Multi-clue Consistency Learning to Bridge Gaps Between General and Oriented Object in Semi-supervised Detection

Chenxu Wang[1], Chunyan Xu[1,*], Xiang Li[2,*], Yuxuan Li[2], Xu Guo[1], Ziqi Gu[1], Zhen Cui[1]

[1]Nanjing University of Science and Technology, Nanjing, China.
[2] Nankai University, Tianjin, China.
facias914@gmail.com, xiang.li.implus@nankai.edu.cn, zcablii@ucl.ac.uk
{cyx, xu.guo, ziqigu, zhen.cui}@njust.edu.cn

Abstract

While existing semi-supervised object detection methods perform well in general scenes, they struggle with oriented objects. Our experiments reveal two inconsistency issues that stem from the gaps between general and oriented object detection in semi-supervised learning: 1) Assignment inconsistency: First, the common label assignment is inadequate for oriented objects with larger aspect ratios when selecting positive labels from labeled data. Second, balancing the precision and localization quality of oriented pseudo-boxes presents greater challenges, introducing more noise when selecting positive labels from unlabeled data. 2) Confidence inconsistency: there exists more mismatch between the predicted classification and localization qualities when considering oriented objects, affecting the selection of pseudo-labels. Therefore, we propose a Multi-clue Consistency Learning (MCL) framework to bridge the gaps between general and oriented objects in semi-supervised detection. Specifically, Gaussian Center Assignment is designed to select shape-aware positive labels from labeled data for oriented objects with larger aspect ratios, while Scale-aware Label Assignment is introduced to select scale-aware pixel-level pseudo-labels instead of unreliable pseudo-boxes from unlabeled data. The Consistent Confidence Soft Label is adopted to further boost the detector by maintaining the alignment of the predicted confidences. Experiments on DOTA-v1.5 and DOTA-v1.0 benchmarks demonstrate that the MCL can achieve state-of-the-art performance in the semi-supervised oriented object detection.

Code — https://github.com/facias914/sood-mcl

Introduction

Labeling adequate accurate annotations is a critical factor in enhancing the performance of object detection algorithms, but it is also labour-consuming, especially for oriented objects. To reduce the annotation burden, the semi-supervised object detection (SSOD) attempts to leverage both labeled data and unlabeled data to boost the detector performance. Existing advanced SSOD methods primarily adopt the self-training framework (Liu et al. 2021; Liu, Ma, and Kira 2022; Wang et al. 2023) to exploit unlabeled data and achieve significant performance in general scenes. However, they encounter various challenges in handling semi-supervised oriented object detection (SOOD). What makes the performance gap between general and oriented object detection when performing semi-supervised learning? In this work, we delve into this issue and identify the underlying causes. With comprehensive investigations, we find that the performance of SOOD is significantly hindered by the assignment and confidence inconsistency problems that stem from the gaps between general and oriented object detection.

The assignment inconsistency is evident in two key aspects. First, for labeled data training branch, the positive labels selected from labeled data using general label assignment strategies are inconsistent with the supervision information required by oriented objects with larger aspect ratios. This inconsistency causes the discriminative information at the edges of oriented objects to be misclassified as background, affecting the optimization of the detector. Furthermore, the sub-optimal label assignment strategies lead to insufficient utilization of valuable labeled data, impairing the performance of semi-supervised learning. Second, for unlabeled data training branch, the oriented pseudo-boxes predicted by teacher model from unlabeled data are difficult to guide the semi-supervised learning process, resulting in noisy label assignment. In general, most advanced SSOD methods (Wang et al. 2023; Zhang, Sun, and Wei 2023) follow the pseudo-boxes paradigm (Sohn et al. 2020) that apply the pseudo-boxes predicted by the teacher model as supervised signals to optimize the student model. However, the introduction of angle parameters in rotated pseudo-boxes makes their localization quality less controllable than that of horizontal ones. Under the guidance of low-quality rotated pseudo-boxes, noisy labels are inevitably introduced during the label assignment process. Moreover, even with the aid of NMS (Non-Maximum Suppression) and threshold filtering, it is impossible to guarantee there are not redundant boxes which will bring additional noisy labels.

On the other hand, the confidence inconsistency refers to the mismatch between the predicted classification and localization qualities, affecting the selection of pseudo-labels. Prior researches (Sun et al. 2021; Xu et al. 2021; Li et al. 2022a) have validated that the classification score cannot effectively measure the localization quality, and the lack of interaction between the two confidences causes an inconsistency in predictions, leading to sub-optimal pseudo-label se-

*Corresponding authors.

lection. Furthermore, while numerous researches have studied the proxy localization quality of horizontal boxes, there is a lack of research for oriented boxes. Our experiments indicate that there remains a gap in the proxy representation of localization quality between horizontal and oriented boxes, and sub-optimal proxy localization quality also negatively impacts the performance of oriented object detection.

Based on these analysis, we propose a simple yet effective framework named Multi-clue Consistency Learning (MCL) to enhance the performance of SOOD by alleviating the inconsistency. Concretely, to tackle assignment inconsistency, we first introduce the Gaussian Center Assignment (GCA) strategy to select more accurate positive labels from the limited annotated data. It assigns positive labels based on the $2D$ Gaussian Distribution modeled according to the ground truth boxes, considering the aspect ratios of oriented objects to effectively mine credible positive supervision signals. Second, we propose the Scale-aware Label Assignment (SLA) method to select pixel-level pseudo-labels from unlabeled data instead of unreliable pseudo-boxes. It adopts a divide-and-rule strategy to develop different pseudo-labels selection rules for objects with different scales.

