SOOD: Towards Semi-Supervised Oriented Object Detection

Wei Hua[*1], Dingkang Liang[*1], Jingyu Li[1], Xiaolong Liu[1], Zhikang Zou[2], Xiaoqing Ye[2], Xiang Bai[†1]

[1]Huazhong University of Science and Technology, {whua_hust, dkliang, xbai}@hust.edu.cn
[2] Baidu Inc., China

Abstract

Semi-Supervised Object Detection (SSOD), aiming to explore unlabeled data for boosting object detectors, has become an active task in recent years. However, existing SSOD approaches mainly focus on horizontal objects, leaving multi-oriented objects that are common in aerial images unexplored. This paper proposes a novel Semi-supervised Oriented Object Detection model, termed SOOD, built upon the mainstream pseudo-labeling framework. Towards oriented objects in aerial scenes, we design two loss functions to provide better supervision. Focusing on the orientations of objects, the first loss regularizes the consistency between each pseudo-label-prediction pair (includes a prediction and its corresponding pseudo label) with adaptive weights based on their orientation gap. Focusing on the layout of an image, the second loss regularizes the similarity and explicitly builds the many-to-many relation between the sets of pseudo-labels and predictions. Such a global consistency constraint can further boost semi-supervised learning. Our experiments show that when trained with the two proposed losses, SOOD surpasses the state-of-the-art SSOD methods under various settings on the DOTA-v1.5 benchmark. The code will be available at https://github.com/HamPerdredes/SOOD.

1. Introduction

Sufficient labeled data is essential for fully-supervised object detection. However, the data labeling process is time-consuming and expensive. Recently, Semi-Supervised Object Detection (SSOD), where object detectors are learned from labeled data as well as easy-to-obtain unlabeled data, has attracted increasing attention. Existing SSOD methods [16, 24, 44, 50] mainly focus on detecting objects with horizontal bounding boxes in general scenes. Nevertheless, in more complex scenes, such as aerial scenes, objects usually need to be annotated with oriented bounding boxes. Considering the higher annotation cost of oriented boxes*, semi-supervised oriented object detection is worth studying.

(a) Arbitrary rotating objects (b) Small and dense objects

Figure 1. Arbitrary rotating (a), small and dense (b) objects are common in aerial scenes, which are often regularly arranged on the image. From a global perspective, this pattern indicates that an aerial can be regarded as a layout.

Compared with general scenes, the main characteristics of objects in aerial scenes (or aerial objects for short) are three-fold: arbitrary orientations, small scales, and agglomeration, as shown in Fig. 1. The mainstream SSOD methods are based on the pseudo-labeling framework [3, 35, 36] consisting of a teacher model and a student model. The teacher model, an Exponential Moving Average (EMA) of the student model at historical training iterations, generates pseudo-labels for unlabeled images. Thus, the student model can learn from both labeled and unlabeled data. To extend the framework to oriented object detection, we think the following two aspects need to be addressed: 1) As orientation is an essential property of multi-oriented objects, how to use the orientation information when guiding the student with pseudo-labels is critical. 2) As aerial objects are often dense and regularly distributed in an image, we can utilize the layout to facilitate the learning of each pair instead of treating them individually.

This paper proposes the first Semi-supervised Oriented

*Equal contribution. †Corresponding author.
Work done when Dingkang Liang was an intern at Baidu.

*Annotation cost of an oriented box is about 36.5% (86$ vs. 63$ per 1k at 2022.11) more than a horizontal box according to https://cloud.google.com/ai-platform/data-labeling/pricing.

Object Detection method, termed SOOD. Following [50], SOOD is built upon the dense pseudo-labeling framework, where the pseudo labels are filtered from the raw pixel-wise predictions (including box coordinates and confidence scores). The key design is two simple yet effective losses that enforce the instance-level and set-level consistency between the student's and the teacher's predictions.

To be specific, considering that the pseudo-label-prediction pairs are not equally informative, we propose the Rotation-aware Adaptive Weighting (RAW) loss. It utilizes the orientation gap of each pair, which reflects the difficulty of this sample in a way, to weight the corresponding loss dynamically. In this manner, we can softly pick those more useful supervision signals to guide the learning of the student. In addition, considering that the layout of an aerial image can potentially reflect components' overall status (e.g., objects' density and location distribution) and help the detection process, we propose the Global Consistency (GC) loss. It measures the similarity of the pseudo-labels and the predictions from a global perspective, which can alleviate the disturbance of noise in pseudo-labels and implicitly regularizes the mutual relations between different objects.

We extensively evaluate SOOD under various settings on DOTA-v1.5, a popular aerial object detection benchmark. Our SOOD achieves consistent performance improvement when using 10%, 20%, 30%, and full of labeled data, compared with the state-of-the-art SSOD methods (using the same oriented object detector). The ablation study also verifies the effectiveness of the two losses.

In summary, this paper makes an early exploration of semi-supervised learning for oriented object detection. By analyzing the distinct characteristics of oriented objects from general objects, we propose two novel loss functions to adapt the pseudo-label framework to this task. We hope that this work can provide a good starting point for semi-supervised oriented object detection and serve as a simple yet strong baseline for future research.

