaches, overlooking the gap in training sample selection. In response, we propose the Adaptive Fine-grained Sample Mining (AFSM) strategy, which improves the selection of fine-grained training samples in HBox-supervised methods. AFSM assigns the best-matching prediction RBox to each ground truth (GT) HBox and selects positive samples based on these paired boxes. Furthermore, to effectively filter the best-matching prediction RBox for AFSM, we introduce the Prediction Rbox Assignment (PRA) scheme, employing Kullback-Leibler Divergence (KLD) as a localization quality metric. Additionally, we introduce an improved self-supervised branch loss (L_{ss^*}) to address the symmetry of weakly-supervised branch prediction boxes. Incorporating these core components (AFSM, PRA, and L_{ss^*}), we develop an end-to-end network architecture (BGHR) to further bridge the gap between HBox- and RBox-supervised oriented object detection. Extensive experiments on DOTA-v1.0 and DIOR-R demonstrate that BGHR achieves state-of-the-art performance compared to HBox-supervised methods without additional overhead. Even when benchmarked against fully supervised FCOS, our method still exhibits a slight performance advantage.

Code — https://github.com/clin-fu/BGHR.git

Introduction

Object detection has been extensively studied, with early research primarily focused on horizontal detection (Liu et al. 2020; Zhao et al. 2019). However, when fine-grained bounding boxes are required, oriented object detection emerges as the preferred approach, especially in complex scenes such as aerial images, scene text, retail scenes, and industrial inspection (Wen et al. 2023). This fine-grained representation enables the model to more accurately delineate regions of interest by reducing extraneous background interference.



(a) Original image (b) Training sample selection (c) Detection results

Figure 1: (a) is the original image in which the region of interest is locally magnified. (b) illustrates the differences in training sample selection between the H2RBox (Yang et al. 2023) and AFSM. The red, green, and yellow rectangles represent ground truth HBox, best-matching prediction RBox, and detection RBox, respectively, while the red dots represent the selected samples. (c) displays the detection results output by H2RBox (top row) and our BGHR (bottom row).

Although re-annotating labeled datasets allows for the training of oriented detectors, two limitations remain: HBox annotations are more commonly available in existing datasets, and RBox or Mask annotations are more costly to produce. Yang et al. (Yang et al. 2023) explored these limitations and proposed the first general solution called HBox-to-RBox (H2RBox) in 2023. H2RBox gives an effective paradigm and outperforms potential alternatives HBox-Mask-RBox (generating RBox from segmentation mask) such as BoxInst (Tian et al. 2021), SAM (Kirillov et al. 2023) and BoxLevelSet (Li et al. 2022b). Furthermore, EIE (Wang et al. 2024) proposed a bounding box regression method based on the Tanimoto coefficient to address the issue of inconsistency in evaluating rotation variations.

Supervised information plays two main roles in object detection: 1) providing the optimization objective, and 2) selecting training samples. However, previous methods have predominantly focused on the gap in orientation optimization objectives between HBox-supervised and RBox-supervised approaches, neglecting the gap in training sam-

*Corresponding author.

ple selection. As shown in Figure 1, the training samples selected by HBox contain more background noise, which impedes the network's perception of the region of interest and affects the detection performance.

To address this challenge, we propose an Adaptive Fine-grained Sample Mining (AFSM) strategy, which automatically mines fine-grained training samples from GT HBox. The AFSM assigns the best-matching prediction RBox to each GT HBox and selects positive samples based on these paired boxes, with labels derived from the corresponding GT HBox. To identify the best-matching prediction RBox for AFSM, we introduce the Prediction Rbox Assignment (PRA) scheme, utilizing Kullback-Leibler Divergence (KLD) as a metric of localization quality (requiring the conversion of bounding box $B(x_c, y_c, w, h, \theta)$ into a 2D-Gaussian distribution $\mathcal{N}(\mu, \Sigma)$). Based on this, we design two assignment strategies KLD-Only (KO) assignment and Predictions-Mixture (PM) assignment. Furthermore, we propose an improved self-supervised branch loss L_{ss^*} to address the symmetry of the weakly-supervised branch prediction boxes. Altogether, the H2Rbox detection paradigm, AFSM, PRA, and L_{ss^*} constitute our proposed approach, named Bridging the Gap Between HBox-supervised and RBox-supervised Oriented Object Detection via Adaptive Fine-grained Sample Mining (BGHR). Extensive experiments showcase that our BGHR achieves state-of-the-art performance compared to the HBox-supervised approaches without introducing any overhead. Moreover, compared to the RBox-supervised FCOS, BGHR still achieves a slight performance advantage.

Our main contributions can be summarized as follows:

- To the best of my knowledge, our Adaptive Fine-grained Sample Mining (AFSM) is the first to implement a label assignment for HBox-supervised oriented object detection. Compared to previous methods, AFSM is capable of effectively mining fine-grained training samples to narrow the gap with RBox-supervised methods.
- We propose the Prediction Rbox Assignment (PRA) to select the best-matching RBox for AFSM, which includes two assignment schemes: KLD-Only (KO) assignment and Predictions-Mixture (PM) assignment.
- We introduce an improved self-supervised (SS) branch loss L_{ss^*} to tackle the symmetry of weakly-supervised (WS) branch prediction boxes.
- Extensive experiments on DOTA-v1.0 and DIOR-R show that our BGHR outperforms HBox-supervised methods without introducing any computational overhead. It also offers a slight performance advantage over RBox-supervised FCOS.