For confidence inconsistency, we apply the soft label (Li et al. 2020) that the value at the ground-truth category index indicates its corresponding localization quality to replace the one-hot label to supervise the classification branch learning. However, the representation of proxy localization quality of oriented boxes has not been investigated. Through experimental analysis, we find that centerness (Tian et al. 2019) can effectively represent the localization quality of the oriented box and outperforms the predicted IoU (Intersection over Union), which dominates the horizontal box (Li et al. 2020; Liu et al. 2023a). Thereby we propose the Consistent Confidence Soft Label (CCSL) based on centerness with a scale factor to further promote the selection of pseudo-labels through mitigating the confidence inconsistency. In conclusion, our contributions are summarized as follows:

- Indicating the performance of semi-supervised oriented object detection is hindered by the assignment and confidence inconsistency that stem from the gaps between general and oriented object detection.
- Designing a Multi-clue Consistency Learning framework to enhance the performance of semi-supervised oriented object detection by alleviating the inconsistency.
- Achieving the state-of-the-art performance on DOTA-v1.5 and DOTA-v1.0 benchmark under SOOD task and demonstrating the importance of the analytics.

Related Work

Semi-supervised Object Detection

Most of previous methods (Yang et al. 2021; Li et al. 2022c; Liu et al. 2023b) are inherited from the Mean Teacher paradigm (Tarvainen and Valpola 2017), where the teacher model is updated via Exponential Moving Average (EMA) of the student weights and produce pseudo labels over unlabeled data as ground truth for training the student model. Under this scheme, the quality and precision of pseudo labels play a substantial role. ISMT (Yang et al. 2021) maintains a memory bank to ensure the consistency of pseudo labels across various iteration stages. Unbiased Teacher (Liu et al. 2021) applies Focal Loss (Lin et al. 2017b) to tackle the class-imbalance pseudo labels issues. Soft Teacher (Xu et al. 2021) employs a box jittering argumentation to select reliable pseudo labels. For misleading instances, Unbiased Teacher v2 (Liu, Ma, and Kira 2022) selects pseudo labels through utilizing uncertainty predictions. Consistent Teacher (Wang et al. 2023) dynamically adjusting the threshold for pseudo labels filtering. Mix Teacher (Liu et al. 2023b) enhances the quality of pseudo labels through mixed-scale prediction. Different from these, some methods (Zhou et al. 2022; Li et al. 2022b; Liu et al. 2023a) directly use the dense predictions of the teacher model as pixel-level pseudo-labels for semi-supervised learning. However, since the usage scenarios of oriented object detection are more complex, there still exist the performance gaps when applying these SSOD methods in semi-supervised oriented object detection.

Semi-supervised Oriented Object Detection

Oriented object detection requires predicting rotated bounding boxes by adding an angle parameter in the regression task, posing significant challenges for controlling the quality of pseudo boxes. SOOD (Hua et al. 2023) uses the absolute distance of the predicted angle between the teacher and student models to weight the regression loss. Additionally, it introduces the optimal transport cost (Monge 1781) to evaluate the global similarity of layouts between the predictions of teacher and student. PST (Wu, Wong, and Wu 2024) adopts two teacher models to mutually inspect their predictions, expecting to generate high-quality pseudo-labels. However, they fails to recognize the root issues hindering the performance of semi-supervised learning in oriented object detection, which lies in how to eliminate inconsistency issues caused by the gaps between general and oriented object detection. In this work, we dive into the gaps and propose a multi-clue consistency learning method for semi-supervised learning in oriented object detection.

Inconsistency and Gaps Analysis

In this section, we begin by introducing the classic SSOD framework. Under this framework, the gaps and resulting inconsistency issues between general and oriented object detection in the semi-supervised learning are investigated. To guarantee the generality, the representative general dataset COCO (Lin et al. 2014) and oriented dataset DOTA-v1.5 (Xia et al. 2018) are adopted to dissect their gaps.

The Basic SSOD Framework

Existing SSOD methods (Zhang, Pan, and Wang 2022; Zhang, Sun, and Wei 2023) frequently use the pseudo-labeling framework inherited from Mean Teacher (Tarvainen and Valpola 2017) as the baseline, which comprises two stages: burn-in stage and self-training stage. In the burn-in stage, the student model is pre-trained under the supervision of ground truth labels from labeled data, optimizing by minimizing the supervised loss $\mathcal{L}^s$. Simultaneously, the learned weights of the student model are mirrored onto the