2. Related works

Semi-Supervised Object Detection. In the past few years, semi-supervised learning (SSL) [2, 32] has achieved impressive performance in image classification. These works leverage unlabeled data by using pseudo-label [9,15,17,42], consistency regularization [2,36,41], data augmentation [4, 31] and even adversarial training [26]. Compared to semi-supervised image classification, SSOD requires instance-level predictions and additional bounding boxes regression sub-task, which makes it more challenging. In [28, 51], pseudo-labels are assembled from predictions of data with different augmentations. CSD [14] only utilizes the horizontal flipping augmentation and applies consistency loss to constrain the model, but the weak augmentation limits its performance. STAC [33] trains an object detection with labeled data and generates pseudo-labels on unlabeled data with this detector offline. After that, some studies [24, 35, 44, 45] adopt EMA from Mean Teacher [36] to update the teacher model after each training iteration. ISMT [45] obtains more accurate pseudo-labels by fusing current pseudo-labels and history labels. Unbiased Teacher [24] replaces cross-entropy loss with focal loss [22] to solve the class-imbalance issue and filters pseudo-labels by threshold. Soft Teacher [44] uses the classification scores to adaptively weight the loss of each pseudo-box and proposes box jittering to select reliable pseudo-labels. Unbiased Teacher v2 [25] adopts an anchor-free detector and uses uncertainty predictions to select pseudo-labels for the regression branch. Dense Teacher [50] replaces post-processed instance-level pseudo-labels with dense pixel-level pseudo-labels, which successfully removes the influence of thresholds and post-processing hyper-parameters. However, none of these works are designed for oriented object detection in aerial scenes. This paper aims to fill this blank and offer a starting point for future research.

Orient Object Detection. Different from general object detectors [8, 23, 29, 30], oriented object detectors represent objects with Oriented Bounding Boxes (OBBs). Typical oriented objects include aerial objects and multi-oriented scene texts [12, 19, 20, 34]. In recent years, many oriented object detection methods have been proposed to boost the performance for this area. CSL [46] formulates the angle regression problem as a classification task to address the out-of-bound issue. R^3Det [47] predicts Horizontal Bounding Boxes (HBBs) at the first stage to improve detection speed and align the feature in the second stage to predict oriented objects. Oriented R-CNN [43] proposes a concise multi-oriented region proposal network and uses the midpoint offsets to represent arbitrarily oriented objects. ReDet [12] proposes a Rotation-equivariant detector to extract rotation-invariant from rotation-equivariant for accurate aerial object detection. Oriented RepPoints [18] proposes a quality assessment module and samples assignment scheme for adaptive points learning, which can obtain non-axis features from neighboring objects and neglect background noises. Different from the above works that focus on the supervised paradigm, this paper makes an early exploration of semi-supervised oriented object detection, which can reduce the annotation cost and boost detectors with unlabeled data.

3. Preliminary

In this section, we revisit the mainstream pseudo-labeling paradigm in SSOD and Monge-Kantarovich optimal transport theory [27] as preliminary.

3.1. Pseudo-labeling Paradigm

Pseudo-labeling frameworks [25, 44, 50] inherited designs from the Mean Teacher [36] structure, which consists

of two parts, i.e., a teacher model and a student model. The teacher model is an Exponential Moving Average (EMA) of the student model. They are learned iteratively by the following steps. 1) Generate pseudo-labels for the unlabeled data in a batch. The pseudo-labels are filtered from the teacher's predictions, e.g., the box coordinates and the classification scores. Meanwhile, the student makes predictions for both labeled and unlabeled data in the batch. 2) Compute loss for the student model's predictions. It consists of two parts, the unsupervised loss $\mathcal{L}_u$ and supervised loss $\mathcal{L}_s$. They are computed for the unlabeled data with the pseudo-labels and the labeled data with the ground truth (GT) labels, respectively. The overall loss $\mathcal{L}$ is the sum of them. 3) Update the parameters of the student model according to the overall loss. The teacher model is updated simultaneously in an EMA manner. In this way, based on the mutual learning mechanism, both models evolve as the training goes on.

Based on the sparsity of pseudo-labels, pseudo-labeling frameworks can be further categorized into sparse pseudo-labeling [25, 44] and dense pseudo-labeling [50], termed SPL and DPL, respectively. The SPL selects the teacher's predictions after the post-processing operations, e.g., non-maximum suppression and score filtering. It obtains sparse labels to supervise the student, e.g., bounding boxes and categories. The DPL directly samples the post-sigmoid logits predicted by the teacher, which are dense and informative. Compared with SPL, DPL bypasses those lengthy post-processing methods, reserving more details from the teacher than its pseudo-box counterpart.

3.2. Optimal Transport

The Monge-Kantorovich Optimal Transport (OT) [27] aims to solve the problem of simultaneously moving items from one set to another set with minimum cost. It has been widely explored in various computer vision tasks [1, 7, 39, 49]. The mathematical formulations of OT are described as follows in detail.

Let $\mathbb{X} = \{x_i | x_i \in \mathbb{R}^d\}_i^N$ and $\mathbb{Y} = \{y_j | y_j \in \mathbb{R}^d\}_j^N$ denote two sets of N d-dimensional vectors. Their discrete distributions $\hat{\mathbb{X}}$ and $\hat{\mathbb{Y}}$ are formulated as:

$$\hat{\mathbb{X}} = \textstyle\sum_i^N \hat{x}_i \delta_{f_i} \tag{1}$$

$$\hat{\mathbb{Y}} = \textstyle\sum_j^N \hat{y}_j \delta_{g_j}, \tag{2}$$

where $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ are the discrete probability vectors, δ is the Dirac delta function. Therefore, the OT cost is measured between these two probabilities, $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$. The possible transportation plans from $\hat{\mathbf{x}}$ to $\hat{\mathbf{y}}$ are formed as:

$$\boldsymbol{P} = \{\boldsymbol{p} \in \mathbb{R}^{N \times N} | \boldsymbol{p}\mathbf{1}_N = \hat{\mathbf{x}}, \boldsymbol{p}^T \mathbf{1}_N = \hat{\mathbf{y}}\}, \tag{3}$$