Related Work

Oriented Object Detection

Oriented object detection can be regarded as fine-grained object detection. Notable approaches in this field include the anchor-based detector Rotated RetinaNet (Lin et al. 2017a), the anchor-free detector Rotated FCOS (Tian et al. 2019), and two-stage detectors such as Oriented R-CNN(Xie et al. 2021), RoI Transformer (Ding et al. 2019), and ReDet (Han et al. 2021). To address the boundary problem caused by the periodicity of angles, (Qian et al. 2021) provides a modulation loss to alleviate loss jumps. CSL (Yang and Yan 2020) and DCL (Yang et al. 2021a) convert the angle into boundary-free coded data. GWD (Yang et al. 2021c), KLD (Yang et al. 2021d), and KFIoU (Yang et al. 2022) propose Gaussian-based losses that convert RBox into a Gaussian distribution. PSC (Yu and Da 2023) proposes a Phase-Shifting Coder that encodes the orientation angle into periodic phases. RepPoint-based approaches (Yang et al. 2019; Hou et al. 2023; Li et al. 2022a) provide new alternatives for oriented object detection by predicting a set of sample points that bounds the spatial extent of an object. In this study, we focus on the more challenging task of using HBox as supervised information for oriented object detection.

HBox-supervised Oriented Object Detection

Utilizing HBox annotations, which are more commonly available, to predict fine-grained RBox can improve the compatibility of oriented object detection with more datasets and expedite its development. HBox-supervised instance segmentation methods (Tian et al. 2021; Li et al. 2022b; Kirillov et al. 2023) employ the HBox-Mask-RBox pipeline to derive RBox from segmentation mask, though this is less cost-effective. A pioneering approach, H2RBox (Yang et al. 2023), bypasses the segmentation step and directly detects RBox from HBox annotations. By leveraging HBox annotations for the same object in various orientations, geometric constraints narrow down the possible angles, making detection more efficient. Furthermore, EIE (Wang et al. 2024) leverages various contrastive cues related to angle prediction, facilitating the learning of equivariance between boxes. It proposes a vectorized bounding box regression method and incorporates the Tanimoto Coefficient to calculate the regression loss after vectorization. These approaches bridge the orientation information gap between HBox-supervised and RBox-supervised methods. Building on these studies, we focus on another key gap: training sample selection between HBox-supervised and RBox-supervised methods.

Label Assignment in Object Detection

In object detection, label assignment is crucial for assigning priors to object proposals, ensuring the accuracy and efficiency of the detection model. Classical anchor-based detectors (Lin et al. 2017a,b; Sun et al. 2021) assign positive samples by calculating the IoU between proposals and Ground Truth (GT). Anchor-free detectors (Tian et al. 2019; Liu et al. 2019) treat grid locations in a fixed region at the center of GT as positive samples. Additionally, ATSS (Zhang et al. 2020) bridges the gap between anchor-based and anchor-free strategies by employing a statistical approach to generate adaptive thresholds for categorizing positive and negative instances. In this work, we develop label assignment strategies for HBox-supervised oriented object detection that adaptively mine fine-grained training samples from GT HBox. To the best of my knowledge, we are the first to conduct relevant research on training sample mining in HBox-supervised oriented object detection.

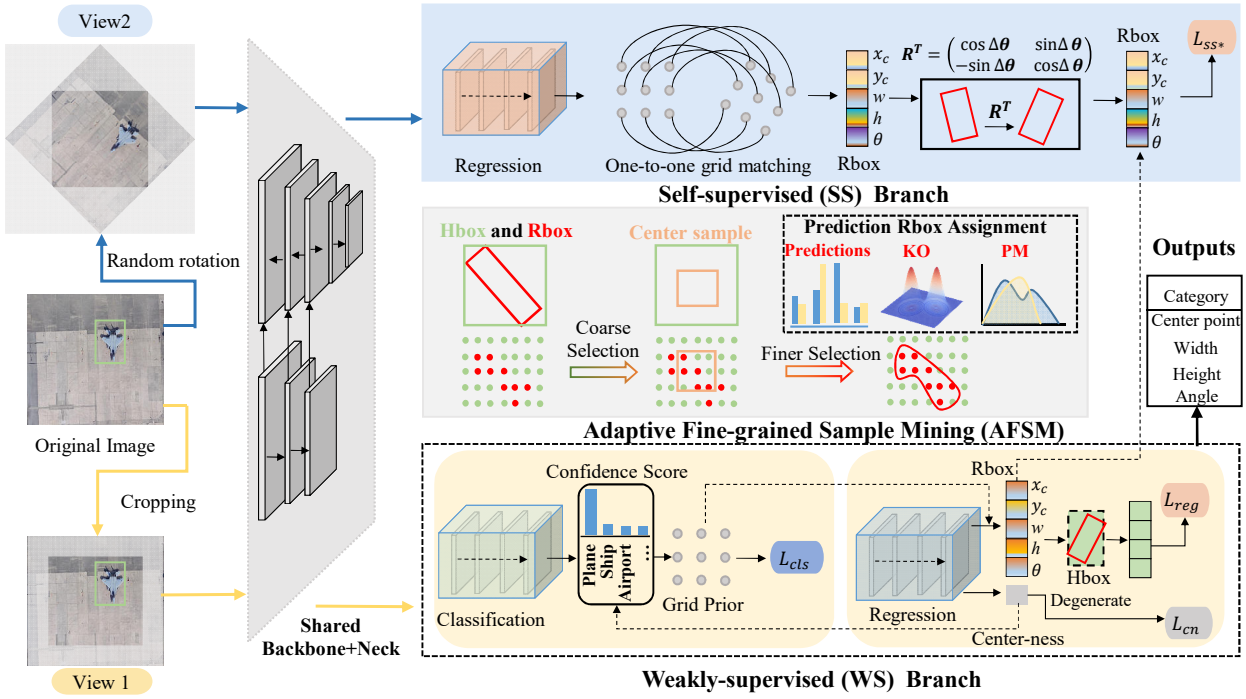


Figure 2: Illustration of the proposed BGHR. The shared backbone+neck were extracted from the two augmented views and then fed into the Weakly-supervised (WS) Branch and Self-supervised (SS) Branch, respectively. Adaptive Fine-grained Sample Mining (AFSM) adaptively mines fine-grained training samples for the network throughout the training process. The Prediction RBox Assignment (PRA) in AFSM is responsible for selecting the best-matching prediction RBox for GT.

