Denser Teacher: Rethinking Dense Pseudo-Label for Semi-supervised Oriented Object Detection

Tong Zhao, Qiang Fang, Xin Xu

***Abstract*—Oriented object detection, which aims to detect multi-oriented objects, is a fundamental task for visual analysis in complex scenarios, such as aerial images. However, powerful detection performance relies on abundant and accurate annotations. Therefore, semi-supervised oriented object detection, which utilizes unlabeled data to improve performance, is a promising method to address this problem. In this work, we explore Dense Pseudo-Label (DPL), which directly selects pseudo labels from the original output of the teacher model without any complicated post-processing steps, and expose the shortcomings of existing methods. Through analysis, we identify that the imbalance between obtaining potential positive samples and removing the interference of inaccurate pseudo labels hinders the effectiveness of DPL. To further improve DPL efficiency, we propose Denser Teacher, a new semi-supervised oriented object detection method. In this method, we design a simple yet effective adaptive mechanism called global dynamic k estimation to guide the selection of DPLs in densely-distributed scenes. Additionally, to improve scale adaptation, we introduce dense multi-scale learning for DPL, where DPLs from different scales are utilized to bridge the scale gap. We conduct extensive experiments on several benchmarks to demonstrate the effectiveness of our proposed method in leveraging unlabeled data for performance improvement. Our code will be available at https://github.com/Haru-zt/DenserTeacher.**

***Index Terms*—aerial images, semi-supervised learning, object detection**

I. Introduction

ORIENTED object detection is a significant research field for visual analysis in complex scenarios, such as aerial images [1]–[5]. Currently, deep learning-based methods dominate the field and have achieved rapid development. However, the progressive performance of oriented object detection is based on massive annotations. When provided with limited annotations, the performance of oriented object detectors drops severely [6]. Moreover, annotating abundant fully labeled datasets is costly and time-consuming. To effectively leverage abundant unlabeled data, Semi-Supervised Object Detection (SSOD) has garnered extensive attention [7]–[9]. However, existing SSOD works [10]–[13] mainly focus on general object detection, where objects are annotated with horizontal boxes. In some scenes, such as aerial images, horizontal boxes have difficulty efficiently representing objects [1], [14]. In contrast to general scenes, objects in aerial images are typically captured from a bird's-eye view and consequently present additional challenges, including arbitrary orientations, multiple scales, and dense distributions [1]. Therefore, semi-supervised oriented object detection deserves serious consideration.

Existing SSOD methods strongly rely on precise pseudo-labels, which can be divided into Sparse Pseudo-Label (SPL) [10], [12], [15], [16] and Dense Pseudo-Label (DPL) [6], [11], based on the sparsity of pseudo-labels. In SPL, bounding boxes and their labels are provided as supervision information, similar to the ground truth. For DPL, pseudo labels are directly selected from the original output of the teacher model without any complicated post-processing steps. By removing post-processing steps, DPL retains richer information and has thus received extensive attention [11].

However, existing DPL-based methods are inefficient for aerial scenes. Dense Teacher [11] proposes a region selection technique to highlight key information and suppress noise, but it requires a fixed selection ratio to control the number of pseudo labels. This limitation restricts the selection of sufficient pseudo labels in dense scenes and may cause the selected pseudo labels to contain abundant noise in other scenes. SOOD [6] combines DPL with SPL to reduce noise. In SOOD [6], DPLs are randomly sampled from the teacher's predictions, but this approach involves a sequence of post-processing steps with fine-tuned hyper-parameters, which has been shown to be sensitive in dense scenes [11].

In this study, we note that although some DPL-based methods achieve competitive performance in semi-supervised oriented object detection, the potential of DPL-based methods is still largely hindered by the imbalance between obtaining potential positive samples and removing the interference of inaccurate pseudo labels. To verify this phenomenon, we analyze the effectiveness of existing DPL-based methods, as shown in Fig. 1. To simplify the analysis, we calculate the True Positive (TP), False Positive (FP), and False Negative (FN) numbers of DPL-based methods Dense Teacher [11] and SOOD [6] on the DOTA-v1.5 validation set. Note that all models are trained under the DOTA-v1.5 10% partially labeled setting. We observe that Dense Teacher [11] obtains the fewest FNs, indicating its effectiveness in mining potential positives but suffers from insufficient TPs and abundant FPs. We conjecture that the fixed selection ratio causes this problem. SOOD [6] greatly alleviates this problem by introducing SPLs to improve the quality of DPLs (TP +79.9% and FP -32.6%), but consequently results in a significant increase in FN (+166.8%), indicating that SOOD [6] still struggles with obtaining potential positive samples.

Through analysis, we identify that an essential cause hin-

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

(Corresponding authors: Qiang Fang.)

Tong Zhao, Qiang Fang and Xin Xu are with the College of Intelligence Science and Technology, National University of Defense Technology, Changsha, 410000 China. *E-mail:* {*zhaotong, qiangfang, xinxu*}*@nudt.edu.cn*

Manuscript received XXX, 2022; revised XXX.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2024.3518452

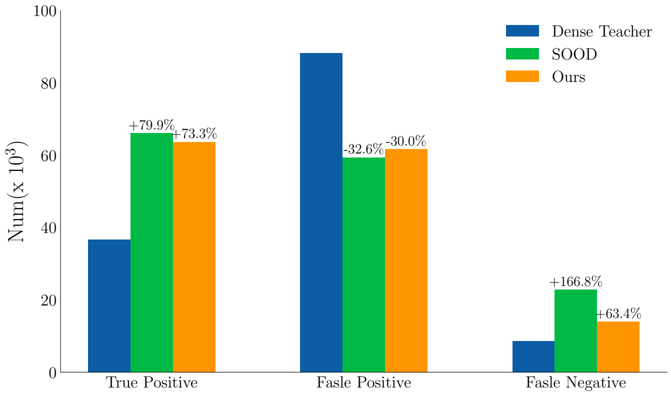


Fig. 1. The True Positive (TP), False Positive (FP), and False Negative (FN) statistics of DPL-based methods in the DOTA-v1.5 10% partially labeled setting. The statistics are measured on the DOTA-v1.5 validation set.

dering the effectiveness of DPL is the imbalance between obtaining potential positive samples and removing the interference of inaccurate pseudo labels. To overcome this problem, we propose integrating potential object information into the DPL selection process. Such design carefully handles dense distribution challenge of oriented object detection, as we follow the natural idea that we should select suitable DPLs according to the quantity of potential objects. Moreover, the scales of oriented objects in aerial images vary significantly across different categories and scenes, which presents a new challenge to DPL-based methods [17], [18]. Existing works have demonstrated that incorporating an extra down-sampled view of the unlabeled image and regularizing the network with consistency constraints at either the feature level or label level can significantly improve performance [13], [16], [19]. However, these methods mostly focus on SPL and do not facilitate DPL, leaving the possibility of building multi-scale learning for DPL.

To address these issues, we propose a novel method called Denser Teacher for semi-supervised oriented object detection. To select proper DPLs in densely-distributed scenes, we design an adaptive mechanism called Global Dynamic K Estimation (GDE) to estimate the quantity of potential objects in an image and use this information to guide the selection of DPLs. Additionally, to mitigate scale variance, we propose Dense Multi-Scale Learning (DMSL) for DPL, in which DPLs with different scales are selected to build a more direct and effective way to improve scale adaptation.