where $\mathbf{1}_N$ is an N-dimensional column vector whose values are all 1. The OT cost is then defined as:

$$\omega_{ot}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \min_{\boldsymbol{p} \in \boldsymbol{P}} \langle \mathbf{C}, \boldsymbol{p} \rangle, \tag{4}$$

where $\mathbf{C} \in \mathbb{R}^{N \times N}$ represents the cost matrix between two sets, $\langle \cdot \rangle$ represents inner product. In common, the OT problem is solved in its dual formulation

$$\begin{aligned} \mathcal{W}_{ot}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \max_{\boldsymbol{\lambda}, \boldsymbol{\mu} \in \mathbb{R}^N} \langle \boldsymbol{\lambda}, \hat{\mathbf{x}} \rangle + \langle \boldsymbol{\mu}, \hat{\mathbf{y}} \rangle, \\ s.t. \ \lambda_i + \mu_j \leq \mathbf{C}_{i,j}, \ \forall i, j, \end{aligned} \tag{5}$$

where $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are the solutions of the OT problem, which can be approximated in an iterative manner [5].

4. Method

Fig. 2 illustrates an overview of our proposed SOOD. Towards multi-oriented object detection in aerial images, we build our approach upon the popular dense pseudo-labeling framework [50], along with the Rotation-aware Adaptive Weighting (RAW) loss and the Global Consistency (GC) loss. In this section, we first describe the overall framework in Sec. 4.1. Then, we describe the key design of the proposed losses, RAW and GC, in the following Sec. 4.2 and Sec. 4.3, respectively.

4.1. The Overall Framework

Currently, the Dense Pseudo-Labeling (DPL) framework achieves the state-of-the-art in SSOD. Hence, we construct a DPL-based end-to-end baseline, including the supervised and unsupervised parts. For the supervised part, the student model is trained with labeled data in a regular manner. For the unsupervised part, we first obtain predicted boxes of the teacher after post-processing. These boxes indicate informative areas in the prediction map, where we randomly sample predictions, forming them as dense pseudo-labels P^t. Note that we also select the predictions P^s at the same correspondence positions from the student.

We use the oriented version of FCOS [37] as the teacher and student models. The basic unsupervised loss consists of three parts: classification loss, regression loss, and center-ness loss, corresponding to the output of FCOS. We adopt smooth l1 loss for regression loss, binary cross-entropy loss for classification and center-ness loss. Based on these losses, we first perform adaptive weighting on them through RAW and further measure the global consistency between the teacher and the student via GC.

4.2. Rotation-aware Adaptive Weighting Loss

Orientation is one essential property of an oriented object. As shown in Fig. 1, even if objects are dense and small, their orientations remain clear. Previous oriented object detection methods have already employed such a property by

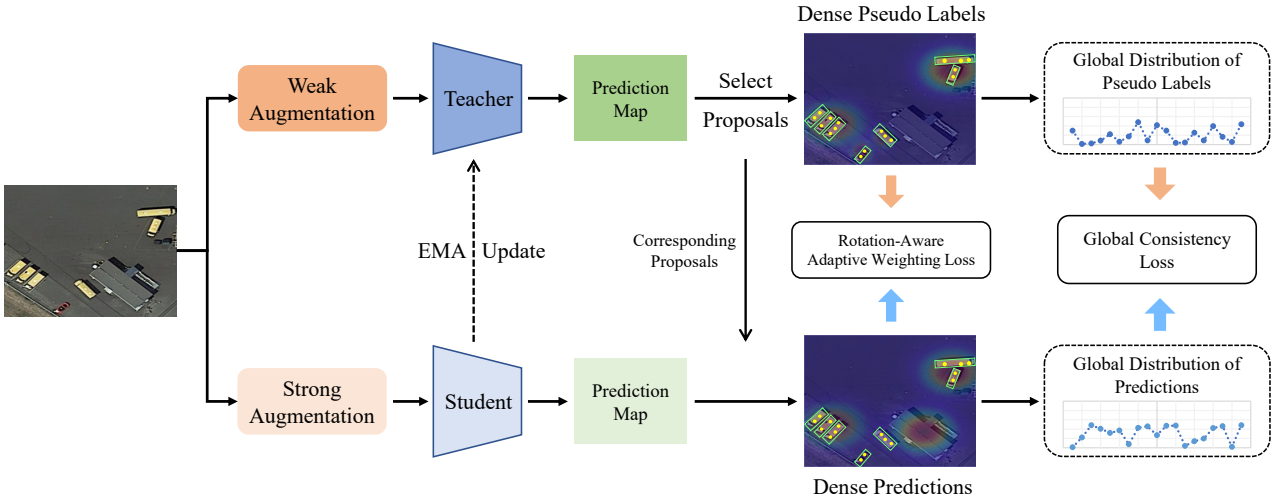

Figure 2. The pipeline of the proposed SOOD. Each training batch consists of both labeled and unlabeled images. Note that the supervised part is hidden for simplicity. For the unsupervised part, we sample dense pseudo-labels from the teacher's prediction map and select the student's predictions at the same positions, obtaining a series of pseudo-label-prediction pairs. We dynamically weigh each pair's unsupervised loss by their orientation difference. Besides, we regard them as two discrete distributions, measuring their similarity in a many-to-many manner via a global consistency loss.

assembling it into loss calculation. However, these works are under the assumption that the angles of the labels are reliable. In this case, it is natural to strictly force the prediction close to the ground truth.