Methodology

Overview

The overall network architecture of BGHR is illustrated in Figure 2. To avoid information leakage and prevent overfitting during training, two augmented views are generated. View1 is created by center region cropping (cropping a $\frac{\sqrt{2}}{2}s \times \frac{\sqrt{2}}{2}s$ area in the center of the image). View2 is generated by applying random rotation and reflection padding (filling the black border area with reflection padding) while ensuring that view2 is cropped the same way as view1.

As shown in Figure 2, our BGHR initially leverages a shared backbone and neck to extract multiscale features from two augmented views. Subsequently, it incorporates two distinct branches dedicated to category prediction and bounding box regression. The weakly-supervised (WS) branch employs an FCOS-based oriented object detector for both training and inference. This branch includes regression and classification sub-networks to predict the RBox, category, and center-ness. The regression loss is calculated between the circumscribed HBox derived from the predicted RBox and the ground truth (GT) HBox.

Complementary to the WS branch, the other branch is trained using self-supervised (SS) learning. This involves two augmented views of the raw input image, promoting consistent RBox predictions between the views. The SS branch contains only one regression sub-network for predicting the RBox in the rotated view2. To calculate the consistency between the predictions of the two branches, we assign the $rbox^{ws}(x_{ws}, y_{ws}, w_{ws}, h_{ws}, \theta ws)$ predicted by the WS branch as the target RBox for the SS branch. Referring to H2RBox, we use one-to-one grid matching to align the grid priors of the two branches. Specifically, the relationship between the grid location (x^*, y^*) of the SS branch and the grid location (x, y) of the WS branch is:

$$(x^*, y^*) = (x - x_c, y - y_c)\mathbf{R}^\top + (x_c, y_c) \tag{1}$$

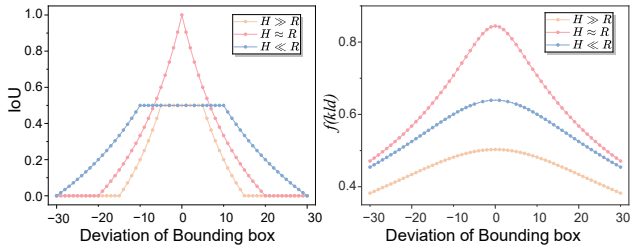
where (x_c, y_c) is the rotation center (i.e. image center). $\mathbf{R}$ represent random rotation transformation with degree $\Delta\theta$ as adopted in View2.

$$\mathbf{R} = \begin{pmatrix} cos\Delta\theta & -sin\Delta\theta \\ sin\Delta\theta & cos\Delta\theta \end{pmatrix} \tag{2}$$

The network structure integrates two augmented views and leverages predictive consistency between the branches, thereby incorporating the missing object orientation information in the HBox-supervised approach. To bridge the sample quality gap, AFSM aids in mining fine-grained samples by assigning an optimal predicted RBox to each GT HBox during training. This allocation is achieved through the PRA scheme. Guided by the optimization function, AFSM adaptively provides high-quality training samples for both branches.

Adaptive Fine-grained Sample Mining Strategy

The proposed AFSM first uses the PRA scheme to assign the best-matching prediction RBox to each GT HBox. Next,



(a) IoU-based (b) KLD-based

Figure 3: (a) and (b) show the deviation curves between HBox and RBox. $H \gg R$, $H \approx R$, and $H \ll R$ indicate that the scale of HBox is much larger than RBox, comparable to RBox, and much smaller than RBox, respectively. $f(\cdot)$ denotes the application of Equation (9) to the KLD.

it uses these paired boxes to select positive samples, with their labels determined by the corresponding GT HBox. our detailed strategy is described below for any pair, GT HBox B^g and prediction RBox B^p.

For a given feature map $\mathbf{F} \in \mathbb{R}^{H\times W\times C}$, where H, W, and C correspond to the height, width, and channel of the feature map, respectively, the set of grid locations for this feature map is represented as $P = \{(x_i, y_i)|i = 1, 2, 3, ..., H \times W\}$. The correspondence between any point (x_i, y_i) in the set P and the original image position (x_i^a, y_i^a) is as follows:

$$(x_i^a, y_i^a) = (\left\lfloor \frac{s}{2} \right\rfloor + x_i \cdot s, \left\lfloor \frac{s}{2} \right\rfloor + y_i \cdot s) \tag{3}$$

where s denote the stride of feature map $\mathbf{F}$. Then calculate the distance $D_i^g(l_i^g, r_i^g, t_i^g, b_i^g)$ and $D_i^p(l_i^p, r_i^p, t_i^p, b_i^p)$ from (x_i^a, y_i^a) to each of the four edges of the boxes, B^g and B^p, where l, r, t, b represent the distances to the left, right, top, and bottom sides of the boxes, respectively. The calculation formula are as follows:

$$\begin{aligned} l &= (x_i^a - x^*)cos\theta + (y_i^a - y^*)sin\theta + \frac{w}{2}, \\ r &= (x_i^a - x^*)cos\theta + (y_i^a - y^*)sin\theta - \frac{w}{2}, \\ t &= -(x_i^a - x^*)sin\theta + (y_i^a - y^*)cos\theta + \frac{h}{2}, \\ b &= -(x_i^a - x^*)sin\theta + (y_i^a - y^*)cos\theta - \frac{h}{2} \end{aligned} \tag{4}$$

where (x^*, y^*) represent the center and w, h and θ are the width, height, and angle of the boxes, respectively.

