ISPRS Journal of Photogrammetry and Remote Sensing 221 (2025) 179–192

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

S^{3}OD: Size-unbiased semi-supervised object detection in aerial images

Ruixiang Zhang [a], Chang Xu [a], Fang Xu [b], Wen Yang [a,*], Guangjun He [c], Huai Yu [a], Gui-Song Xia [b]

[a] School of Electronic Information, Wuhan University, Wuhan 430072, China
[b] School of Artificial Intelligence, Wuhan University, Wuhan 430072, China
[c] State Key Laboratory of Space Information System and Integrated Application, China Academy of Space Technology, Beijing, 100086, China

ARTICLE INFO

Keywords:
Aerial images
Semi-supervised learning
Object detection

ABSTRACT

Aerial images present significant challenges to label-driven supervised learning, in particular, the annotation of substantial small-sized objects is a highly laborious process. To maximize the utility of scarce labeled data alongside the abundance of unlabeled data, we present a semi-supervised learning pipeline tailored for label-efficient object detection in aerial images. In our investigation, we identify three size-related biases inherent in semi-supervised object detection (SSOD): pseudo-label imbalance, label assignment imbalance, and negative learning imbalance. These biases significantly impair the detection performance of small objects. To address these issues, we propose a novel Size-unbiased Semi-Supervised Object Detection (S^{3}OD) pipeline for aerial images. The S^{3}OD pipeline comprises three key components: Size-aware Adaptive Thresholding (SAT), Size-rebalanced Label Assignment (SLA), and Teacher-guided Negative Learning (TNL), all aimed at fostering size-unbiased learning. Specifically, SAT adaptively selects appropriate thresholds to filter pseudo-labels for objects at different scales. SLA balances positive samples of objects at different sizes through resampling and reweighting. TNL alleviates the imbalance in negative samples by leveraging insights from the teacher model, enhancing the model's ability to discern between object and background regions. Extensive experiments on DOTA-v1.5 and SODA-A demonstrate the superiority of S^{3}OD over state-of-the-art competitors. Notably, with merely 5% SODA-A training labels, our method outperforms the fully supervised baseline by 2.17 points. Codes are available at https://github.com/ZhangRuixiang-WHU/S3OD/tree/master.

1. Introduction

Object detection in aerial images primarily relies on fully supervised methods (Wang et al., 2022a; Shi et al., 2020; Zhang et al., 2022c; Liang et al., 2022; Ding et al., 2019; Liu et al., 2021a, 2022a), where meticulous annotation of each region of interest is essential for training aerial object detectors. However, the acquisition of large-scale object-by-object annotation for aerial images is quite laborious and expensive. In particular, the objects' large quantity and small size (Ding et al., 2022) amplify the burden in the annotation process. Take one representative aerial object detection dataset as an example: DOTA-v1.5 (Ding et al., 2022) contains more than 140 objects per image on average while its mean absolute object size is only 34 × 34 pixels, the annotation of each image can take up tens to hundreds of minutes. This motivates us to develop a label-efficient object detection pipeline that reduces the heavy reliance on annotations while maintaining high detection precision.

One approach to address label-efficient aerial object detection is by adapting a state-of-the-art SSOD pipeline to aerial images. Unfortunately, as shown in Fig. 1, this approach shows limited improvement on aerial images, achieving 12% improvement with 5% labels, compared to a 47% improvement on MS COCO (Xu et al., 2021). SSOD's advances are centered on generic object detection, while aerial images exhibit characteristics distinct from generic objects, impeding the performance of SSOD. Notably, we observe that the substantial number of small objects[1] among aerial images will introduce severe bias issues to the existing SSOD pipelines. Although previous works like Unbiased Teacher (Liu et al., 2021b) and Consistent Teacher (Wang et al., 2023) have noticed the bias issue in SSOD, they are constrained to mitigating the *foreground–background bias* or *inter-class bias* issues, while overlooking the *size-induced bias* inherent in aerial images.