Unfortunately, the above assumption does not hold in the semi-supervised setting. In other words, the pseudo-labels may be incorrect. Simply forcing the student to be close to the teacher may cause noise accumulation, harming the model's training process. Hence, we propose to utilize the orientation information softly. Intuitively, as orientation is essential but hard to be accurately predicted, the difference in rotation angles between a prediction and a pseudo-label can reflect the difficulty of the sample in a way. In other words, the orientation difference can be used to dynamically adjust the unsupervised loss. Therefore, we construct a rotation-aware modulating factor, similar to focal loss [22]. This factor can dynamically weight the loss of each pseudo-label-prediction pair by considering their orientation difference.

Specifically, the modulating factor ω_i^{rot} of the i-th pair is formed as:

$$\omega_i^{rot} = 1 + \sigma_i, \tag{6}$$

$$\sigma_i = \alpha \frac{|r_i^t - r_i^s|}{\pi}, r_i^t, r_i^s \in [-\frac{\pi}{2}, \frac{\pi}{2}), \tag{7}$$

where r_i^t and r_i^s are the i-th pseudo-label's and prediction's rotation angle in radians, respectively. α is a hyper-parameter for adjusting orientation's importance, and we set it to 50 empirically. We add a constant to σ_i, maintaining the origin unsupervised loss when pseudo-label and prediction have the same orientation. With the rotation-aware modulating factor, the overall rotation-aware adaptive weighting loss is formulated as:

$$\mathcal{L}_{RAW} = \sum_i^{N_p} \omega_i^{rot} \mathcal{L}_u^i, \tag{8}$$

where N_p is the number of pseudo-labels and $\mathcal{L}_u^i$ is the basic unsupervised loss of the i-th pseudo-label-prediction pair. By using the rotation-aware modulating factor, the RAW loss makes better use of the orientation information and provides more informative guidance, potentially benefiting the semi-supervised learning process.

4.3. Global Consistency Loss

Objects in an aerial image are usually dense and regularly distributed, as depicted in Fig. 1. Similar to texts in a document, the arrangement of the set of objects, i.e., the layout, encodes the mutual relations between them and the global pattern of the image. Ideally, the layout consistency between the student's and the teacher's predictions will be ensured if each pseudo-label-prediction pair is aligned. However, the latter condition is too strict and may hurt performance when there are noises in pseudo-labels. Therefore, it is reasonable to add the consistency between layouts as an additional *relaxed* optimization objective, encouraging the student to learn robust information from the teacher. In this way, the noise disturbance in pseudo-labels can be alleviated. Besides, the relations between different predicted instances from the student can also be regularized implicitly, which provides an additional guide to the student.

We introduce the optimal transport cost [38] to measure the global similarity of layouts between the teacher's and the student's predictions, forming the global consistency loss. To be concrete, we denote the classification scores predicted by the teacher and the student by $\mathbf{s}^t \in \mathbb{R}^{N_p \times K}$ and $\mathbf{s}^s \in \mathbb{R}^{N_p \times K}$ respectively, where K is the number of classes. Then, their global distributions, $\mathbf{d}^t \in \mathbb{R}^{N_p}$ and $\mathbf{d}^s \in \mathbb{R}^{N_p}$ can be formulated by

$$\mathbf{d}_i^t = e^{\mathbf{s}_{i,c(i)}^t}, \tag{9}$$

$$\mathbf{d}_i^s = e^{\mathbf{s}_{i,c(i)}^s}, \tag{10}$$

where $c(i) = \underset{j=1,\dots,K}{\arg\max}\, \mathbf{s}_{i,j}^t$ is the index of the class with the largest score for the i-th pseudo-label.

The global consistency loss is defined as the OT problem's dual formulation

$$\mathcal{L}_{GC}(\mathbf{d}^t, \mathbf{d}^s) = \left\langle \boldsymbol{\lambda}^*, \frac{\mathbf{d}^t}{\|\mathbf{d}^t\|_1} \right\rangle + \left\langle \boldsymbol{\mu}^*, \frac{\mathbf{d}^s}{\|\mathbf{d}^s\|_1} \right\rangle, \tag{11}$$

where we normalize two distributions to form discrete probabilities, by dividing them to their sum. To construct the cost map for solving the OT problem, we consider both the spatial distance and the score difference of each possible matching pair. Specifically, for each prediction, we measure its matching cost with every pseudo-label as follows:

$$C_{i,j} = C_{i,j}^{dist} + C_{i,j}^{score}, \tag{12}$$

$$C_{i,j}^{dist} = \frac{\|\mathbf{z}_i^t - \mathbf{z}_j^s\|_2^2}{\max_{1<=a,b<=N_p} \|\mathbf{z}_a^t - \mathbf{z}_b^s\|_2^2}, \tag{13}$$

$$C_{i,j}^{score} = \frac{\|\mathbf{s}_{i,c(i)}^t - \mathbf{s}_{j,c(j)}^s\|_1}{\max_{1<=a,b<=N_p} \|\mathbf{s}_{a,c(a)}^t - \mathbf{s}_{b,c(b)}^s\|_1}, \tag{14}$$

where $\mathbf{z}_i^t$ and $\mathbf{z}_j^s$ are 2D coordinates of the i-th sample in the teacher and the j_{th} sample in the student.

We solve the OT problem by a fast Sinkhorn distances algorithm [5], obtaining the approximate solution $\boldsymbol{\lambda}^*$ and $\boldsymbol{\mu}^*$. Based on the defined loss, its gradient with respect to $\mathbf{d}^s$ is

$$\frac{\partial \mathcal{L}_{GC}(\mathbf{d}^t, \mathbf{d}^s)}{\partial \mathbf{d}^s} = \frac{\boldsymbol{\mu}^*}{\|\mathbf{d}^s\|_1} - \frac{\langle \boldsymbol{\mu}^*, \mathbf{d}^s \rangle}{\|\mathbf{d}^s\|_1^2}. \tag{15}$$

The gradients can be back-propagated to update the model, enforcing the layout consistency in the framework. Although OT-based loss has been explored before [6,39,49], our goal of using OT is different. Specifically, they focus on utilizing OT to improve the model's generalizability [6, 39] or mitigate the matching constraint [49]. However, our GC aims to model the many-to-many relationship between the teacher and the student, which is complementary to the RAW. In addition, we adopt such a set-to-set matching to alleviate the error in pseudo-label assignment, providing a more loose but stable constraint.