Since GT HBox and GT RBox are co-centred, samples in the central region of the GT HBox also belong to the GT RBox. Thus, we perform center sampling on the GT HBox B^g with a sampling radius of $s\times o$, where s denotes the stride and o denotes the sampling ratio. The sampling result is denoted as B^c. The distance $D_i^c(l_i^c, r_i^c, t_i^c, b_i^c)$ from (x_i^a, y_i^a) to each edge of B^c is also calculated using the aforementioned formula. Subsequently we obtain the set P^g, P^c and P^p belonging to B^g , B^c and B^p in the following way:

$$\begin{aligned} P^g &= \{(x_i, y_i)|i \in \{i|l_i^g > 0, r_i^g > 0, t_i^g > 0, b_i^g > 0\}\}, \\ P^c &= \{(x_i, y_i)|i \in \{i|l_i^c > 0, r_i^c > 0, t_i^c > 0, b_i^c > 0\}\}, \\ P^p &= \{(x_i, y_i)|i \in \{i|l_i^p > 0, r_i^p > 0, t_i^p > 0, b_i^p > 0\}\} \end{aligned} \tag{5}$$

The final positive samples grid locations P^+ are obtained using the following formula:

$$P^+ = (P^g \cap P^c) \cup (P^g \cap P^p) \tag{6}$$

where $P^g \cap P^c$ presents coarse selection process, center sampling allows the selection of a small number of fine-grained samples, removing background noise and ensuring normal training. $P^g \cap P^p$ is further supplemented with fine-grained samples using the prediction RBox. The logical AND operation is used to prevent the introduction of noise from inaccurate prediction RBox during the initial phase of training.

As the network is better trained, an indirect connection forms between prediction RBox and high-quality training samples: The fine-grained samples selected by AFSM enhance the network's accuracy in predicting RBox, which in turn helps the AFSM select even higher-quality samples.

Prediction Rbox Assignment Scheme

The primary function of PRA is to select the best-matching prediction RBox B^p for GT HBox B^g. Typically, B^p includes localization information (measured by IoU) and category information (measured by confidence). Given that RBox is usually contained within HBox, there are scale differences between boxes in the HBox-supervised method.

As shown in Figure 3(a), IoU-based deviation curves remain unchanged in some intervals when scale differences are large ($H \gg R$, $H \ll R$). Conversely, as illustrated in Figure 3(b), the KLD-based deviation curve changes uniformly and smoothly. Hence, the KLD-based metric is used for PRA. Specifically, we first convert a Box $B(x_c, y_c, w, h, \theta)$ into a 2-D Gaussian $\mathcal{N}(\mu, \Sigma)$ by the following equation:

$$\begin{aligned} \mu &= (x_c, y_c)^\top, \\ \Sigma^{1/2} &= \begin{pmatrix} \frac{w}{2}\cos^2\theta + \frac{h}{2}\sin^2\theta & \frac{w-h}{2}\cos\theta\sin\theta \\ \frac{w-h}{2}\cos\theta\sin\theta & \frac{w}{2}\sin^2\theta + \frac{h}{2}\cos^2\theta \end{pmatrix} \end{aligned} \tag{7}$$

The KLD between two probability measures $\mathcal{N}_p(\mu_p, \Sigma_p)$ and $\mathcal{N}_g(\mu_g, \Sigma_g)$ expressed as:

$$\begin{aligned} D_{kl}(\mathcal{N}_p\|\mathcal{N}_g) = \frac{1}{2}(\mu_p - \mu_g)^\top\Sigma_g^{-1}(\mu_p - \mu_g) + \\ \frac{1}{2}\mathrm{Tr}(\Sigma_g^{-1}\Sigma_p) + \frac{1}{2}\ln|\Sigma_g| - \frac{1}{2}\ln|\Sigma_p| - 1 \end{aligned} \tag{8}$$

We then propose two assignment schemes:

1) KLD-Only (KO) assignment: The match score M^{ko} between B^g and the prediction RBox B^p for all grid locations is calculated by the following formula:

$$M^{ko} = \frac{1}{D_{kl}(\mathcal{N}_p\|\mathcal{N}_g) + \tau}, \tau \geq 1 \tag{9}$$

where the hyperparameter τ is configured according to the optimal setting from KLD(Yang et al. 2021d), with a default value of 1.

2) Predictions-Mixture (PM) assignment: We introduce the confidence Z of prediction RBox that corresponds to B^g label, which jointly modulates the overall match score M^{pm}:

$$M^{pm} = Z^{\alpha} \times \left(\frac{1}{D_{kl}(\mathcal{N}_p \| \mathcal{N}_g) + 1} \right)^{\beta} \tag{10}$$

where α and β are control parameters, balancing the effects of localization and classification. The prediction RBox with the highest matching score was selected as the best match.