We summarize our main contributions as follows:

1) We investigate the effectiveness of dense pseudo-labels and expose the shortcomings of existing dense pseudo-label methods.
2) We propose an adaptive mechanism called Global Dynamic K Estimation (GDE) to formulate a direct way to integrate potential objects information into the dense pseudo-label selection process and select suitable pseudo labels.
3) We introduce Dense Multi-Scale Learning (DMSL) for dense pseudo-labels, in which dense pseudo labels from different scales are utilized to improve scale adaptation.
4) Our Denser Teacher contributes significant performance gains on several benchmarks, confirming the effectiveness of our proposed method.

In the following manuscript, Sec. II introduces the related work on semi-supervised oriented object detection; Sec. III discusses the proposed method, Denser Teacher; Sec. IV shows the experimental setting and results; Sec. V presents the discussion; and Sec. VI presents the conclusion.

II. Related Work

A. Oriented Object Detection

Unlike general object detection, oriented object detection represents objects with Oriented Bounding Boxes (OBBs). In recent years, oriented object detection has witnessed significant progress due to the rapid development of deep learning. RoI Transformer [20] proposed an RRoI learner to convert horizontal regions of interest (HRoIs) into rotated regions of interest (RRoIs) and an RPS RoI Align module to extract spatially rotation-invariant feature maps. R^3Det [21] introduced a coarse-to-fine approach to reconstruct feature maps by designing a feature refinement module. ReDet [22] proposed rotation-equivariant networks and RiRoI Align to extract rotation-invariant features. Oriented R-CNN [2] proposed a new rotated object representation based on midpoint offset and designed an oriented RPN to reduce the cost of proposals. LSKNet [3] introduced large and selective kernel mechanisms into oriented object detection to incorporate prior knowledge. Moreover, discontinuity in oriented object detection has received much attention. GWD [23], KLD [24], and KFIoU [25] used Gaussian distributions to represent OBBs and demonstrated effectiveness in alleviating the impact of discontinuity. CSL [26] transformed the angular prediction task from a regression problem to a classification task to solve the issue of discontinuous boundaries. Gliding Vertex [27] explored a new OBB representation by sliding the four vertices of an HBB (Horizontal Bounding Box) to construct an OBB. Transformer-based methods [17], [28] have also been developed for oriented object detection. The above methods enhanced detection performance by fully leveraging the characteristics of oriented objects. However, these methods usually required a large amount of training data with fully labeled annotations, which are costly and time-consuming. Our method aims to improve the performance of semi-supervised oriented object detection and alleviate the demand for abundant annotations.

B. Semi-Supervised Object Detection

Recently, semi-supervised learning (SSL), which aims to improve performance by leveraging a limited amount of labeled data alongside a large volume of unlabeled data, has achieved significant results in image classification. Most existing works in SSL can be roughly categorized into pseudo-labeling and consistency regularization. In contrast, SSOD methods need to make instance-level predictions and regress the corresponding bounding boxes, which makes them more challenging. STAC [29] proposed a multi-stage SSOD training framework that combined pseudo-labeling and consistency training by utilizing weak and strong augmentations inspired

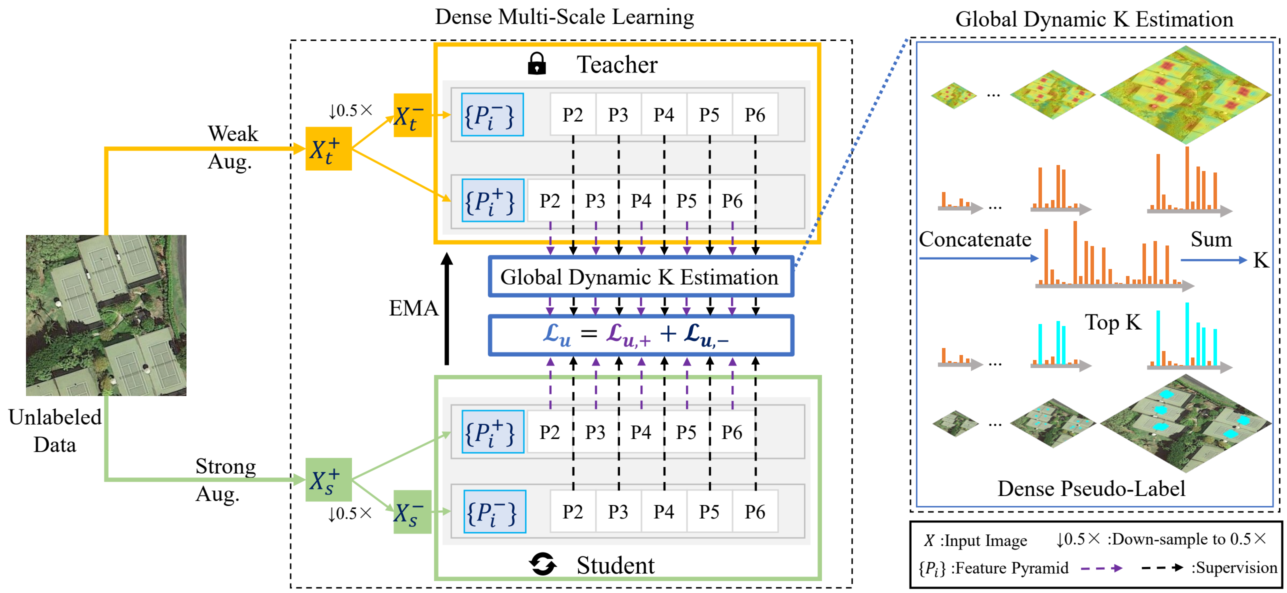


Fig. 2. Overview of the proposed Denser Teacher. It comprises a student model and a teacher model. During training, the teacher model parameters are updated from the student model using Exponential Moving Average (EMA). Global Dynamic K Estimation is employed to select suitable DPLs for unlabeled data according to K, and Dense Multi-Scale Learning is employed to improve the scale adaptation. Note that the labeled part is hidden for simplicity.

by FixMatch [30]. Unbiased Teacher [15] addressed the class imbalance problem in pseudo-labeling by focusing on learning rare classes using focal loss. Soft Teacher [10] adopted classification scores to re-weight pseudo labels and introduced a box jittering technique to select high-quality pseudo labels for the regression branch. Consistent Teacher [12] focused on inconsistencies during training and proposed a unified framework to handle inconsistencies in anchor assignment, feature alignment, and threshold processes. The above methods were all based on SPL. For SPL, various threshold-based techniques were employed to select reliable pseudo-labels, mostly conducted after complex post-processing steps like NMS. In contrast, Dense Teacher [11] introduced DPL and a region selection method to reduce noise and provide finer-grained supervision signals. DPL selects pixel-wise pseudo-labels, eliminating the aforementioned trouble. Moreover, the challenge of scale variation in SSOD has also drawn attention in recent years. PseCo [16] adopted a down-sampled view to make scale-invariant predictions. MixTeacher [13] adopted a similar approach but introduced a mixed view. However, these methods were all based on SPL, leaving DPL unexplored. Moreover, the aforementioned works focused on general object detection. This paper aims to improve the performance of semi-supervised oriented object detection.

C. Semi-Supervised Oriented Object Detection

Recently, SOOD [6] pioneered semi-supervised oriented object detection by introducing global consistency and adaptive weights based on the orientation gap between the teacher and student models, achieving excellent performance. DDPLS [31] introduced a density-guided selection method, achieving some improvement but lacking various dataset validations. PST [32] proposed a new framework called Pseudo-Siamese Teacher, in which two teacher models are used to generate high-quality pseudo annotations. Moreover, PST [33] applied a symmetric and bounded Jensen–Shannon divergence and scale-adaptive knowledge distillation to reduce the unreliability of pseudo annotations in localization, scale, and orientation, achieving significant improvement. Compared to these works, our method focuses on DPL, carefully handling the selection of DPL, and introduces a new multi-scale framework for DPL.