SOOD is trained with the proposed unsupervised losses, RAW and GC, for unlabeled data as well as the supervised loss for labeled data. The overall loss $\mathcal{L}$ is defined as:

$$\mathcal{L} = \underbrace{\mathcal{L}_{RAW} + \mathcal{L}_{GC}}_{\mathcal{L}_u} + \mathcal{L}_s. \tag{16}$$

Note that the supervised loss is the same as defined in FCOS, our designs only modify the unsupervised part.

5. Experiments

We conduct experiments on DOTA-v1.5, which is proposed at DOAI-2019[†]. It contains 2806 large aerial images and 402,089 annotated oriented objects. It includes three subsets: DOTA-v1.5-train, DOTA-v1.5-val and DOTA-v1.5-test, containing 1411, 458, and 937 images, respectively. The annotations of DOTA-v1.5-test is not released.

There are 16 categories in this dataset: 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). Compared with DOTA-v1.0 [40], a previous version, DOTA-v1.5 contains more small instances (less than 10 pixels), which makes it more challenging.

Following conventions in SSOD, we consider two protocols, Partially Labeled Data and Fully Labeled Data, to validate the performance of a method on limited and abundant labeled data, respectively.

Partially Labeled Data. We randomly sample 10%, 20%, and 30% images from DOTA-v1.5-train as labeled data and set the remaining images as unlabeled data. Following DOTA-v1.5-train's data distribution, we provide one fold for each data proportion.

Fully Labeled Data. We set DOTA-v1.5-train as labeled data and DOTA-v1.5-test as unlabeled data.

For all experiments, we perform evaluation on DOTA-v1.5-val and report the performance with the standard mean average precision (mAP) as the evaluation metrics.

5.1. Implementation Details

Without loss of generality, we take FCOS [37] as the representative anchor-free detector, and adopt ResNet-50 [13] with FPN [21] as the backbone for all our experiments. Following the previous works [11, 12, 40], we crop the original images into 1024×1024 patches with a stride of 824, that is, the pixel overlap between two adjacent patches is 200. We utilize asymmetric data augmentation for unlabeled data. Specifically, we use strong augmentation for

[†]The 1st Workshop on Detecting Objects in Aerial Images in conjunction with IEEE CVPR 2019 https://captain-whu.github.io/DOAI2019/dataset.html

Table 1. Experimental results on DOTA-v1.5 under the Partially Labeled Data setting. * and † indicate our implementations with rotated-Faster-RCNN and rotated-FCOS, respectively. Experiments are conducted on 10%, 20% and 30% labeled data settings.

Setting	Method	Publication	10%	20%	30%
Supervised	Faster R-CNN* [30]	NeurIPS 2016	43.43	51.32	53.14
	FCOS† [37]	ICCV 2019	42.78	50.11	54.79
Semi-supervised	Unbiased Teacher* [24]	ICLR 2021	44.51	52.80	53.33
	Soft Teacher* [44]	ICCV 2021	48.46	54.89	57.83
	Dense Teacher† [50]	ECCV 2022	46.90	53.93	57.86
	SOOD† (**ours**)	-	**48.63**	**55.58**	**59.23**

Table 2. Experimental results on DOTA-v1.5 under the Fully Labeled Data setting. * and † indicate our implementations with rotated-Faster-RCNN and rotated-FCOS respectively. Numbers in font of the arrow indicate the supervised baseline.

Method	Publication	mAP
Unbiased Teacher* [24]	ICLR 2021	66.12 $\xrightarrow{-1.27}$ 64.85
Soft Teacher* [44]	ICCV 2021	66.12 $\xrightarrow{+0.28}$ 66.40
Dense Teacher† [50]	ECCV 2022	65.46 $\xrightarrow{+0.92}$ 66.38
SOOD† (**ours**)	-	65.46 $\xrightarrow{+2.24}$ **67.70**

Table 3. The effectiveness of SOOD on other methods under Fully Labeled Data setting. * means based on RetinaNet [22]. SOOD is able to generalize to other detectors and boost their performance.

Detector	Publication	Method	mAP
CFA [10]	CVPR 2021	Supervised	65.75
		Ours	**67.07**
KLD* [48]	NeurIPS 2021	Supervised	62.21
		Ours	**64.62**

the student model and weak augmentation for the teacher model. Random flipping is used for weak augmentation, while strong augmentation contains random flipping, color jittering, random grayscale, and random Gaussian blur. All models are trained for 180k iterations on 2 RTX3090 GPUs. With the SGD optimizer, the initial learning rate of 0.0025 is divided by 10 at 120k and 160k. The momentum and the weight decay are set to 0.9 and 0.0001, respectively. Each GPU takes 3 images as input, where the proportion between unlabeled and labeled data is set to 1:2. The pseudo-label sampling ratio is set as 0.25 by default. Following previous SSOD works [24, 50], we use the "burn-in" strategy to initialize the teacher model.

5.2. Main Results

In this section, we compare our method with the state-of-the-art SSOD methods [3, 24, 44] on DOTA-v1.5. For a fair comparison, we re-implement these methods on oriented object detectors with the same augmentation setting.