The Overall Loss

Loss for WS Branch. Since the WS branch structure is based on FCOS, the losses in this part mainly include the regression L_{reg}, classification L_{cls}, and center-ness L_{cn}. The loss function for the WS branch, L_{ws}, is defined as follows:

$$\begin{aligned} L_{ws} = & \frac{\mu_1}{N_{pos}} \sum_i L_{cls}(c_i^*, c_i) + \frac{\mu_2}{N_{pos}} \sum_i L_{cn}(cn_i^*, cn_i) \\ & + \frac{\mu_3}{\sum cn_{pos}} \sum_{(i)} \mathbb{I}_{\{c_i > 0\}} cn_i L_{reg} \left\{ r2h(B_i^{ws}), B_i^g \right\} \end{aligned} \tag{11}$$

where L_{cls} is the focal loss (Lin et al. 2017b), L_{cn} is cross-entropy loss, and L_{reg} is IoU loss. N_{pos} denotes the number of positive samples. c^* and c denote the probability distribution of various classes calculated by Sigmoid function and target category. B^{ws} and B^g represent the predicted RBox in the WS branch and the GT HBox, respectively. cn_i^* and cn_i indicate the predicted and target center-ness. $\mathbb{I}_{\{c_i > 0\}}$ is the indicator function, being 1 if $c_i > 0$ and 0 otherwise. The $r2h(\cdot)$ function converts the RBox to its corresponding horizontal circumscribed rectangle. We set the hyperparameters $\mu_1 = 1$, $\mu_2 = 1$ and $\mu_3 = 1$, the same as H2RBox.

Loss for SS Branch. We apply the rotation transformation to B^{ws} to obtain B^{ws^*} and then compute the SS loss L_{ss^*} between $B^{ws^*}(x_c^*, y_c^*, w^*, h^*, \theta^*)$ and the RBox $B^{ss}(x_c^{ss}, y_c^{ss}, w^{ss}, h^{ss}, \theta^{ss})$ predicted by SS branch:

$$L_{ss^*} = \frac{1}{\sum cn_{pos}^*} \sum_{i^*} \mathbb{I}_{\{c_{i^*} > 0\}} cn_{i^*} L_{reg}^*(B_{i^*}^{ws^*}, B_{i^*}^{ss}) \tag{12}$$

There are two possible prediction RBox for the WS branch, which are symmetric about the x-axis. Therefore we design the regression loss L_{reg}^* of SS branch as:

$$\begin{aligned} Loss_1 &= L_{iou}(H_{ws}, H_{ss}^1) + L_1(\sin(\theta^* - \theta^{ss}), 0), \\ Loss_2 &= L_{iou}(H_{ws}, H_{ss}^2) + L_1(\sin(\theta^* + \theta^{ss}), 0), \\ L_{reg}^*(B^{ws^*}, B^{ss}) &= \min\{Loss_1, Loss_2\} \end{aligned} \tag{13}$$

where $H_{ws}(-w^*, -h^*, w^*, h^*)$ denotes the projection of the width and height of B^{ws^*} in the horizontal direction. Considering the loss discontinuity caused by boundary issues, $H_{ss}^1(-w^{ss}, -h^{ss}, w^{ss}, h^{ss})$ and $H_{ss}^2(-h^{ss}, -w^{ss}, h^{ss}, w^{ss})$ denote the two projections of the width and height of B^{ss} in the horizontal direction.

The overall loss L_{total} is calculated as a sum of the WS loss L_{ws} and the SS loss L_{ss^*} by the following equation:

$$L_{total} = L_{ws} + \lambda L_{ss^*} \tag{14}$$

where λ set to 0.4 by default, the same as H2RBox.

Experiment

Datasets and Experimental Settings

Datasets. We select two datasets, DOTA-v1.0 (Xia et al. 2018) and DIOR-R (Cheng et al. 2022), as they provide concurrent annotations in both HBox and RBox formats.

Evaluation Metrics. This paper employs AP_{50} as well as the stricter metrics AP_{75} and AP. The default AP refers to $AP_{50:95}$ in object detection community.

Implementation Details. All models employ ResNet50 (He et al. 2016), trained on Tesla A100 GPUs with AdamW (Loshchilov and Hutter 2017) and a mini-batch size of 2.

Comparison with State-of-the-art Methods

Results on DOTA-v1.0. Table 1 and Table 2 present a comprehensive comparison of BGHR with existing methods on DOTA-v1.0. These methods include HBox- and RBox-supervised detectors. Compared to HBox-supervised methods, BGHR outperforms the HBox-Mask-RBox optimal SAM-based detector by 7.5%. When benchmarked against the baseline H2RBox, BGHR achieves a 3.62% higher AP_{50} score than H2RBox. When multi-scale (MS) training and testing is applied, the AP_{50} score reaches 77.65%, surpassing the performance of other HBox-supervised models. When more rigorous metrics AP_{75} and AP are used, our approach consistently outperforms H2RBox by 4.38% and 3.12% in AP_{75} and AP, respectively. With MS applied, the improvements increase to 5.35% and 3.99%.

As a pioneer of HBox-supervised detectors, H2RBox achieves competitive results (AP_{50} score is 2.96% below FCOS) with RBox-supervised methods, relying solely on horizontal supervision. Building on this research, we analyze the impact of supervised information on detection performance, improving it by mining fine-grained training samples. Our experimental results indicate significant improvements across all metrics with our method. Encouragingly, our BGHR is also 0.66% AP_{50} score higher than the RBox-supervised FCOS, which reinforces the fact that our BGHR can further bridge the gap between HBox- and RBox-supervised methods.

Results on DIOR-R. To assess the robustness of our BGHR, we also compared BGHR against state-of-the-art methods using the DIOR-R dataset, as detailed in Table 3. Our BGHR achieves optimal performance compared to other HBox-supervision methods, with AP_{50}, AP_{75}, and AP scores of 59.20%, 33.00%, and 34.11% respectively. In the default setup, our BGHR outperforms baseline H2RBox by 2.20% (59.20% vs. 57.00%). Additionally, compared to RBox-supervised methods like KLD (Yang et al. 2021d) and FCOS, our approach still achieves competitive results.