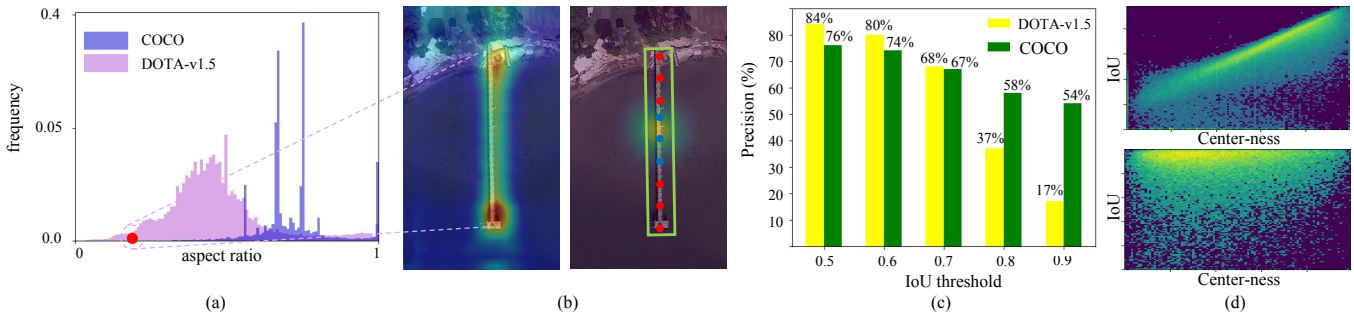


Figure 1: Gaps analysis between COCO and DOTA-v1.5 dataset. (a) Statistics of object aspect ratio distribution on both datasets. (b) The left is the class activation mapping of an object on DOTA-v1.5 with large aspect ratio. The right is the selected results by the common label assignment, where the red points and blue points represent negatives and positives respectively. (c) The pseudo-box precision of the two datasets under different IoU thresholds. (d) The top heatmap illustrates the consistency of centerness and IoU on the DOTA-v1.5 dataset, while the bottom one is for the COCO dataset.

teacher model. In the self-training stage, the unlabeled data with weak data augmentation is fed into the teacher model to generate pseudo-labels. Subsequently, the student model is further supervised by these generated pseudo-labels and optimized by unsupervised loss $\mathcal{L}^u$, while the input data is strongly augmented. Meanwhile, the supervised learning with $\mathcal{L}^s$ is maintained on the student network. The overall loss function is thus formulated as $\mathcal{L} = \mathcal{L}^s + \beta\mathcal{L}^u$, where β is a hyper-parameter regulating the impact of unsupervised learning. The network parameters of the student model are updated onto the teacher network through an exponential moving average (EMA) mechanism at each iteration.

Nevertheless, there is a performance gap between general and oriented object detection under this framework. We find it is caused by the assignment and confidence inconsistency problems that stem from the gaps between general and oriented object detection. The detailed analysis is as follows.

Assignment Inconsistency

First, there is a gap between general and oriented objects in aspect ratio. We statistically analyze the aspect ratio gaps of objects on the COCO and DOTA-v1.5 datasets as shown in Figure 1(a). Most of the objects are with the aspect ratios exceeding 0.5 on the COCO dataset, while more than 70% of objects exhibit aspect ratios less than 0.5 on the DOTA-v1.5 dataset, especially for some objects with extreme aspect ratios (seen in Figure 1(b)). This significant gap in aspect ratios can lead to assignment inconsistency when performing label assignment under the supervision of ground truth labels from labeled data. Taking the center sampling used in FCOS as an example, only the blue points can be selected as positive labels, while these red points would be seen as negative labels, leading to incorrect guidance for the model optimization. Therefore, the inconsistency between the shape-agnostic center sampling and objects with large aspect ratios would cause the performance degradation in the SOOD task.

Second, compared with the horizontal object detection in the general scenes, the localization of oriented objects is more challenging due to their arbitrary rotation angles. To elucidate this, we conduct evaluations with the FCOS detector on the DOTA-v1.5 and COCO datasets, and then utilize the Intersection over Union (IoU) metric between predicted boxes and the corresponding ground-truth boxes to assess the localization quality of the pseudo-boxes. As shown in Figure 1(c), we use different IoU thresholds to obtain the pseudo-box precision under the same recall rate on the DOTA-v1.5 and COCO datasets, respectively. With a lower IoU threshold, we can achieve a higher precision on the DOTA-v1.5 dataset compared to the COCO dataset, but the pseudo-boxes with the poor localization quality would affect the detection optimization. Under the guidance of low-quality rotated pseudo-boxes, noisy labels are inevitably introduced during the label assignment process. When improving the requirement of localization quality by increasing the IoU threshold (from 0.5 to 0.9) on both datasets, the detection precision decreases from 76% to 54% on the COCO dataset. However, the precision would severely drop from 84% to 17% on the DOTA-v1.5 dataset. A large number of redundant noisy boxes caused by reduced precision can also introduce the noisy labels. Hence, it's hard to achieve a trade-off between the precision and localization quality of rotated pseudo-boxes.

Confidence Inconsistency

Prior research (Li et al. 2020) has validated that the confidence inconsistency problem is existing between the classification and localization qualities. To eliminate the inconsistency, most previous horizontal object detectors (Li et al. 2020; Liu et al. 2023a) use the IoU-based soft label to supervise the classification branch. However, predicting a confidence IoU value is difficult for the oriented objects due to the uncertainty introduced by the rotation angle parameter. Meanwhile, we discover that centerness can well represent the localization quality of oriented bounding boxes. As in Figure 1(d), we statically analyze the IoU and centerness between all ground truth boxes and their corresponding true positive boxes on the DOTA-v1.5 dataset (the top subfigure) and the COCO dataset ((the bottom subfigure). Compared to the COCO dataset, the correlation between the IoU and centerness values is significantly higher on the DOTA-v1.5 dataset. Therefore, we utilize the predicted centerness value to measure the localization quality of oriented objects.