Partially Labeled Data. We evaluate our method under different labeled data proportions, and the results are shown in Tab. 1. Our SOOD achieves state-of-the-art performance under all proportions. Specifically, it obtains 48.63, 55.58, and 59.23 mAP on 10%, 20%, and 30% proportions, respectively, surpassing our supervised baseline by +5.85, +5.47, and +4.44 mAP. We also surpass the state-of-the-art anchor-free method Dense Teacher [50] by +1.73, +1.65, and +1.37 under various proportions. We provide two anchor-based methods for comparison, Unbiased Teacher [24] and Soft Teacher [44]. On 10% and 20% proportions, our SOOD achieves higher performance than Soft Teacher, even though our baseline is weaker than Soft Teacher's. Under 30% data proportion, our SOOD surpasses Soft Teacher and Unbiased Teacher by at least 1.40 mAP.

The qualitative results of our method compared with supervised baseline and Dense Teacher [50] are shown in Fig. 3. With the help of our RAW and GC, SOOD is able to exploit more potential semantic information from the unlabeled data, helping reduce false predictions and improve the detection quality.

Fully Labeled Data. We also compare our SOOD with the other SSOD methods [24, 44, 50] on fully labeled data setting. Since the reported methods are based on different detectors, we report the results of the compared methods and their baseline in Tab. 2. Our SOOD surpasses previous methods by at least 1.30 points. Compared to our baseline, we obtain +2.24 mAP improvement, which further demonstrates our method's ability to learn from unlabeled data. We notice that the performance of Unbiased Teacher [24] drops after adding unlabeled data. The reason might be that Unbiased Teacher does not apply unsupervised losses for bounding box regression, which is important for oriented object detection.

Generalization on other detectors. To further validate the effectiveness of our method, we evaluate our method on other oriented object detectors, CFA [10] and KLD [48], under the Fully Labeled Data setting. As shown in Tab. 3,

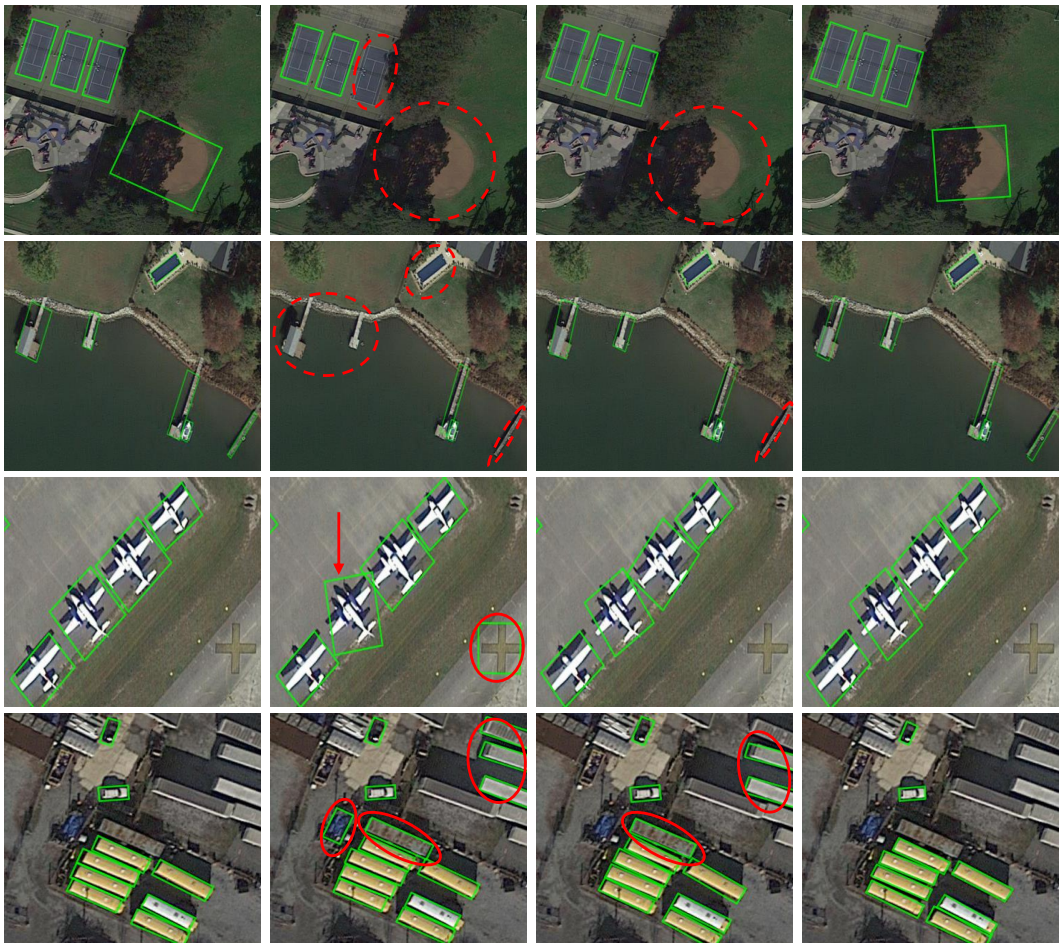

Figure 3. Some visualization examples from the DOTA-v1.5 dataset. From left to right, each column shows ground truth, results of the supervised baseline (rotated-FCOS [37]), Dense Teacher [3], and our SOOD. The green rectangles indicate predictions. The red dashed circle, solid red circle, and red arrow represent false negative, false positive, and inaccurate orientation prediction, respectively.

although CFA is a strong detector, our method still results in an improvement of +1.32 mAP and reaches 67.07 mAP. On the KLD detector, our method brings an improvement of +2.41 mAP. The above results validate the generalization ability of our method.