III. Method

A. Overview

As shown in Sec. I, we identify that a primary cause hindering the effectiveness of DPL-based methods is the imbalance between obtaining potential positive samples and removing the interference of inaccurate pseudo labels. To address this problem, we propose a DPL-based method called Denser Teacher. An overview of our method is shown in Fig. 2. Unlike previous DPL-based methods, we design a new DPL selection mechanism called Global Dynamic k Estimation (GDE) to adaptively select suitable DPLs in densely-distributed scenes, as shown in Sec. III-C. Compared with the previous DPL-based methods, GDE directly integrates potential object information into the dense pseudo-label selection process, addressing the dense distribution challenge inherent in oriented object detection. Additionally, scale variation has been widely explored in semi-supervised object detection in recent years. Existing works have demonstrated that incorporating an extra down-sampled view of the unlabeled image and regularizing the network with consistency constraints at either the feature level or label level can significantly improve performance [16], [19]. However, few works focus on the scale variation of DPL. Since the scale variance problem represents a significant challenge in oriented object detection, we propose a new multi-scale learning framework called Dense Multi-Scale Learning (DMSL) for DPL, which will be detailed in Sec. III-D. Moreover, we offer preliminary in Sec. III-B.

B. Preliminary

In semi-supervised oriented object detection, a model is trained with a labeled set $D_l = \{(X_i^l, Y_i^l)|_{i=1}^{N_l}\}$ and an unlabeled image set $D_u = \{X_i^u|_{i=1}^{N_u}\}$, where N_l and N_u are the numbers of labeled and unlabeled data, respectively. For each labeled image X_i^l, the annotation Y_i^l consists of a set of rotated boxes and corresponding category labels for the instances that appear in the image. Following common practice in previous work [6], [10], [16], we adopt the pseudo-labeling framework under the teacher-student paradigm as our basic training framework. Specifically, the training images are sampled from both labeled and unlabeled datasets, and the overall objective comprises these two parts to update the student model. Due to the lack of ground truth in unlabeled images, the teacher model provides pseudo labels for the student, whose weights are updated by the exponential moving average of the student model.

$$\theta_{t+1}^T = (1-\lambda)\theta_t^S + \lambda\theta_t^T \tag{1}$$

where θ^T and θ^S denote the parameters of the teacher model and student model, respectively, and the subscript denotes the training iteration. λ is the momentum to maintain the difference between the teacher model and student model.

In every training iteration, the training objective on labeled data follows a regular manner, fully supervised by the ground truth labels. For the unlabeled data, the teacher model first generates pseudo labels on a weakly augmented view of the image, which provides supervision signals for a strongly augmented view of the image for the student model. Subsequently, the student model is updated with the objective from the labeled data and a strongly augmented view of the image with pseudo labels. The overall training objective can be formulated as:

$$\mathcal{L} = \mathcal{L}_s + \alpha\mathcal{L}_u \tag{2}$$

where $\mathcal{L}_s$ and $\mathcal{L}_u$ denote the supervised loss of labeled images and the unsupervised loss of unlabeled images, respectively. α controls the contribution of the unsupervised loss.

C. Global Dynamic K Estimation

In the dense pseudo-labeling framework, the selection of DPLs is a key problem. While DPLs contain rich information, they also contain noise. In Dense Teacher [11], the selection process relies on a fixed selection ratio determined by dataset analysis. While this global approach may be effective for datasets like COCO [34], where object distribution is relatively uniform, it may not be sufficient for scenarios with extreme imbalanced distribution, such as in aerial images. SOOD [6] uses SPLs as the basis for selection, employing random sampling to select reliable DPLs, but its performance is thus limited by the SPLs. Moreover, Fig. 1 shows that these methods still struggle with the abundance of low-quality DPLs or inefficiency in finding potential DPLs.

To alleviate this problem, inspired by OTA [35], we build a simple yet effective selection mechanism called Global Dynamic K Estimation (GDE), where we carefully handle the dense distribution challenge of oriented object detection. In OTA [35], the IoU values over the candidate bag are summed up to represent the number of positive samples. For labeled data, the candidate bag is based on ground truth, which is missing in unlabeled data. Moreover, the IoU calculated between the prediction and pseudo label is inaccurate. As we cannot obtain accurate local ground truth, we seek an approximate method to estimate the positive samples or DPLs from the entire image. Intuitively, the number of DPLs selected for an image varies. Many factors can affect the selection, such as object distribution, object size, and occlusion conditions. It is difficult to build a function that could take all of these factors into consideration, especially in unlabeled data. Therefore, in GDE, we roughly estimate the number of DPLs in an image according to the dense predictions. Specifically, for an image, we sum up the classification scores of dense predictions and represent the estimated quantity of DPLs as K. We define K as:

$$K = \sum_{l=1}^{M}\sum_{i=1}^{W_l}\sum_{j=1}^{H_l} S_{lij} \tag{3}$$

$$S_{lij} = \max_c y_{lij,c} \tag{4}$$

where $y_{lij,c}$ is the probability of category c in the l-th Feature Pyramid Network (FPN) layer at location (i, j) in the corresponding feature map. M is the number of FPN layers. As a result, the DPLs are selected as follows:

$$\vec{d_{lij}} = \begin{cases} 1, \text{if} \quad S_{lij} \quad \text{in} \quad \text{top} \quad K, \\ 0, \text{otherwise} \end{cases} \tag{5}$$

where $\vec{d_{lij}}$ is the symbol deciding the selection of a DPL in the l-th FPN layer at location (i, j) in the corresponding feature map. Note that we round down the K in practice. Moreover, GDE can be directly applied to anchor-free detectors like FCOS [36]. An empirical study in Fig. 3 demonstrates our hypothesis. The estimated K has a positive correlation with the relative number of pseudo labels selected. GDE's adaptive mechanism directly incorporates potential object information into the dense pseudo-label selection process, carefully selecting suitable DPLs in densely distributed scenes. This approach effectively addresses the dense distribution challenge inherent in oriented object detection, enhancing efficiency compared to previous methods and addressing an issue largely overlooked by prior works.

After selecting suitable DPLs, to handle continuous values (values between 0 and 1), we use Quality Focal Loss [37] as the classification objective for unlabeled data. Let y^T and y^S denote the teacher's and student's predictions of the classification head. We calculate the classification loss as:

$$\mathcal{L}_u^{cls} = -|y^T - y^S|^{\gamma} \times [y^T log(y^S) + (1-y^T) log(1-y^S)] \tag{6}$$

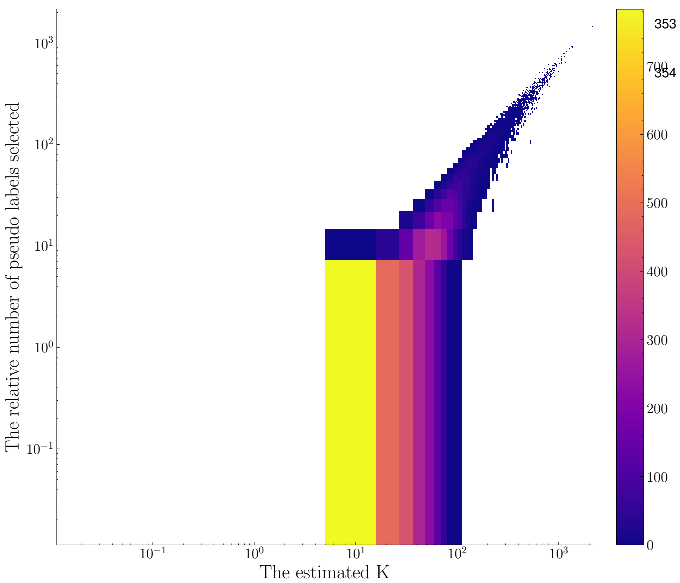


Fig. 3. The correlation between the estimated K and the relative number of pseudo labels selected under the DOTA-v1.5 10% partially labeled setting. Relative number indicates the sum of confidence of pseudo labels selected.