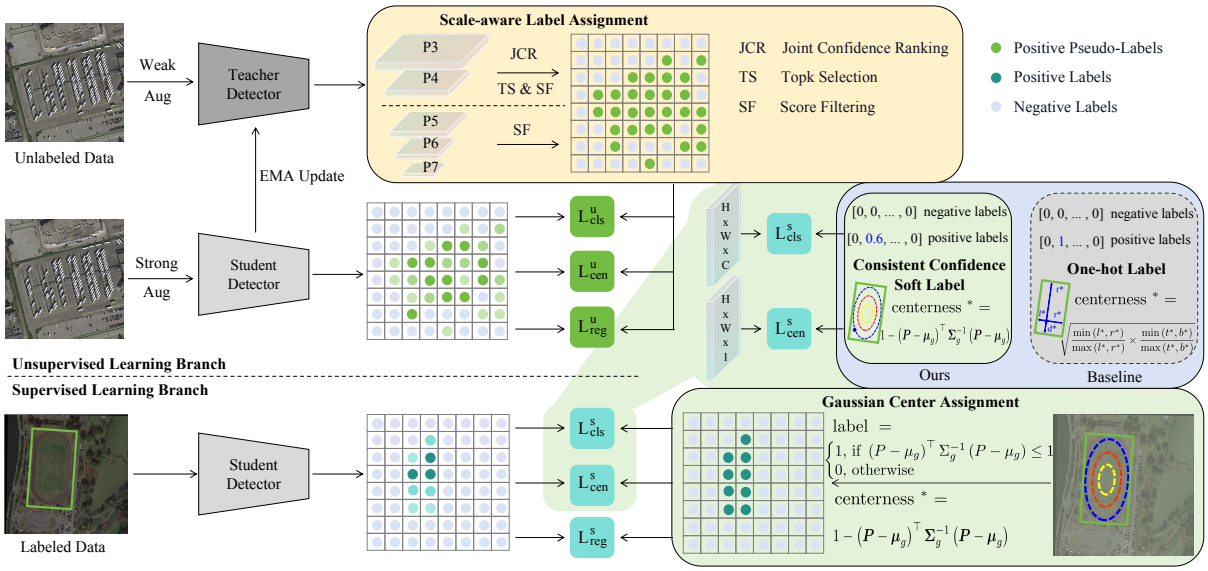


Figure 2: The pipeline of the MCL. The GCA is introduced to select shape-aware positive labels from labeled data. The SLA is proposed to select scale-aware pixel-level pseudo-labels. The CCSL is adopted to mitigate the confidence inconsistency.

Methodology

To mitigate the inconsistency and bridge the gaps, we subsequently propose the Multi-clue Consistency Learning (MCL) framework, which comprises three modules: Gaussian Center Assignment, Scale-aware Label Assignment and Consistent Confidence Soft Label.

Gaussian Center Assignment

To bridge the gap in aspect ratios and boost the SOOD performance, GCA aims to select more accurate positive labels by: 1) considering the shape information of oriented objects and 2) making full use of valuable labeled data.

Concretely, given an oriented box $B = (x_c, y_c, w, h, \theta)$, where θ is the angle parameter; (x_c, y_c), w, and h represent the center coordinates, width, and height of the oriented box respectively, we model the object as a $2D$ Gaussian distribution (Yang et al. 2022) that can well reflect the shapes and directions of the oriented objects. The coordinate of the center point (x_c, y_c) serves as the mean vector $\boldsymbol{\mu} = (x_c, y_c)^\top$, and the co-variance matrix $\boldsymbol{\Sigma}$ can be formulated as:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \frac{w^2}{4} & 0 \\ 0 & \frac{h^2}{4} \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \tag{1}$$

We then define a point coordinate as $\boldsymbol{P} = (x, y)^\top$ and determine whether it is positive based on the following formula:

$$\text{label} = \begin{cases} 1, & \text{if } (\boldsymbol{P}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{P}-\boldsymbol{\mu}) \le 1 \\ 0, & \text{otherwise .} \end{cases} \tag{2}$$

This label assignment ensures that positive labels cover the majority of the object's region, thereby avoiding the omission of discriminative features for objects with large aspect ratios. Moreover, a large number of pixels within the objects provide sufficient supervision information for optimizing the detector. According to the prior that pixels located around the center of the object are more representative than those near the object boundaries, we establish a new normalized distinction in the object region to represent the pixel-wise localization quality:

$$\text{centerness}^* = 1 - (\boldsymbol{P}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{P}-\boldsymbol{\mu}). \tag{3}$$

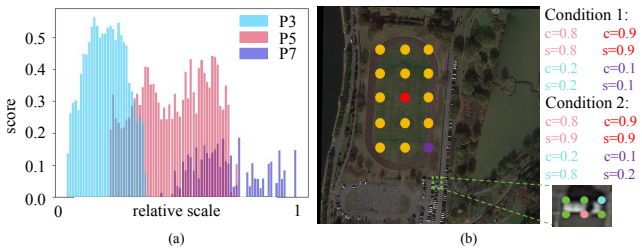
The GCA can select comprehensive supervision signals from the precious labeled data and further improve the performance for objects with large aspect ratios.