5.3. Ablation Study

In this section, we conduct extensive studies to validate our key designs. Unless specified, all the ablation experiments are performed using 10% of labeled data.

The effect of each component. We study the effects of the proposed two losses, Rotation-aware Adaptive Weighting (RAW) loss and Global Consistency (GC) loss. Note that our SOOD degrades to the vanilla dense pseudo-labeling framework without these two losses. As shown in Tab. 4, both losses are proved effective and complementary under all three settings: RAW and GC can each bring performance gain, and the baseline is further improved when equipped with two losses. It indicates that the local constraint built by RAW and the global constraint built by GC can benefit the semi-supervised learning process, boosting the model by constructing one-to-one and many-to-many relationships between the teacher and the student.

Table 4. The effects of Rotation-aware Adaptive Weighting (RAW) loss and Global Consistency (GC) loss. Experiments are conducted on 10%, 20% and 30% labeled data settings.

Setting	RAW	GC	mAP		
			10%	20%	30%
I	-	-	47.24	54.07	57.74
II	✓	-	47.82	55.21	58.93
III	-	✓	47.71	54.72	58.70
IV	✓	✓	**48.63**	**55.58**	**59.23**

The influence of sampling ratio. In this part, we discuss

Table 5. The effect of the sampling ratio for instance-level dense pseudo-labeling. Experiments are conducted at 10% setting, and the method is equipped with both RAW and GC losses.

Setting	Sample Ratio	mAP
I	0.125	48.27
II	0.25	**48.63**
III	0.5	47.91
IV	1.0	47.69

Table 6. The effects of different compositions in the cost map of GC. Experiments are conducted at 10% setting, and the method is equipped with RAW loss. The results indicate that both distance and score are essential factors of the cost map.

Setting	Distance	Score	mAP
I	-	-	47.82
II	-	✓	48.10
III	✓	-	47.94
IV	✓	✓	**48.63**

the influence of the ratios in sampling pseudo-labels. The results with different sampling ratios are shown in Tab. 5. The best performance, 48.36 mAP, is achieved when the sample ratio is set to 0.25. Setting it to other values degrades the performance. We hypothesize that this value ensures a good balance between noises (e.g., false positives) and valid predictions (e.g., true positives). Increasing it will introduce more noise that harms the training process, while decreasing it leads to information loss and failure in learning the representation of objects.

The effect of different compositions in cost map of GC. Here, we study the effects of the spatial distance and the score difference when constructing the cost map of optimal transport in GC loss. The results of different settings are shown in Tab. 6. We get at most +0.28 mAP improvement when using only one of them, indicating that the information from only one side is inadequate for learning the global prior. When considering both the score difference and spatial distance, the performance gain brought by GC is further improved to +0.81 mAP. It indicates that the information from score difference and spatial distance are complementary. With their help, RAW can effectively model the many-to-many relationship between the teacher and the student, providing an informative guide to the model.

The effect of RAW's hyper-parameter α. Here, we study the influence of the hyper-parameter α in RAW. As shown in Tab. 7, we set α to 1.0 and get the performance of 47.77 mAP. As α increases, the performance of our method improves when α varies from 1 to 50. However, further increasing it to 100.0 slightly hurt the performance. Therefore, we set it to 50 by default. For this observation, we conjecture that increasing the weight α will enlarge the influence of orientation information, but also amplify the impact of teacher's inaccurate labels.

Table 7. The effect of hyper-parameter α in the RAW loss. Experiments are conducted at 10% setting, and the method is equipped with GC loss.

Setting	α	mAP
I	1	47.77
II	10	47.87
III	50	**48.63**
IV	100	47.95

5.4. Limitation and Discussion

Although our method achieves satisfactory results on semi-supervised oriented object detection, the usage of aerial objects' characteristics is limited. Apart from orientation and global layout, many other properties of aerial objects should be considered, e.g., scale variations and large aspect ratios. Apart from that, we separately consider orientation and global layout by constructing two different constraints, which can be integrated into one unified module to utilize both information simultaneously. We also find that oriented objects and even complex objects wildly appear in other tasks, such as 3D object detection and text detection, leaving much room for further exploration.

6. Conclusion

In this paper, we have presented an effective solution for semi-supervised oriented object detection, which is important but neglected. Focusing on oriented objects' characteristics in aerial scenes, we have designed two novel losses, rotation-aware adaptive weighting (RAW) loss and global consistency (GC) loss. The former considers the importance of rotation information for oriented objects, dynamically weighting each pseudo-label-prediction pair by their rotation difference. The latter introduces the global layout concept to SSOD, measuring the global similarity between the teacher and the student in a many-to-many manner. To validate the effectiveness of our method, we have conducted extensive experiments on the DOTA-v1.5 benchmark. Compared with state-of-the-art methods, SOOD achieves consistent performance improvement on partially and fully labeled data.

Acknowledgements. This work was supported by the National Science Fund for Distinguished Young Scholars of China (Grant No.62225603) and the Young Scientists Fund of the National Natural Science Foundation of China (Grant No.62206103).