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	AP_{50}
RBox-supervised:																
RepPoints	86.67	**81.12**	41.62	62.04	76.23	56.32	75.71	90.66	80.80	85.29	63.29	66.64	59.13	67.57	33.67	68.45
RetinaNet	87.87	77.34	39.67	61.38	75.85	54.41	75.56	90.84	77.36	79.70	51.78	61.50	50.77	65.08	35.33	65.63
PSC	89.60	73.61	49.03	62.34	75.18	77.69	88.00	90.85	82.63	72.97	61.48	64.20	65.77	72.88	67.88	72.90
GWD	88.53	74.04	46.39	60.52	80.22	76.91	87.44	90.88	81.98	84.53	55.51	62.11	63.93	70.67	50.94	71.64
KLD	88.74	77.09	47.17	57.18	79.92	77.99	87.46	**90.90**	83.50	84.44	56.32	65.53	64.32	69.67	57.45	72.51
KFIoU	**89.05**	75.17	49.04	69.73	78.06	75.46	86.69	**90.90**	83.65	84.48	62.21	62.87	66.72	65.95	50.20	72.68
FCOS	88.41	75.61	47.98	60.10	79.78	77.81	86.64	90.08	78.23	84.95	52.80	66.25	64.45	68.28	40.31	70.78
HBox-supervised:																
BoxInst-RBox (960)	68.43	40.75	33.07	32.29	46.91	55.43	56.55	79.49	66.81	82.14	41.24	52.83	52.80	65.04	29.99	53.59
BoxLevelSet-RBox (960)	63.48	71.27	39.34	61.06	41.89	41.03	45.83	90.87	74.12	72.13	47.59	62.99	50.00	56.42	28.63	56.44
SAM	78.63	69.15	31.39	56.68	72.23	71.42	77.02	90.53	76.17	83.65	42.46	59.52	51.18	56.24	42.88	63.94
EIE	87.66	70.15	41.50	60.47	80.70	76.25	86.25	90.87	82.63	84.70	53.14	64.51	58.07	70.41	43.83	70.08
H2RBox (baseline)	88.47	73.51	40.81	56.89	77.48	65.42	77.87	90.88	83.19	85.27	55.27	62.90	52.41	63.63	43.26	67.82
BGHR (Ours) (960)	88.44	73.76	42.06	62.35	79.56	74.15	79.31	90.74	82.40	85.71	57.26	66.63	61.08	71.29	54.86	71.31
BGHR (Ours)	88.21	75.65	42.83	60.17	79.39	74.99	85.76	90.86	83.94	85.34	56.77	65.90	60.18	72.45	49.11	71.44
BGHR (Ours) (MS)	88.81	81.01	**53.43**	**72.16**	**81.73**	**83.92**	**88.11**	90.88	**86.71**	**88.80**	**66.09**	**68.62**	**66.80**	**81.05**	**67.67**	**77.65**

Table 1: Results of box the default AP_{50} for each category on the DOTA-v1.0. MS denotes multi-scale. '960' indicates the input images are resized to 960 × 960 during both training and inference, default is 1024 × 1024.

Method	AP	AP_{50}	AP_{75}
RBox-supervised:			
RepPoints (Yang et al. 2019)	37.56	68.45	37.23
RetinaNet (Lin et al. 2017a)	37.90	65.63	37.66
PSC (Yu and Da 2023)	41.51	72.90	39.80
GWD (Yang et al. 2021c)	41.82	71.64	40.95
KLD (Yang et al. 2021d)	40.86	72.51	39.56
KFIoU (Yang et al. 2022)	39.92	72.68	37.85
R^3Det (Yang et al. 2021b)	37.57	68.75	37.23
FCOS (Tian et al. 2019)	39.34	70.78	36.91
HBox-supervised:			
BoxInst-RBox (Tian et al. 2021) (960)	-	53.59	-
BoxLevelSet-RBox (Li et al. 2022b) (960)	-	56.44	-
SAM (Kirillov et al. 2023)	30.16	63.94	22.33
EIE (Wang et al. 2024)	38.54	70.08	36.35
EIE (Wang et al. 2024) (MS)	45.07	75.74	45.52
H2RBox (baseline) (Yang et al. 2023) (960)	36.39	67.51	34.11
H2RBox (baseline) (Yang et al. 2023)	36.63	67.82	33.67
H2RBox (baseline) (Yang et al. 2023) (MS)	43.13	74.53	42.49
BGHR (FCOS-based) (Ours) (960)	39.59	71.31	37.69
BGHR (FCOS-based) (Ours)	39.75	71.44	38.05
BGHR (FCOS-based) (Ours) (MS)	**47.12**	**77.65**	**47.84**

Table 2: Comparison of performance on the DOTA-v1.0 with more rigorous metrics.

Method	AP	AP_{50}	AP_{75}
RBox-supervised:			
RetinaNet (Lin et al. 2017a)	33.47	54.60	33.80
KLD (Yang et al. 2021d)	**35.77**	58.00	**37.00**
FCOS (Tian et al. 2019)	34.16	58.60	31.90
HBox-supervised:			
BoxInst-RBox (Tian et al. 2021)	29.96	56.65	24.36
BoxLevelSet-RBox (Li et al. 2022b)	31.73	57.40	28.10
H2RBox (baseline) (Yang et al. 2023)	33.15	57.00	32.60
EIE (Wang et al. 2024)	33.73	58.50	33.00
BGHR (FCOS-based) (ours)	34.11	59.20	33.30
BGHR (FCOS-based) (ours) (1024)	34.61	**60.70**	33.60

Table 3: Comparison of performance on DIOR-R. '1024' indicates the input images are resized to 1024 × 1024, default is 800 × 800.