Scale-aware Label Assignment

To alleviate the assignment inconsistency, we propose SLA to select pixel-level pseudo-labels based on the feature level predictions of the teacher model rather than the unreliable pseudo-boxes. It considers two imbalance issues: 1) Score imbalance between features of different levels, where the higher level feature map in FPN (Lin et al. 2017a) tend to predict the lower scores(shown in Figure 3(a)). It causes an imbalance in the label allocation of objects of different scales. 2) Value imbalance that the joint confidence(denoted as the multiplication of score and centerness) is dominated by centerness due to its higher value, which is unreliable to determine whether a pixel is positive since the centerness branch lacks supervision from negative instances.

SLA adopts a divide-and-rule strategy to develop different selection rules for objects with different scales. For the low-level feature maps (i.e., P_3, P_4) with higher scores, a coarse-to-fine rule with two stages is designed to predict the small-scale objects. Coarse selection stage is expected to select candidates with high classification and localization qualities. Thus the joint confidence is used to rank all pixels and the top-k sorted pixels will be selected as candidates. Considering the value imbalance, a score threshold is further applied to filter these candidates in the fine selection stage.

For the high-level feature maps (i.e., P_5, P_6, P_7) with the lower predicted scores, we only use the score threshold to fil-



(a) (b)

Figure 3: (a) Score imbalance between feature maps at different levels. (b) For centerness-based soft label, condition 1 introduce ambiguities while condition 2 is more suitable.

ter high-confidence pseudo-labels, thereby avoiding the unbalanced pseudo-labels allocation across various scales. The unsupervised loss consists of three parts:

$$\mathcal{L}_u = \frac{1}{N_{all}} \sum_{i}^{N_{all}} \mathcal{L}_i^{cls} + \frac{\alpha}{N_{pos}} \sum_{j}^{N_{pos}} w_j (\mathcal{L}_j^{cen} + \mathcal{L}_j^{reg}) \quad (4)$$

where α is a weighting parameter, N_{all} represents the number of pixels across five level feature maps, N_{pos} denotes the number of selected pseudo-labels. We apply the quality focal loss (Li et al. 2020), the binary cross-entropy loss and the smooth-l1 loss for guiding the classification, centerness and regression branches, respectively. Besides, the localization weights w are determined by the following formula:

$$w_j = \begin{cases} s_j \times c_j & j \in [P_3, P_4] \\ s_j & j \in [P_5, P_6, P_7] \end{cases} \quad (5)$$

where s and c denote the score and centerness, respectively. Since the existence of false positive labels is inevitable and the localization task is more sensitive to such noisy samples, employing the weight w can reduce the contribution of low-confidence samples to the overall loss.

Consistent Confidence Soft Label

The CCSL is designed to alleviate the confidence inconsistency issue. It has to satisfy two conditions: (1) being positively correlated with centerness; (2) the score variance among all points within the target should not be too large, especially for small targets. Therefore, we first employ the soft label (Li et al. 2020) to establish correlations between score and centerness confidences. The core idea is to replace the value of one-hot label at corresponding category index with a float value $y \in [0, 1]$, which satisfies the condition that the covariance calculated with centerness equal to 1. The positive value $y \in [0, 1]$ is given as follows:

$$\mathrm{y} = [1 - (\boldsymbol{P} - \boldsymbol{\mu}_g)^\top \boldsymbol{\Sigma}_g^{-1} (\boldsymbol{P} - \boldsymbol{\mu}_g)]^\gamma \quad (6)$$

where γ is a scale factor designed for the second condition. As shown in Figure 3(b), all points belonging to a small object have smaller stride, resulting in very similar point features. We hope that the score values of all points within the target are very close. Therefore, we weight them with the scale factor γ, which is formulated as $\sqrt[\beta]{(h \times w)/(H \times W)}$. h and w denote the height and width of the oriented box, H and W denote the height and width of the whole image, while β controls the smoothing degree of γ. γ can prevent confusion during model training caused by excessive variance in point scores within the target.

Task	Method	mAP(%) ↑		
		10%	20%	30%
OD	Faster RCNN	43.43	51.32	53.14
	FCOS	42.78	50.11	54.79
	RetinaNet	38.96	45.87	50.25
SSOD	Unbaised Teacher°	44.51	52.80	53.33
	Soft Teacher°	48.46	54.89	57.83
	Dense Teacher†	46.90	53.93	57.86
	PseCo°	48.04	55.28	58.03
	DualPolish°	49.02	55.17	58.44
	ARSL†	48.17	55.34	59.02
SOOD	SOOD*†	48.63	55.58	59.23
	PST°	49.63	57.39	60.40
	MCL‡(Ours)	**51.53**	**58.26**	**61.23**
	MCL†(Ours)	**52.98**	**59.63**	**62.63**

Table 1: Comparison with other state-of-the-art methods on the DOTA-v1.5-Partial dataset. °, † and ‡denotes the base detector is Faster RCNN, FCOS and RetinaNet respectively.