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proc. of Intl. Conf. on Machine Learning*, pages 214–223. PMLR, 2017. 3

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Proc. of Advances in Neural Information Processing Systems*, 32, 2019. 2

[3] Binghui Chen, Pengyu Li, Xiang Chen, Biao Wang, Lei Zhang, and Xian-Sheng Hua. Dense learning based semi-supervised object detection. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 4815–4824, 2022. 1, 6, 7

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. of Intl. Conf. on Machine Learning*, pages 1597–1607. PMLR, 2020. 2

[5] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Proc. of Advances in Neural Information Processing Systems*, 26, 2013. 3, 5

[6] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a wasserstein loss. *Advances in neural information processing systems*, 28, 2015. 5

[7] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 303–312, 2021. 3

[8] Ross Girshick. Fast r-cnn. In *Porc. of IEEE Intl. Conf. on Computer Vision*, pages 1440–1448, 2015. 2

[9] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Proc. of Advances in Neural Information Processing Systems*, 17, 2004. 2

[10] Zonghao Guo, Chang Liu, Xiaosong Zhang, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. Beyond bounding-box: Convex-hull feature adaptation for oriented and densely packed object detection. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 8792–8801, 2021. 6

[11] Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2021. 5

[12] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Redet: A rotation-equivariant detector for aerial object detection. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 2786–2795, 2021. 2, 5

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5

[14] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. *Proc. of Advances in Neural Information Processing Systems*, 32, 2019. 2

[15] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Proc. of Intl. Conf. on Machine Learning*, volume 3, page 896, 2013. 2

[16] Gang Li, Xiang Li, Yujie Wang, Shanshan Zhang, Yichao Wu, and Ding Liang. Pseco: Pseudo labeling and consistency training for semi-supervised object detection. In *Proc. of European Conference on Computer Vision*, 2022. 1

[17] Jingyu Li, Zhe Liu, Jinghua Hou, and Dingkang Liang. Dds3d: Dense pseudo-labels with dynamic threshold for semi-supervised 3d object detection. *Proc. of IEEE Intl. Conf. on Robotics and Automation*, 2023. 2

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

[19] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. In *Proc. of the AAAI Conf. on Artificial Intelligence*, volume 34, pages 11474–11481, 2020. 2

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

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

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Porc. of IEEE Intl. Conf. on Computer Vision*, pages 2980–2988, 2017. 2, 4, 6

[23] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Proc. of European Conference on Computer Vision*, pages 21–37. Springer, 2016. 2

[24] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *Proc. of International Conference on Learning Representations*, 2021. 1, 2, 6

[25] Yen-Cheng Liu, Chih-Yao Ma, and Zsolt Kira. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 9819–9828, 2022. 2, 3

[26] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2018. 2

[27] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781. 2, 3

[28] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *Proc. of IEEE Intl. Conf. on Com-*

puter Vision and Pattern Recognition, pages 4119–4128, 2018. 2

[29] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 779–788, 2016. 2

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Proc. of Advances in Neural Information Processing Systems*, 28, 2015. 2, 6

[31] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Proc. of Advances in Neural Information Processing Systems*, 29, 2016. 2

[32] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Proc. of Advances in Neural Information Processing Systems*, 33:596–608, 2020. 2

[33] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 2

[34] Jingqun Tang, Wenqing Zhang, Hongye Liu, MingKun Yang, Bo Jiang, Guanglong Hu, and Xiang Bai. Few could be better than all: Feature sampling and grouping for scene text detection. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 4563–4572, 2022. 2

[35] Yihe Tang, Weifeng Chen, Yijun Luo, and Yuting Zhang. Humble teachers teach better students for semi-supervised object detection. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 3132–3141, 2021. 1, 2

[36] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Proc. of Advances in Neural Information Processing Systems*, 30, 2017. 1, 2

[37] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Porc. of IEEE Intl. Conf. on Computer Vision*, pages 9627–9636, 2019. 3, 5, 6, 7

[38] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009. 5

[39] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai Nguyen. Distribution matching for crowd counting. *Proc. of Advances in Neural Information Processing Systems*, 33:1595–1607, 2020. 3, 5

[40] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 3974–3983, 2018. 5

[41] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Proc. of Advances in Neural Information Processing Systems*, 33:6256–6268, 2020. 2

[42] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020. 2

[43] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented r-cnn for object detection. In *Porc. of IEEE Intl. Conf. on Computer Vision*, pages 3520–3529, 2021. 2

[44] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Porc. of IEEE Intl. Conf. on Computer Vision*, pages 3060–3069, 2021. 1, 2, 3, 6

[45] Qize Yang, Xihan Wei, Biao Wang, Xian-Sheng Hua, and Lei Zhang. Interactive self-training with mean teachers for semi-supervised object detection. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 5941–5950, 2021. 2

[46] Xue Yang and Junchi Yan. On the arbitrary-oriented object detection: Classification based approaches revisited. *International Journal of Computer Vision*, 130(5):1340–1365, 2022. 2

[47] Xue Yang, Junchi Yan, Ziming Feng, and Tao He. R3det: Refined single-stage detector with feature refinement for rotating object. In *Proc. of the AAAI Conf. on Artificial Intelligence*, volume 35, pages 3163–3171, 2021. 2

[48] Xue Yang, Xiaojiang Yang, Jirui Yang, Qi Ming, Wentao Wang, Qi Tian, and Junchi Yan. Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. *Proc. of Advances in Neural Information Processing Systems*, 2021. 6

[49] Fangneng Zhan, Yingchen Yu, Kaiwen Cui, Gongjie Zhang, Shijian Lu, Jianxiong Pan, Changgong Zhang, Feiying Ma, Xuansong Xie, and Chunyan Miao. Unbalanced feature transport for exemplar-based image translation. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 15028–15038, 2021. 3, 5

[50] Hongyu Zhou, Zheng Ge, Songtao Liu, Weixin Mao, Zeming Li, Haiyan Yu, and Jian Sun. Dense teacher: Dense pseudo-labels for semi-supervised object detection. In *Proc. of European Conference on Computer Vision*, 2022. 1, 2, 3, 6

[51] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. *Proc. of Advances in Neural Information Processing Systems*, 33:3833–3845, 2020. 2