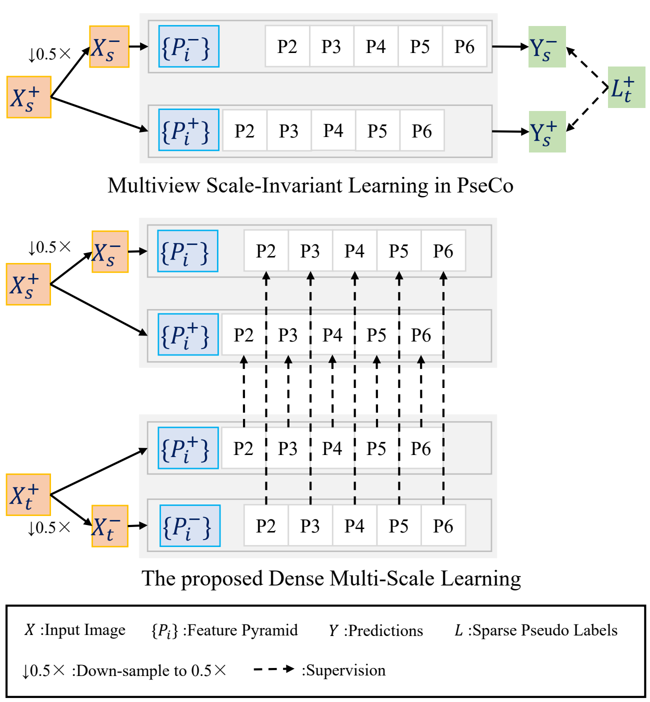


Fig. 4. Comparison of multi-scale learning between SPL-based methods (PseCo [16] as example) and our DPL-based Dense Multi-Scale Learning.

where γ is the suppression factor. For the regression head and auxiliary head (like the centerness branch in FCOS [36]), we employ Smooth L1 Loss [38], following the employment in SOOD [6]. Thus, the overall loss of unlabeled data is:

$$\mathcal{L}_u = \mathcal{L}_u^{cls} + \mathcal{L}_u^{reg} + \mathcal{L}_u^{aux} \tag{7}$$

where, $\mathcal{L}_u^{reg}$ and $\mathcal{L}_u^{aux}$ represent the regression loss and auxiliary loss, respectively, for the unlabeled data.

D. Dense Multi-Scale Learning

Scale variation across object instances remains a key challenge in object detection tasks [39], [40], especially in oriented object detection in aerial images. Despite the remarkable progress made by modern detection models, this challenge is particularly evident in the semi-supervised setting. Existing works have demonstrated that incorporating an extra down-sampled view of the unlabeled image and regularizing the network with consistency constraints can improve the performance of semi-supervised object detection [16]. However, previous works [13], [16] mainly focus on methods based on the SPL framework, where label-level scale learning is easy to deploy. However, for DPL, as the DPLs are selected from the original output of the model without any post-processing method, it is difficult to build label-level scale learning, as shown in Fig. 4. For DPL-based methods, SED [41] and DSL [19] construct a distillation method to utilize multi-scale information where all the original outputs are used without a selection process. While DPL contains rich information, it also retains many low-scoring predictions due to the absence of a threshold operation. Since those low-scoring predictions usually involve the background regions, the knowledge encompassed in them is intuitively less informative. Previous works find that learning to mimic the teacher's response in those regions hurts performance [11]. As far as we are aware, no existing work focuses on directly building multi-scale learning for DPLs.

Based on the above observation, to mitigate scale variance in semi-supervised oriented object detection, we propose a new framework called Dense Multi-Scale Learning (DMSL) for DPLs, which also leverages the down-sampled view but resorts to building a more convenient method for multi-scale DPL learning. Given an image, most detectors first extract multi-scale features P_i with decreasing spatial sizes, which constitute a feature pyramid $\mathbb{P}$. In the case of FPN, the spatial sizes of adjacent levels in the feature pyramid always differ by 2×, resulting in $P_2 - P_6$ layers with spatial sizes from $1/2^2$ to $1/2^6$ with respect to the size of the input image.

In this work, we first extract two feature pyramids from the regular view and the down-sampled view of the input image, denoted as $\mathbb{P}^+ = \{P_2^+, ..., P_6^+\}$ and $\mathbb{P}^- = \{P_2^-, ..., P_6^-\}$, respectively. Notice that with a 0.5× down-sample ratio, the network produces a small-scale feature pyramid. Unlike previous works, we utilize the down-sampled view of the teacher model and constrain the consistency across different scales. Consequently, the training objective for unlabeled data in Equation 7 extends to:

$$\mathcal{L}_u = \mathcal{L}_{u,+} + \mathcal{L}_{u,-} \tag{8}$$
$$= \mathcal{L}_{u,+}^{cls} + \mathcal{L}_{u,+}^{reg} + \mathcal{L}_{u,+}^{aux} + \mathcal{L}_{u,-}^{cls} + \mathcal{L}_{u,-}^{reg} + \mathcal{L}_{u,-}^{aux} \tag{9}$$

where $\mathcal{L}_{u,+}$ and $\mathcal{L}_{u,-}$ represent the loss for the unlabeled data in the regular view and down-sampled view, respectively. Through DMSL, DPLs with different scales are selected to build a more direct and effective way to improve scale adaptation.

TABLE I
EXPERIMENTAL RESULTS ON DOTA-V1.5 WITH PARTIALLY LABELED SETTING. THE BEST RESULTS ARE IN BOLD. - INDICATES THAT THE RESULT WAS NOT REPORTED IN THE LITERATURE. $*$ AND † INDICATE IMPLEMENTATIONS WITH ROTATED-FASTER R-CNN AND ROTATED-FCOS, RESPECTIVELY.

Setting	Methods	Partially Labeled Data				
		1%	5%	10%	20%	30%
Supervised	Faster RCNN [43]	13.22	33.95	43.43	51.32	53.14
Semi-supervised	Unbiased Teacher* [15]	-	-	44.51	52.80	53.33
	Soft Teacher* [10]	-	-	48.46	54.89	57.83
	PseCo* [16]	-	-	48.04	55.28	58.03
	DualPolish* [44]	-	-	49.02	55.17	58.44
	PST* [33]	-	41.39	49.63	57.39	60.40
Supervised	FCOS [36]	15.67	33.38	42.78	50.11	54.79
Semi-supervised	Dense Teacher† [11]	18.38	40.27	46.90	53.93	57.86
	SOOD† [6]	17.12	40.02	48.63	55.58	59.23
	Denser Teacher (Ours)†	**20.98**	**43.40**	**52.05**	**57.49**	**60.40**

IV. EXPERIMENT

A. Dataset and Evaluation Protocol

DOTA [14] is one of the largest datasets for oriented object detection in aerial scenes. We conducted experiments on DOTA-v1.5 and DOTA-v1.0. Compared to DOTA-v1.0, the images in DOTA-v1.5 remain unchanged, but there are additional annotations for small objects (less than 10 pixels) and an extra category, Container crane. These additional annotations for small objects make the dataset more challenging and better reflect the characteristics of real-world aerial imagery objects. Both DOTA-v1.5 and DOTA-v1.0 comprise 2,806 large-scale aerial images and are divided into three sets. The training set consists of 1,411 images, the validation set has 458 images, and the test set contains 937 images. We adopt the standard mean Average Precision (mAP) as the evaluation metric for the DOTA datasets.