Method	GFLOPs	Params(MB)	FPS	AP_{50}
H2RBox(baseline)	206.91	31.92	**27.2**	67.82
BGHR(ours)	206.91	31.92	26.6	**71.44**

Table 4: Results of computational cost and detection speed.

Computational Cost and Speed. Table 4 shows that our method incurs no additional computational overhead or difference in detection speed compared to the baseline.

Ablation Study

Effect of Components. Table 5 presents the ablation study of the components. Here, AFSM and L_{ss^*} represent our proposed mining strategy and loss of SS branch, respectively. L_{ws} and L_{ss} are combined to represent the performance of the baseline model. The AP_{75} score improved by 2.85% and 1.97% when adding AFSM and Lss^*, respectively, to the baseline. When all the components work together, the score improvement reaches 4.38%. These excellent performances are attributed to the fine-grained training samples mined by AFSM and the effectiveness of the L_{ss^*} in addressing the symmetry of WS branch prediction boxes.

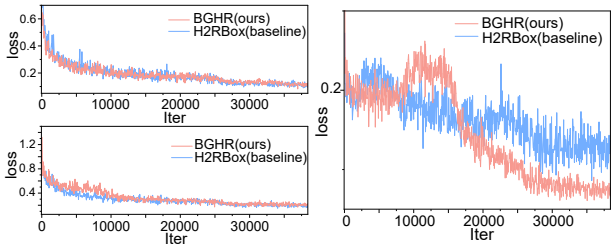
Effect of Hyper-parameters. We investigate the performance of PRA using different values of α and β in Equation (10), which regulate the impact of confidence and localization quality on assignment RBox. Through a coarse search shown in Table 6, we adopt $\alpha = 1$ and $\beta = 2$. The experimental results above indicate that PRA achieves optimal detection by balancing localization quality and confidence. This success is because the RBox selected by the mixed metrics exhibits a degree of categorical regression consistency, ensuring the quality of the best-matching RBox.

AFSM	L_{ws}	L_{ss}	L_{ss^*}	AP	AP_{50}	AP_{75}
	✓			12.63	37.13	7.54
	✓	✓		36.63	67.82	33.67
	✓		✓	38.20	69.63	35.64
✓	✓	✓		38.61	69.82	36.52
✓	✓		✓	**39.75**	**71.44**	**38.05**

Table 5: Ablation for each component.

α	β	AP	AP_{50}	AP_{75}
0.5	1	38.47	70.94	35.13
0.5	2	24.83	57.24	17.55
1.0	1	30.79	61.68	25.65
1.0	2	**39.75**	71.44	**38.05**
1.5	1	24.01	57.03	16.71
1.5	2	39.37	**71.69**	37.01

Table 6: Ablation for hyper-parameters of Equation (10).



(a) WS branch (b) SS branch

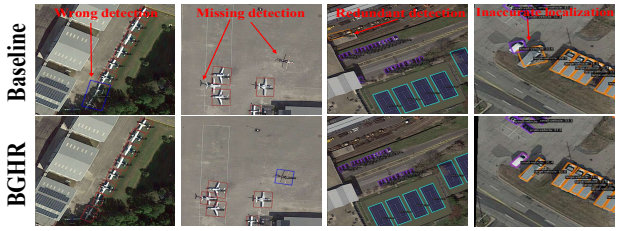
Figure 4: Comparison of the loss convergence curves for our BGHR (red) and H2RBox (blue).

Dataset	PRA		AP	AP_{50}	AP_{75}
	KO	PM			
DOTA-V1.0	✓		38.69	**71.56**	35.96
		✓	**39.75**	71.44	**38.05**
DIOR-R	✓		34.10	59.40	32.80
		✓	**34.61**	**60.70**	**33.60**

Table 7: Ablation study for KO and PM assignment strategy.

Radius	AP	AP_{50}	AP_{75}
-	24.19	57.62	15.88
1.40	38.38	70.90	35.61
1.45	**39.75**	**71.44**	**38.05**
1.50	39.04	70.59	37.50
2.00	38.06	69.92	34.47

Table 8: Analysis of different center sampling radii.



(a) (b) (c) (d)

Figure 5: Comparison of detection results with baseline.

Effect of Assignment Strategy. Table 7 presents the ablation experiments of the two assignment strategies of our PRA on different datasets. PM outperforms KO except for a slightly lower AP_{50} score on the DOTA-v1.0 dataset. Intuitively, the localization quality of RBox is directly related to the label assignment of the FCOS-based detector. However, experimental results indicate that the PM strategy, incorporating a confidence mixing factor, demonstrates better overall performance. This suggests that classification confidence can influence the quality of the prediction RBox.

Effect of Center Sampling Radius. Table 8 studies the impact of different center sampling radii for AFSM. Center sampling is introduced to select samples near the geometric center of the object, providing the network with a small number of fine-grained samples when the prediction RBox is inaccurate. Our method achieves optimal results with a radius of 1.45. Compared to not using center sampling, our optimal parameter settings improve the AP, AP_{50} and AP_{75} scores by 19.56%, 13.82%, and 22.17%, respectively. The experimental results show that center sampling can significantly improve detection performance.