Experiment

Experimental Setup

Dataset: The experiments are conducted on DOTA-v1.5 and DOTA-v1.0 (Xia et al. 2018). DOTA-v1.5 comprises 16 categories which contains 400k annotated oriented instances. DOTA-v1.5-train, DOTA-v1.5-val, and DOTAv1.5-test contain 1411, 458, and 937 images, respectively. DOTA-v1.0 uses the same images as DOTA-v1.5 and the number of annotated instances is half the DOTA-v1.5's, with targets having a pixel area smaller than 10 being ignored. We include three evaluation protocols: 1) DOTA1.5-Partial. Following SOOD method (Hua et al. 2023), we randomly sample10%, 20%, and 30% images from DOTA-v1.5-train as labeled data and set the remaining images as unlabeled data. 2) DOTA1.5-Full. We set DOTA-v1.5-train as labeled data, DOTA-v1.5-test as unlabeled data and perform the evaluation on the DOTA-v1.5-val. 3) DOTA1.0-Full. The DOTA-v1.0-train and DOTA-v1.0-test are set as labeled data and unlabeled data respectively, while the DOTA-v1.0-val is set as the evaluation dataset. For all evaluation protocols, the mAP (Xia et al. 2018) is adopted as the evaluation metric.

Implementation Details: We adopt the FCOS (Tian et al. 2019) as the detector and ResNet-50 (He et al. 2016) pre-trained on ImageNet (Deng et al. 2009) as the backbone. The MCL is trained on 2 RTX4090 GPUs with 3 images per GPU (2 labeled images and 1 unlabeled image). To ensure a fair comparison, the image crop, data augmentation and training optimization strategies are consistent with previous work (Hua et al. 2023). The burn-in strategy is employed to initialize the parameters of the teacher model and the EMA strategy is applied to update its parameters at each iteration.

Comparison with State-of-the-Arts

To validate the effectiveness of MCL, we compare it with previous methods under DOTA1.5-Partial, DOTA1.5-Full

Method	DOTA-v1.5	DOTA-v1.0
Unbiased Teacher	66.12 → 64.85	-
Soft Teacher	66.12 → 66.40	-
Dense Teacher	65.46 → 66.38	66.72 → 70.30
SOOD*	65.46 → 67.70	66.72 → 70.81
MCL(Ours)	**65.46 → 69.08**	**66.72 → 73.63**

Table 2: Experiments on DOTA-v1.5-Full and DOTA-v1.0-Full, while the numbers in font of the arrow indicate the result of supervised baseline.

and DOTA1.0-Full evaluation protocols. For distinguishing with SOOD task, the SOOD (Hua et al. 2023) method is termed as SOOD*. Moreover, some SSOD methods with their oriented version also participate in the comparison.

DOTA-v1.5-Partial: The results under the DOTA-v1.5-Partial protocol are presented in Table 1. In SOOD task, SSOD methods typically underperform the SOOD methods, which exhibits the necessity of bridging the gap between horizontal boxes and oriented boxes in semi-supervised detection. SOOD* (Hua et al. 2023) and PST (Wu, Wong, and Wu 2024) are specifically designed for the characteristics of oriented objects, yet the performance gains are marginal. By bridging the gaps and alleviating the inconsistency, our MCL achieves remarkable enhancements over the base detector FCOS and outperforms all SOOD methods under all proportions. It achieves 52.98, 59.63 and 62.63 mAP on 10% 20% 30% experiments respectively, largely surpassing the state-of-the-arts method PST by $2 \sim 3.5$ mAP, validating the necessity of our analyses and methods.

DOTA-v1.5-Full and DOTA-v1.0-Full: Table 2 gives the comparison results on the DOTA-v1.5-Full and DOTA-v1.0-Full settings. Compared to the SOOD* using the same detector, the MCL can achieve the notable improvement: 1.38% on the the DOTA-v1.5-Full setting, and 2.82% on the DOTA-v1.0-Full setting, exhibit its effectiveness.

Generalization of MCL: We apply MCL in RetinaNet (Lin et al. 2017b) to verify its scalability. Concretely, we preset one anchor on each pixel of feature map and select positives through GCA based on the center location of the anchor, instead of inefficient IoU-based label assignment. SLA can be employed to select anchors in the unlabeled data training branch, and CCSL can be used by adding a lightweight convolution layer for predicting centerness. The result shown in Table 1 demonstrates that MCL is effective for both anchor-free and anchor-based detectors.

Ablation Studies

Subsequently, we delve into the analysis of each module. All the ablation experiments are performed on the 30% setting of DOTA-v1.5-Partial without special instructions. We set the Dense Teacher (Zhou et al. 2022) as the baseline. It also uses pixel-level pseudo labels by sorting all pixels according to classification scores and selects the top ratio(%) ones. The original parameter setting sets ratio to 1%, and we set ratio to 3% by default to maximize the performance.

The effect of each module: We study the effectiveness of our proposed three modules under different experiment settings (10% 20% 30%). The results are shown in Table 3. After comprehensively extracting supervision information from annotated data, GCA facilitates a performance enhancement for MCL. Furthermore, SLA enhances the quality of unsupervised information extracted from unlabeled data and also contributes to performance improvement. The CCSL can further improve the precision of pseudo-labels by mitigating the inconsistency between classification and localization confidence, thereby the participation of the CCSL leads to additional performance improvement for MCL.

	GCA	SLA	CCSL	mAP(%) ↑		
				10%	20%	30%
Base	×	×	×	49.71	55.64	59.65
MCL	√	×	×	51.12	56.74	60.62
	×	√	×	51.84	57.16	61.32
	√	√	×	52.02	57.28	61.46
	√	√	√	52.98	59.63	62.63

Table 3: Performance comparison with different modules.

	ratio	thr	topk	R	P	mAP
s	1%	-	-	40.08	30.76	57.86
	3%	-	-	69.44	17.76	59.65
s × c	3%	-	-	70.38	18.01	60.45
SLA	-	0.01	1500	90.35	12.74	59.97
		0.02		88.76	19.50	60.61
		0.03		87.15	24.03	60.59
	-	0.01	2000	93.72	12.39	60.34
		0.02		91.89	19.13	61.32
		0.03		90.04	23.69	60.89

Table 4: Ablation studies on SLA. s and c denote score value and centerness value respectively. R and P denote Recall and Precision respectively.