DIOR-R [42] is a challenging dataset with oriented objects annotated on the DIOR dataset. The DIOR-R dataset includes 11,725 and 11,738 images as the trainval set and test set, respectively, with a uniform size of 800×800, covering 20 categories. We also adopt mAP as the evaluation metric for the DIOR-R dataset. Compared with DOTA dataset, the DIOR-R dataset carefully collects data with uniform size and thus features a more balanced distribution of object sizes and densities, with fewer extreme variations compared to the DOTA dataset, which contains a wider range of object sizes and more variable densities.

To be closer to the actual application scenario, we mainly consider a partially labeled setting to confirm the effectiveness of our proposed method on limited data.

DOTA Partially Labeled. In DOTA-v1.5, following SOOD [6], we randomly sample 10%, 20%, and 30% of images from the training set as labeled data and set the remaining images as unlabeled data. For each protocol, we provide a fold with a similar distribution as the training set to avoid distribution mismatching [33]. To further evaluate our method in more severe situations, we extend this setting to 1% and 5%. Note that in the 1% setting, only 14 images are provided as labeled data. For DOTA-v1.0, we use the same setting as in DOTA-v1.5.

DIOR-R Partially Labeled. Similarly to the setting in DOTA, we randomly sample 1%, 5%, 10%, 20%, and 30% of images from the trainval set of DIOR-R as labeled data and keep the remaining data as unlabeled data.

B. Implementation Details

We use Rotated-FCOS [36] as the base rotated object detector and ResNet-50 [45] with FPN [39] as the backbone. The implementation of the base detector follows the MMRotate framework [46].

DOTA Partially Labeled. The model is trained for 120k iterations on two NVIDIA RTX3090 GPUs with three images per GPU. We use SGD with the learning rate initialized to 0.0025. The weight decay and momentum are set to 0.0001 and 0.9, respectively. For a fair comparison, we set the data sample ratio between the labeled and unlabeled data to 2:1, following the setting in SOOD [6]. Following previous work [6], [33], we split the original images into 1024×1024 patches with a pixel overlap of 200 between adjacent patches.

DIOR-R Partially Labeled. We follow the same implementation as in DOTA.

We adopt the same asymmetric data augmentation used in SOOD [6]. Specifically, we use strong augmentation for the student model and weak augmentation for the teacher model. Strong augmentation includes random flipping, color jittering, random grayscale, and random Gaussian blur, while weak augmentation only includes random flipping. Following previous works [6], [12], we use the "burn-in" strategy to initialize the teacher model. For the α in Equation 2, which balances the contributions of the supervised and unsupervised losses, we initially adopt α to 1, following the prior work [6]. However, with the introduction of DMSL, which incorporates two unsupervised losses at different scales, we adjust α to 0.5. This adjustment ensures equal contributions from the supervised and unsupervised components, maintaining a balanced influence across all loss terms.

C. Main Results

In this section, we compare our method with SOTA semi-supervised oriented object detection methods [6], [33] and

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2024.3518452

TABLE II
EXPERIMENTAL RESULTS ON DOTA-V1.0 WITH PARTIALLY LABELED SETTING. THE BEST RESULTS ARE IN BOLD.

Setting	Methods	Partially Labeled Data				
		1%	5%	10%	20%	30%
Supervised	FCOS [36]	15.55	34.34	43.03	51.40	55.30
Semi-supervised	Dense Teacher [11]	**20.05**	42.57	49.53	55.76	58.07
	SOOD [6]	17.52	43.00	50.18	56.47	60.37
	Denser Teacher (Ours)	19.45	**45.84**	**52.62**	**59.20**	**62.82**

TABLE III
EXPERIMENTAL RESULTS ON DIOR-R WITH PARTIALLY LABELED SETTING. THE BEST RESULTS ARE IN BOLD.

Setting	Methods	Partially Labeled Data				
		1%	5%	10%	20%	30%
Supervised	FCOS [36]	19.33	37.45	43.66	47.96	52.23
Semi-supervised	Dense Teacher [11]	**26.98**	44.45	51.05	55.22	57.51
	SOOD [6]	25.02	41.56	48.18	52.61	55.47
	Denser Teacher (Ours)	26.88	**46.46**	**52.87**	**55.93**	**58.73**

re-implement some SOTA SSOD methods on oriented object detectors for reference. In the experiments, for a fair comparison, we apply the same augmentation settings in the re-implemented experiments.

1) Quantitative Analysis:

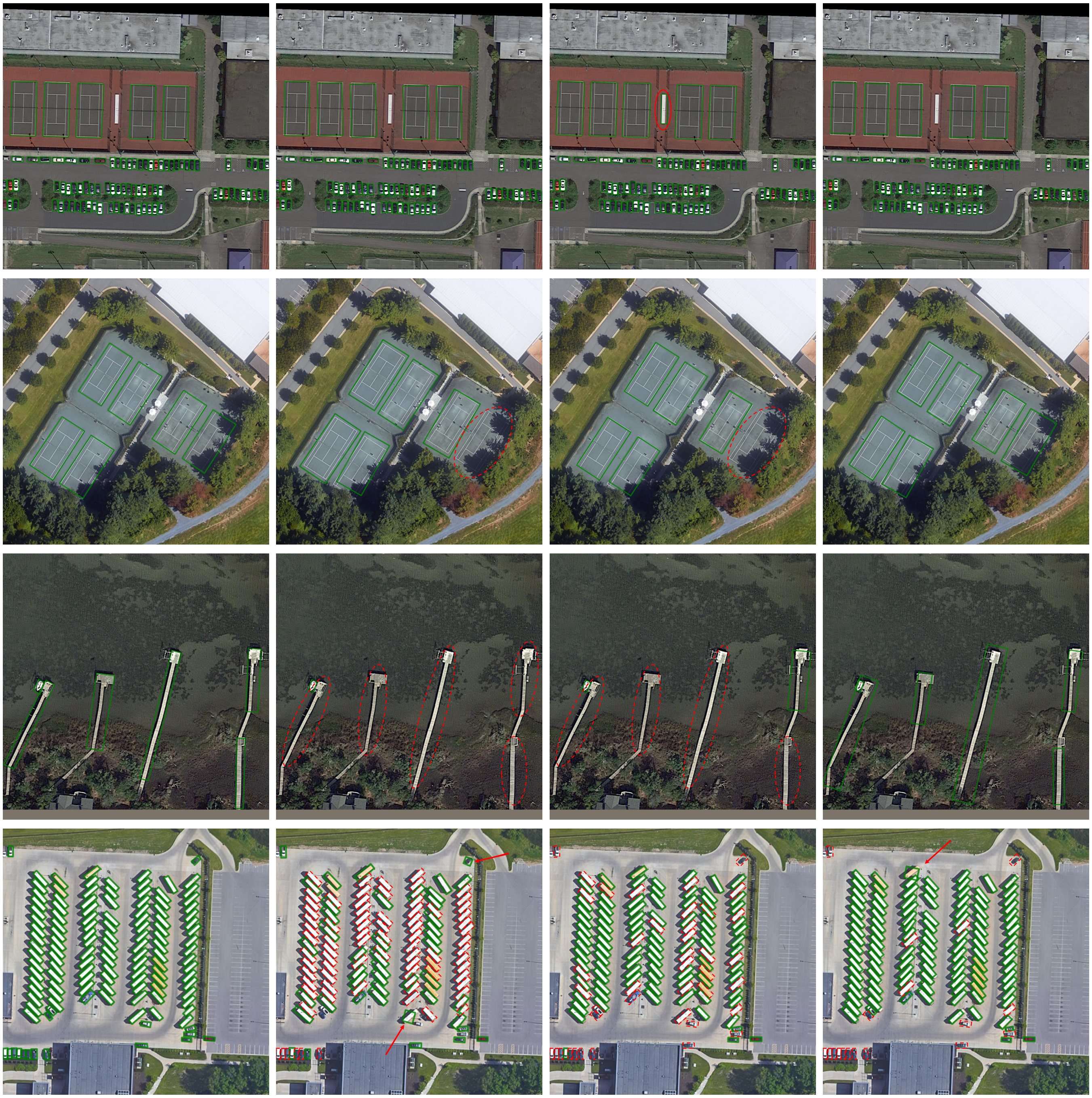
DOTA Partially Labeled. We compare our proposed Denser Teacher with existing SOTA methods on the DOTA-v1.5 and DOTA-v1.0 datasets. The results are shown in Table. I and Table. II. In the DOTA-v1.5 dataset, our method, Denser Teacher, achieves the best performance under the 1%, 5%, 10%, 20%, and 30% proportions, reaching 20.98 mAP, 43.40 mAP, 52.05 mAP, 57.49 mAP, and 60.40 mAP, respectively. This outperforms the supervised baseline by 5.31 points, 10.02 points, 9.27 points, 7.38 points, and 5.61 points, respectively. Similarly, our method also surpasses or equals the previous SOTA method PST [33], especially when labeled data are scarce. For example, it outperforms PST [33] by 2.01 points and 2.42 points in the 5% and 10% settings, confirming the effectiveness of our proposed method on severely limited data. Moreover, among DPL-based methods, our method also shows excellent performance and surpasses the SOTA method SOOD [6] by a large margin. Furthermore, we compare our proposed method, Denser Teacher, with re-implemented DPL-based methods in the DOTA-v1.0 dataset. As shown in Table. II, our proposed Denser Teacher achieves optimal performance in most settings, except in the 1% setting. Specifically, our method achieves a performance of 19.45 mAP, which is 0.60 points behind Dense Teacher [11]. In other settings, our method clearly exceeds previous DPL-based methods, showing outstanding performance in semi-supervised oriented object detection.