Convergence Analysis. To examine the mutual influence between our L_{ss^*} and L_{ws} during model training, we compare the convergence curves of H2RBox and our BGHR, as shown in Figure 4. It can be seen that our improved SS branch loss can converge to the minimum value (as indicated in Figure 4(b)) and has no effect on L_{ws} (as indicated in Figure 4(a)). This indicates that our improved SS branch loss can achieve the lowest loss value and enhance the prediction ability without affecting the convergence of WS branch loss.

Qualitative Comparison

We present a qualitative comparison between our BGHR and the baseline H2RBox in Figure 5. Our BGHR mitigates the occurrence of wrong detection and false negatives/positives (as depicted in Figure 5(a), (b), (c)) and achieves improved alignment with oriented foreground objects (as shown in Figure 5(d)). This observation indicates that mined fine-grained training samples can enhance detector performance.

Conclusion and Limitation

In this paper, we introduce a novel and efficient HBox-supervised oriented object detector aimed at further bridging the gap between HBox- and RBox-supervised methods. Our AFSM can adaptively mine the fine-grained training samples for HBox-supervised oriented detector by combining GT HBox and prediction RBox. Moreover, we propose an improved SS branch loss to address the symmetry of WS branch prediction boxes. Extensive experiments demonstrate the effectiveness of our approach. However, our approach still uses the GT HBox to compute the center-ness target. Fine-grained center-ness might further improve the model performance, and we leave that as future work.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62072318, in part by the Key Project of Department of Education of Guangdong Province under Grant 2023ZDZX1016, and in part by Shenzhen Science and Technology Program under Grant 20220810142553001.

References

Cheng, G.; Wang, J.; Li, K.; Xie, X.; Lang, C.; Yao, Y.; and Han, J. 2022. Anchor-free oriented proposal generator for object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–11.

Ding, J.; Xue, N.; Long, Y.; Xia, G.-S.; and Lu, Q. 2019. Learning roi transformer for oriented object detection in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2849–2858.

Han, J.; Ding, J.; Xue, N.; and Xia, G.-S. 2021. Redet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2786–2795.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hou, L.; Lu, K.; Yang, X.; Li, Y.; and Xue, J. 2023. G-rep: Gaussian representation for arbitrary-oriented object detection. *Remote Sensing*, 15(3): 757.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Li, W.; Chen, Y.; Hu, K.; and Zhu, J. 2022a. Oriented reppoints for aerial object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1829–1838.

Li, W.; Liu, W.; Zhu, J.; Cui, M.; Hua, X.-S.; and Zhang, L. 2022b. Box-supervised instance segmentation with level set evolution. In *European conference on computer vision*.

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

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

Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; and Pietikäinen, M. 2020. Deep learning for generic object detection: A survey. *International journal of computer vision*, 128: 261–318.

Liu, W.; Liao, S.; Ren, W.; Hu, W.; and Yu, Y. 2019. High-level semantic feature detection: A new perspective for pedestrian detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5187–5196.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Qian, W.; Yang, X.; Peng, S.; Yan, J.; and Guo, Y. 2021. Learning modulated loss for rotated object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 2458–2466.

Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.; Wang, C.; et al. 2021. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14454–14463.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9627–9636.

Tian, Z.; Shen, C.; Wang, X.; and Chen, H. 2021. Boxinst: High-performance instance segmentation with box annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5443–5452.

Wang, L.; Zhan, Y.; Lin, X.; Yu, B.; Ding, L.; Zhu, J.; and Tao, D. 2024. Explicit and Implicit Box Equivariance Learning for Weakly-Supervised Rotated Object Detection. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

Wen, L.; Cheng, Y.; Fang, Y.; and Li, X. 2023. A comprehensive survey of oriented object detection in remote sensing images. *Expert Systems with Applications*, 224: 119960.

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

Xie, X.; Cheng, G.; Wang, J.; Yao, X.; and Han, J. 2021. Oriented R-CNN for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3520–3529.

Yang, X.; Hou, L.; Zhou, Y.; Wang, W.; and Yan, J. 2021a. Dense label encoding for boundary discontinuity free rotation detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15819–15829.

Yang, X.; and Yan, J. 2020. 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*, 677–694. Springer.

Yang, X.; Yan, J.; Feng, Z.; and He, T. 2021b. R3det: Refined single-stage detector with feature refinement for rotating object. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 3163–3171.

Yang, X.; Yan, J.; Ming, Q.; Wang, W.; Zhang, X.; and Tian, Q. 2021c. Rethinking rotated object detection with gaussian wasserstein distance loss. In *International Conference on Machine Learning*, 11830–11841. PMLR.

Yang, X.; Yang, X.; Yang, J.; Ming, Q.; Wang, W.; Tian, Q.; and Yan, J. 2021d. Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. *Advances in Neural Information Processing Systems*, 34: 18381–18394.

Yang, X.; Zhang, G.; Li, W.; Wang, X.; Zhou, Y.; and Yan, J. 2023. H2rbox: Horizontal box annotation is all you need for oriented object detection. *International Conference on Learning Representations*.

Yang, X.; Zhou, Y.; Zhang, G.; Yang, J.; Wang, W.; Yan, J.; Zhang, X.; and Tian, Q. 2022. The KFIoU loss for rotated object detection. *arXiv preprint arXiv:2201.12558*.

Yang, Z.; Liu, S.; Hu, H.; Wang, L.; and Lin, S. 2019. Reppoints: Point set representation for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9657–9666.

Yu, Y.; and Da, F. 2023. Phase-shifting coder: Predicting accurate orientation in oriented object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13354–13363.

Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; and Li, S. Z. 2020. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9759–9768.

Zhao, Z.-Q.; Zheng, P.; Xu, S.-t.; and Wu, X. 2019. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11): 3212–3232.