All Sampling vs GCA: To evaluate the effective of the GCA in SOOD task, we compare GCA with the center sampling strategy used in FCOS (Tian et al. 2019) and the all sampling strategy that labels all pixels insides ground-truth boxes as positives. The results are shown in Table 5. Although the aim is to fully utilize labeled data to extract more positive supervision information, all sampling strategy still shows performance degradation compared to center sampling. We speculate that this is due to it introduces excessive background at the target's edges. In contrast, GCA can provide sufficient positive supervision information while accounting for the target's aspect ratio without introducing excessive noise, thereby improving the performance.

The analysis of SLA: To verify the impact of SLA hyperparameters and demonstrate the advantages of SLA over other selection strategies, we conduct the experiments with the pixel-level recall and precision as the metrics. In detail, we denote the selected points from the ground truth boxes as ground truth points. The selected pseudo-labels that coincide with ground truth points are denoted as true positives and the remaining ones are false positives. The computation

Sampling strategy	mAP
CS	59.65
AS	58.49
GCA	**60.62**

Table 5: Performance comparison between GCA and All Sampling(AS) and Center Sampling(CS).

Method	mAP
FCOS(base)	54.79
Mean Teacher	26.02
Dense Teacher	59.65

Table 6: Performance comparison between Mean Teacher and Dense Teacher.

	HB		BD		mAP
	Recall	mAP	Recall	mAP	
FCOS	81.7	71.8	60.8	40.1	65.5
+GCA	84.6	74.1	66.5	45.7	67.2
RetinaNet	77.1	61.5	62.8	41.0	61.1
+GCA	82.5	72.9	66.1	45.2	64.2

Table 7: The impacts of GCA on large aspect ratio objects.

forms of pixel-level recall and precision are consistent with the previous work (Lin et al. 2014). The experimentation is shown in Table 4. It can be observed that the performance of semi-supervised learning improves 1.79% mAP when recall increases from 40.08% to 69.44%, demonstrating that maintaining high recall of pseudo-labels is crucial for ensuring the performance of SOOD task. In addition, replacing score selection used in Dense Teacher (Zhou et al. 2022) with joint confidence selection can slightly enhances both recall and precision, thereby improving the performance from 59.65 mAP to 60.45 mAP. However, it is still not the optimal solution. By considering the scale and value imbalance, SLA significantly enhances recall while maintaining satisfactory precision, thereby markedly improving SOOD performance.

The analysis of CCSL: To validate our analysis that predicted centerness is a more suitable proxy localization quality of oriented box than predicted IoU, we perform the comparative experiment between them. As shown in Table 8, compared to predicting the uncontrollable IoU, predicting the centerness with considering the scale information improves the SOOD performance. In addition, the scale factor γ is also a critical component. By combining the scale factor γ, the variance of the target internal center distribution can be dynamically adjusted according to the target scale. Therefore, compared to vanilla centerness, centerness with the scale factor can better enhance the detection performance.

Inconsistency Mitigation Investigation

Assignment Inconsistency: We first conduct additional experiments to demonstrate the effectiveness of GCA for objects with large aspect ratios. The models are trained on the DOTA-v1.5-train and evaluated on the DOTA-v1.5-val. We show the results of two categories with large aspect ratios in Table 7, including HB (harbor), BD (bridge). Supported by GCA, there is a notable enhancement in recall for these objects with large aspect ratios, which verifies that GCA can effectively alleviating the assignment inconsistency in labeled data training stage.

Localization Quality	β	mAP
IoU	-	60.98
Centerness	0	60.87
	0.1	61.38
	0.2	62.63
	0.25	62.26

Table 8: Ablation studies on CCSL.

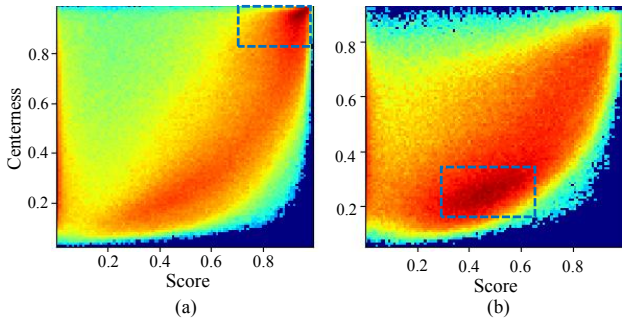


Figure 4: Illustrated analysis of CCSL. (a) and (b) are the heatmaps of the scores and centerness of positive samples trained based on CCSL and one-hot label respectively.

To verify the necessity of using pixel-level pseudo-labels in unlabeled data training stage as our analysis, we compare the pseudo-boxes-based method Mean Teacher(the vanilla pseudo-boxes framework) with the pixel-level pseudo-labels-based method Dense Teacher (Zhou et al. 2022) on FCOS. As shown in Table 6, Mean Teacher exhibits a significant performance drop due to the fact that dense object detectors adopt a binary label assignment strategy, making them highly sensitive to the overall quality of the oriented pseudo-boxes which is difficult to guarantee. In contrast, pixel-level pseudo-labels alleviate the assignment inconsistency, thereby improving the performance.