DIOR-R Partially Labeled. To further evaluate our method on various datasets, we compare our Denser Teacher method with re-implemented DPL-based methods on DIOR-R. The results are shown in Table. III. Our proposed Denser Teacher achieves the best performance in most cases. Specifically, it reaches 26.88 mAP, 46.46 mAP, 52.87 mAP, 55.93 mAP, and 58.73 mAP under the 1%, 5%, 10%, 20%, and 30% labeled data settings, surpassing the supervised baseline by 7.55 points, 9.01 points, 9.21 points, 7.97 points, and 6.5 points, respectively. Compared with SOOD [6], our method shows a significant improvement across different data ratios. Moreover, we notice that Dense Teacher [11] also surpasses SOOD [6] by a large margin. We conjecture that since the object distribution in DIOR-R is not as extreme as in DOTA, the disadvantage of using a fixed selection ratio is greatly compensated, resulting in similar performance. Nevertheless, our method still surpasses Dense Teacher in most settings, showing great adaptation in changeable scenarios.

We observed that on the DOTA-v1.0 and DIOR-R datasets, our method significantly outperforms the SOOD and FCOS methods under the 1% labeled data setting. Although its accuracy is slightly lower than that of the Dense Teacher method, the performance remains comparable. This minor performance gap may be attributed to the relatively weaker base detector used in our framework, which struggles to distinguish between foreground and background regions effectively. This challenge indirectly impacts the quality of DPLs generated during training. Despite this, our method achieves optimal performance in all other labeled data settings, demonstrating its effectiveness and robustness for semi-supervised oriented object detection across various datasets.

2) Qualitative Analysis: Fig. 5 presents a qualitative comparison between Denser Teacher and other SOTA methods. We find that our method excels in detecting multi-scale dense objects, indicating that multi-scale object information and abundant supervision signals are effectively learned. Moreover, the visual results demonstrate that introducing the proposed mechanism significantly reduces false negatives (dashed circles) and false positives (solid circles), indicating more robust learning of the objects and a significant contribution to semi-supervised oriented object detection.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2024.3518452



(a) Ground Truth (b) Supervised Baseline (c) SOOD (d) Ours

Fig. 5. Some visualization examples from the DOTA-v1.5 dataset. The green rectangles indicate predictions. The red dashed circle, solid red circle, and red arrow represent false negatives, false positives, and inaccurate orientation predictions, respectively.

D. Ablation Study

In this section, we conduct ablation experiments to validate our key designs. Unless specified, all ablation experiments are performed under the 10% partially labeled setting in DOTA-v1.5.

1) Component Analysis: The contributions of different components of our proposed Denser Teacher are listed in Table. IV. In the DOTA-v1.5 10% partially labeled setting, the Rotated FCOS supervised baseline achieves 42.97 mAP.

TABLE IV
COMPONENT ANALYSIS OF THE PROPOSED METHOD.

Methods	GDE	DMSL	DOTA-1.5	DIOR-R
Supervised	-	-	42.97	43.66
Denser Teacher	✓	-	51.00	52.15
	✓	✓	52.05	52.87

TABLE V
COMPARISONS OF DIFFERENT DENSE PSEUDO-LABEL SELECTION METHODS.

	Selection Strategies	mAP
I	Learning Region	46.90
II	Instance-level DPL	47.18
III	GDE (Ours)	51.00

(a) Ground Truth (b) Dense Teacher

(c) SOOD (d) Denser Teacher (Ours)

Fig. 6. Visualization of different dense pseudo-label selection methods. Different color represents different category. Note that in SOOD [6], dense pseudo-labels are selected by random sampling in the prediction of teacher model filtered by fixed threshold 0.5 .

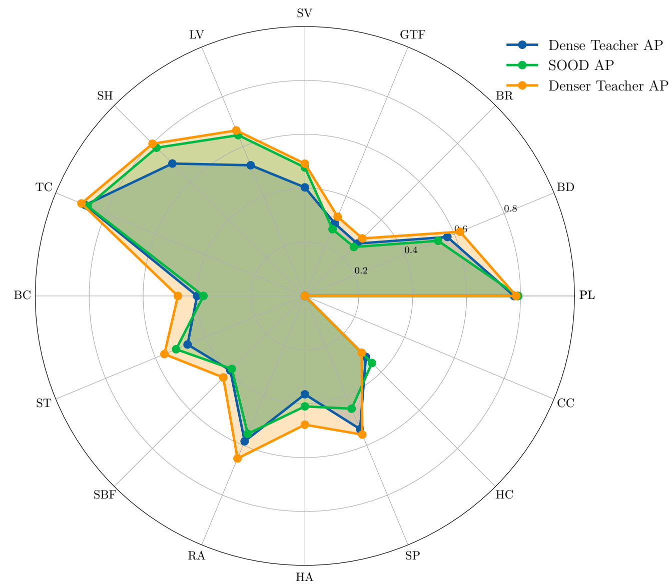


Fig. 7. Class-wise AP in different methods. Plane (PL), Baseball Diamond (BD), Bridge (BR), Ground Track Field (GTF), Small Vehicle (SV), Large Vehicle (LV), Ship (SH), Tennis Court (TC), Basketball Court (BC), Storage Tank (ST), Soccer-Ball Field (SBF), Roundabout (RA), Harbor (HA), Swimming Pool (SP), Helicopter (HC), and Container Crane (CC).

By using GDE, the performance can be significantly improved from 42.97 to 51.00 mAP, already surpassing the SOTA method in Table. I. By adopting DMSL, the performance can be further improved to 52.05 mAP, indicating that the model becomes more robust and has higher accuracy. Similarly, in the DIOR-R 10% partially labeled setting, by using GDE, the performance can be significantly improved from 43.66 to 52.15. By adopting DMSL, the performance can be further improved to 52.87 mAP. The ablation studies in Table. IV verify the effectiveness of each module in Denser Teacher in various dataset.

2) Comparisons of Different Dense Pseudo-Label Selection Methods: The selection of DPL is one of the key components of DPL-based methods. To further verify the effectiveness of our proposed selection method, we conduct a comparison of different selection methods, including: the learning region used in Dense Teacher [11], the instance-level DPL selection method used in SOOD [6], and our GDE. For a fair comparison, we remove the other components in the methods. The results are shown in Table. V. We also provide a visualization of the selection results of different methods in Fig. 6. Dense Teacher [11] involves a learning region strategy based on Feature Richness Score [47], but requires a static hyper-parameter to control the number of selections, resulting in deficient DPL selection. Such a design also brings challenges in complex scenarios where extreme distribution is common. SOOD [6] improves the quality of DPLs by randomly sampling from the SPLs. This helps the model concentrate on high-quality supervision but makes its performance highly dependent on the results of SPLs, which have been confirmed to be sensitive in complex scenes [11]. In Fig. 6, we present the SPLs used in SOOD [6], filtered by a fixed threshold of 0.5, to provide an intuitive understanding. In contrast, our proposed selection method, GDE, shows an obvious advantage in selecting suitable DPLs, and thus achieves the best performance gain. This demonstrates the effectiveness of our proposed selection method in complex scenes, clearly setting it apart from existing methodologies.

3) Multi-scale Learning: DMSL provides a straightforward and effective approach to achieving multi-scale learning, and it is distinctly differentiated from existing methodologies through the incorporation of DPL. To further demonstrate our method's effectiveness in multi-scale learning, we select several representative categories, including Ship (SH), Plane (PL), Small Vehicle (SV), Large Vehicle (LV), Harbor (HA), Swimming Pool (SP), and Basketball Court (BC), and report the results of our method. We also re-implement some DPL-based methods for reference. Results are shown in Table. VI, where our method shows significant improvement compared with the supervised baseline in all selected categories. For small objects like ships, small vehicles, and large vehicles, our method shows great improvement compared with SOOD [6]. In fact, our method surpasses SOOD in all selected categories except for the Plane. SOOD has a slight improvement in this category. To better demonstrate our method's multi-scale

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2024.3518452

TABLE VI
THE PERFORMANCE OF THE PROPOSED DENSER TEACHER AND OTHER DPL-BASED METHODS ON SEVERAL REPRESENTATIVE CATEGORIES IN THE VALIDATION SET OF DOTA-V1.5. THE BEST RESULTS ARE IN BOLD.

Setting	Methods	SH	PL	SV	LV	HA	SP	BC	mAP
Supervised	FCOS [36]	0.779	0.782	0.429	0.578	0.308	0.493	0.295	42.97
Semi-supervised	Dense Teacher [11]	0.695	0.776	0.403	0.525	0.365	0.535	0.400	46.78
	SOOD [6]	0.778	**0.791**	0.477	0.646	0.410	0.453	0.376	47.93
	Denser Teacher (Ours)	**0.799**	0.784	**0.491**	**0.665**	**0.478**	**0.557**	**0.471**	**52.05**

TABLE VII
EXTENSION TO QUERY-BASED BACKBONE. EXPERIMENTS ARE CONDUCTED AT 10% SETTING.

Setting	Method	mAP
Supervised	FCOS [36]	43.34
Semi-supervised	Dense Teacher [11]	46.46
	SOOD [6]	47.73
	Denser Teacher (Ours)	48.99

TABLE VIII
TIME COST ANALYSIS. EXPERIMENTS ARE CONDUCTED AT 10% SETTING.

Setting	Method	mAP	Seconds
Supervised	FCOS [36]	42.97	0.20
Semi-supervised	Dense Teacher [11]	46.90	0.36
	SOOD [6]	48.63	0.54
	Denser Teacher (Ours)	52.05	0.59

learning ability, we visualize the class-wise AP of our method and other DPL-based methods in Fig. 7. The results show that our method significantly improves scale adaptation, thus achieving better performance.

4) Extension to other backbone: We also validate the effectiveness of the proposed method on other backbone. Specifically, we take Swin Transformer [48] as backbone, and implement our proposed method under the same experimental setting. We also implement Dense Teacher and SOOD as comparison. As shown in Table. VII, when extending to query-based backbone, our proposed method still achieves obvious improvement, showing great effectiveness.

5) Time cost analysis: We report the time cost analysis of our method. Moreover, Dense Teacher [11] and SOOD [6] are also evaluated for comparison. The results are shown in Table. VIII. Our proposed method slightly increases the computational cost compared with SOOD but achieves obvious performance improvement. Moreover, our proposed method adopts teacher model for inference and thus no extra computational expense is introduced compared to the base model in the inference stage.

V. DISCUSSION

Our method demonstrates strong performance in semi-supervised oriented object detection, particularly in addressing multi-scale learning challenges with the novel DMSL framework tailored for DPLs, which has been largely overlooked in previous works. However, its usage of the distinctive characteristics of aerial objects remains limited. Specifically, our method primarily leverages the dense distribution and multi-scale characteristics of aerial objects, which contribute to its success. However, other distinctive characteristics, such as large scale ratios and complex backgrounds, are not explicitly addressed, potentially limiting the method's applicability in more diverse aerial scenarios. Moreover, the proposed GDE method might face limitations in scenarios with sparse distributions, as shown in Fig. 3. The results indicate that the estimation of K becomes less accurate in such scenes, potentially affecting overall model performance. Future work could explore adaptive strategies to enhance GDE's robustness in handling sparse or heterogeneous distributions.

VI. CONCLUSION

In this paper, we analyze the shortcomings of existing DPL-based methods in semi-supervised oriented object detection and identify that these methods suffer from an imbalance in obtaining potential positive samples and removing the interference of inaccurate pseudo labels. To overcome this problem, we introduce Denser Teacher, a novel method for semi-supervised oriented object detection. In Denser Teacher, we propose Global Dynamic K Estimation (GDE) to leverage the information of potential objects to guide the selection of DPLs in densely-distributed scene and mitigate scale variance by introducing Dense Multi-Scale Learning (DMSL). Through these designs, our Denser Teacher achieves significant improvements compared with the SOTA methods. Extensive experiments demonstrate the effectiveness of our proposed method.

REFERENCES

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

[2] J. Yi, P. Wu, B. Liu, Q. Huang, H. Qu, and D. Metaxas, "Oriented object detection in aerial images with box boundary-aware vectors," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2150–2159.

[3] Y. Li, Q. Hou, Z. Zheng, C. Ming-Ming, Y. Jian, and L. Xiang, "Large selective kernel network for remote sensing object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 16 794–16 805.

[4] L. Dai, H. Liu, H. Tang, Z. Wu, and P. Song, "Ao2-detr: Arbitrary-oriented object detection transformer," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[5] Z. Guo, X. Zhang, C. Liu, X. Ji, J. Jiao, and Q. Ye, "Convex-hull feature adaptation for oriented and densely packed object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5252–5265, 2022.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2024.3518452

[6] W. Hua, D. Liang, J. Li, X. Liu, Z. Zou, X. Ye, and X. Bai, "Sood: Towards semi-supervised oriented object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 558–15 567.