Confidence Inconsistency: To assess the effectiveness of CCSL in eliminating confidence inconsistency, we train two FCOS detectors on DOTA-v1.5-train, one of which is supervised with one-hot label as usual and the other with CCSL. We visualize the score-centerness heatmaps of all ground truth points on the DOAT-v1.5-val set, which is shown in Figure 4. As highlighted in the blue squares, with the incorporation of CCSL, there is an improvement in the consistency between scores and centerness, which demonstrates that the CCSL is effective in eliminating the inconsistency between classification and localization qualities.

Conclusion

In this work, we investigate the gaps between general and oriented object detection in semi-supervised learning. To address the resulting inconsistency issues, the Multi-clue Consistency Learning framework is proposed, which consists of Gaussian Center Assignment, Scale-aware Label Assignment and Consistent Confidence Soft Label. Eexperiments on the DOTA-v1.0 and DOTA-v1.5 datasets verify its effectiveness to alleviate the inconsistency.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grants Nos. 62372238,62476133)

References

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. Ieee.

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.

Hua, W.; Liang, D.; Li, J.; Liu, X.; Zou, Z.; Ye, X.; and Bai, X. 2023. SOOD: Towards semi-supervised oriented object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 15558–15567.

Li, G.; Li, X.; Wang, Y.; Wu, Y.; Liang, D.; and Zhang, S. 2022a. Pseco: Pseudo labeling and consistency training for semi-supervised object detection. In *European Conference on Computer Vision*, 457–472.

Li, G.; Li, X.; Wang, Y.; Yichao, W.; Liang, D.; and Zhang, S. 2022b. Dtg-ssod: Dense teacher guidance for semi-supervised object detection. *Advances in Neural Information Processing Systems*, 35: 8840–8852.

Li, H.; Wu, Z.; Shrivastava, A.; and Davis, L. S. 2022c. Rethinking pseudo labels for semi-supervised object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1314–1322.

Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; and Yang, J. 2020. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*, 33: 21002–21012.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017a. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2117–2125.

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.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European Conference Computer Vision*, 740–755.

Liu, C.; Zhang, W.; Lin, X.; Zhang, W.; Tan, X.; Han, J.; Li, X.; Ding, E.; and Wang, J. 2023a. Ambiguity-resistant semi-supervised learning for dense object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 15579–15588.

Liu, L.; Zhang, B.; Zhang, J.; Zhang, W.; Gan, Z.; Tian, G.; Zhu, W.; Wang, Y.; and Wang, C. 2023b. Mixteacher: Mining promising labels with mixed scale teacher for semi-supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7370–7379.

Liu, Y.-C.; Ma, C.-Y.; He, Z.; Kuo, C.-W.; Chen, K.; Zhang, P.; Wu, B.; Kira, Z.; and Vajda, P. 2021. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*.

Liu, Y.-C.; Ma, C.-Y.; and Kira, Z. 2022. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9819–9828.

Monge, G. 1781. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, 666–704.

Sohn, K.; Zhang, Z.; Li, C.-L.; Zhang, H.; Lee, C.-Y.; and Pfister, T. 2020. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*.

Sun, P.; Jiang, Y.; Xie, E.; Shao, W.; Yuan, Z.; Wang, C.; and Luo, P. 2021. What makes for end-to-end object detection? In *International Conference on Machine Learning*, 9934–9944. PMLR.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 9627–9636.

Wang, X.; Yang, X.; Zhang, S.; Li, Y.; Feng, L.; Fang, S.; Lyu, C.; Chen, K.; and Zhang, W. 2023. Consistent-Teacher: Towards Reducing Inconsistent Pseudo-Targets in Semi-Supervised Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3240–3249.

Wu, W.; Wong, H.-S.; and Wu, S. 2024. Pseudo-Siamese Teacher for Semi-Supervised Oriented Object Detection. *IEEE Transactions on Geoscience and Remote Sensing*.

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.

Xu, M.; Zhang, Z.; Hu, H.; Wang, J.; Wang, L.; Wei, F.; Bai, X.; and Liu, Z. 2021. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3060–3069.

Yang, Q.; Wei, X.; Wang, B.; Hua, X.-S.; and Zhang, L. 2021. Interactive self-training with mean teachers for semi-supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5941–5950.

Yang, X.; Zhang, G.; Yang, X.; Zhou, Y.; Wang, W.; Tang, J.; He, T.; and Yan, J. 2022. Detecting rotated objects as gaussian distributions and its 3-d generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4335–4354.

Zhang, F.; Pan, T.; and Wang, B. 2022. Semi-supervised object detection with adaptive class-rebalancing self-training. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 3252–3261.

Zhang, L.; Sun, Y.; and Wei, W. 2023. Mind the gap: Polishing pseudo labels for accurate semi-supervised object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3463–3471.

Zhou, H.; Ge, Z.; Liu, S.; Mao, W.; Li, Z.; Yu, H.; and Sun, J. 2022. Dense teacher: Dense pseudo-labels for semi-supervised object detection. In *European Conference on Computer Vision*, 35–50.