[7] Y. Tang, Z. Cao, Y. Yang, J. Liu, and J. Yu, "Semi-supervised few-shot object detection via adaptive pseudo labeling," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[8] L. Yang, X. Zhang, J. Li, L. Wang, M. Zhu, C. Zhang, and H. Liu, "Mix-teaching: A simple, unified and effective semi-supervised learning framework for monocular 3d object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[9] D. Zhang, J. Han, G. Guo, and L. Zhao, "Learning object detectors with semi-annotated weak labels," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 12, pp. 3622–3635, 2018.

[10] M. Xu, Z. Zhang, H. Hu, J. Wang, L. Wang, F. Wei, X. Bai, and Z. Liu, "End-to-end semi-supervised object detection with soft teacher," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3060–3069.

[11] H. Zhou, Z. Ge, S. Liu, W. Mao, Z. Li, H. Yu, and J. Sun, "Dense teacher: Dense pseudo-labels for semi-supervised object detection," in *Proceedings of the European Conference on Computer Vision*. Springer, 2022, pp. 35–50.

[12] X. Wang, X. Yang, S. Zhang, Y. Li, L. Feng, S. Fang, C. Lyu, K. Chen, and W. Zhang, "Consistent-teacher: Towards reducing inconsistent pseudo-targets in semi-supervised object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3240–3249.

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

[14] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3974–3983.

[15] Y.-C. Liu, C.-Y. Ma, Z. He, C.-W. Kuo, K. Chen, P. Zhang, B. Wu, Z. Kira, and P. Vajda, "Unbiased teacher for semi-supervised object detection," *International Conference on Learning Representations*, pp. 1–17, 2021.

[16] G. Li, X. Li, Y. Wang, Y. Wu, D. Liang, and S. Zhang, "Pseco: Pseudo labeling and consistency training for semi-supervised object detection," in *Proceedings of the European Conference on Computer Vision*. Springer, 2022, pp. 457–472.

[17] C. Zhang, T. Liu, and K.-M. Lam, "Angle tokenization guided multi-scale vision transformer for oriented object detection in remote sensing imagery," in *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2022, pp. 3063–3066.

[18] R. Zhang, C. Xu, F. Xu, W. Yang, G. He, H. Yu, and G.-S. Xia, "Rethinking scale imbalance in semi-supervised object detection for aerial images," *arXiv preprint arXiv:2310.14718*, 2023.

[19] B. Chen, P. Li, X. Chen, B. Wang, L. Zhang, and X.-S. Hua, "Dense learning based semi-supervised object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4815–4824.

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

[21] X. Yang, J. Yan, Z. Feng, and T. He, "R3det: Refined single-stage detector with feature refinement for rotating object," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3163–3171.

[22] J. Han, J. Ding, N. Xue, and G.-S. Xia, "Redet: A rotation-equivariant detector for aerial object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2786–2795.

[23] X. Yang, J. Yan, Q. Ming, W. Wang, X. Zhang, and Q. Tian, "Rethinking rotated object detection with gaussian wasserstein distance loss," in *International Conference on Machine Learning*. PMLR, 2021, pp. 11 830–11 841.

[24] X. Yang, X. Yang, J. Yang, Q. Ming, W. Wang, Q. Tian, and J. Yan, "Learning high-precision bounding box for rotated object detection via kullback-leibler divergence," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 381–18 394, 2021.

[25] X. Yang, Y. Zhou, G. Zhang, J. Yang, W. Wang, J. Yan, X. Zhang, and Q. Tian, "The kfiou loss for rotated object detection," *arXiv preprint arXiv:2201.12558*, 2022.

[26] X. Yang and J. 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*. Springer, 2020, pp. 677–694.

[27] Y. Xu, M. Fu, Q. Wang, Y. Wang, K. Chen, G.-S. Xia, and X. Bai, "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 4, pp. 1452–1459, 2020.

[28] Y. Zeng, X. Yang, Q. Li, Y. Chen, and J. Yan, "Ars-detr: Aspect ratio sensitive oriented object detection with transformer," *arXiv preprint arXiv:2303.04989*, 2023.

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

[30] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in Neural Information Processing Systems*, vol. 33, pp. 596–608, 2020.

[31] T. Zhao, Q. Fang, S. Shi, and X. Xu, "Density-guided dense pseudo label selection for semi-supervised oriented object detection," *arXiv preprint arXiv:2311.12608*, 2023.

[32] W. Wu, H.-S. Wong, and S. Wu, "Pseudo-siamese teacher for semi-supervised oriented object detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[33] W. Wu, H.-S. Wong, and S. Wu, "Pseudo-siamese teacher for semi-supervised oriented object detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[34] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[35] Z. Ge, S. Liu, Z. Li, O. Yoshie, and J. Sun, "Ota: Optimal transport assignment for object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 303–312.

[36] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9627–9636.

[37] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 002–21 012, 2020.

[38] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[39] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[40] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection snip," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3578–3587.

[41] Q. Guo, Y. Mu, J. Chen, T. Wang, Y. Yu, and P. Luo, "Scale-equivalent distillation for semi-supervised object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 522–14 531.

[42] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 296–307, 2020.

[43] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[44] L. Zhang, Y. Sun, and W. Wei, "Mind the gap: Polishing pseudo labels for accurate semi-supervised object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 3463–3471.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[46] Y. Zhou, X. Yang, G. Zhang, J. Wang, Y. Liu, L. Hou, X. Jiang, X. Liu, J. Yan, C. Lyu *et al.*, "Mmrotate: A rotated object detection benchmark using pytorch," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 7331–7334.

[47] D. Zhixing, R. Zhang, M. Chang, S. Liu, T. Chen, Y. Chen *et al.*, "Distilling object detectors with feature richness," *Advances in Neural Information Processing Systems*, vol. 34, pp. 5213–5224, 2021.

[48] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

Tong Zhao received the B.S. degree in Automation from Central South University. He is currently pursuing a Master's degree in Control Science and Engineering at the National University of Defense Technology. His research interests include semi-supervised learning and target detection in remote sensing images.

Fang Qiang received the B.S. degree in Automation from Xidian University, Xi'an, China, in 2007. He then earned his M.S. and Ph.D. degrees in Control Science and Engineering from the National University of Defense Technology, Changsha, China, in 2009 and 2013, respectively. Currently, he is an Associate Professor at the College of Intelligence Science and Engineering, National University of Defense Technology. His research interests include robotics and unmanned aerial vehicles, with a focus on object detection, reinforcement learning, and 6D pose estimation.

Xin Xu (Senior Member, IEEE) received the BS degree in electrical engineering from the Department of Automatic Control, National University of Defense Technology, and the PhD degree in control science and engineering from the College of Mechatronics and Automation, National University of Defense Technology. He has been a visiting scientist for cooperation research in the Hong Kong Polytechnic University, University of Alberta, and the University of Strathclyde, respectively. Currently, he is a full professor and the director of the Department of Intelligent Science and Technology with the National University of Defense Technology. His main research areas include: reinforcement learning and intelligent vehicles, learning control, robotics and machine learning. He has coauthored four books and published more than 150 papers in international journals and conferences. He is an associate editor of Information Sciences, CAAI Transactions on Intelligence Technology, Acta Automatica Sinica, Intelligent Automation and Soft Computing. He has also been a guest editor of IEEE Transactions on System, Man and Cybernetics: Systems, International Journal of Adaptive Control and Signal Processing. He is one of the recipients received the 2nd class National Natural Science Award of China, in 2012, the 1st class Natural Science Award from Hunan Province, P. R. China, in 2009 and the Fork Ying Tong Youth Teacher Fund of China, in 2008. He is a committee member of the IEEE Technical Committee on Approximate Dynamic Programming and Reinforcement Learning and the IEEE Technical Committee on Robot Learning. He has served as a PC member or session chair in many international conferences.