* Corresponding author.
E-mail addresses: zhangruixiang@whu.edu.cn (R. Zhang), xuchangeis@whu.edu.cn (C. Xu), xufang@whu.edu.cn (F. Xu), yangwen@whu.edu.cn (W. Yang), hegj@spacestar.com.cn (G. He), yuhuai@whu.edu.cn (H. Yu), guisong.xia@whu.edu.cn (G.-S. Xia).

[1] According to MS COCO (Lin et al., 2014), objects with area smaller than 1024 (32 × 32) pixels are defined as small objects, objects with areas larger than 1024 pixels are called large objects in this paper.

https://doi.org/10.1016/j.isprsjprs.2025.01.037
Received 7 February 2024; Received in revised form 5 August 2024; Accepted 28 January 2025
Available online 12 February 2025

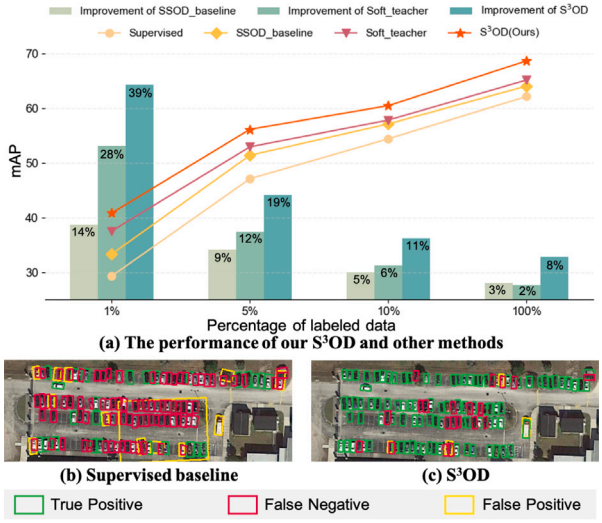

Fig. 1. (a) The line chart shows a comparison between the performance of SSOD-baseline, Soft-teacher, and our S^{3}OD on DOTA-v1.5 validation set under different proportions of labeled training data. The bar chart shows the relative improvement over the supervised baseline. (b) Detection results of supervised setting under 1% labeling rate. (c) Detection results of our S^{3}OD of semi-supervised setting under 1% labeling rate.

In short, the limited number of pixels and visual quality make small objects easily suppressed and regarded as outliers, *i.e.*, existing methods are prone to be biased towards confident large objects while neglecting less confident small objects. In a typical SSOD pipeline (*e.g.* Soft Teacher (Xu et al., 2021)), bias induced by small objects can be found in three components: pseudo-labeling, label assignment, and negative learning (see Fig. 2). First, *pseudo-labels between small and large objects are unbalanced*, as the predicted confidence of small objects tends to be lower than large objects. Consequently, applying a fixed threshold for pseudo-label selection will lead to a drastic loss of supervision for small objects. Conversely, although lowering the threshold can compensate more pseudo labels for small objects, it will introduce much low-quality and inaccurate supervision for large objects. Existing SSOD method (Hua et al., 2023) for aerial images lacks a size-aware pseudo-labeling strategy adaptive to all-size objects. Second, *small objects suffer from positive sample insufficiency issues when assigning labels between pseudo boxes and anchors*, exacerbating the size-biased learning. For small objects, a slight location deviation between the pseudo box and the anchor will lead to severe mismatch issues (Xu et al., 2022a,b), reducing the number of positive anchors assigned to small objects. Previous works (Li et al., 2022a; Wang et al., 2023) ignore the effect of size when optimizing the assignment of inaccurate pseudo labels. In addition, *negative learning tends to mistakenly suppress foreground small objects*. The recall of small objects is usually much lower than large objects, there remain many undiscovered small objects in the "background" region. Hence, during negative learning, the foreground smaller objects are more likely to be wrongly suppressed as the background, which in turn reduces the detection performance. By reducing the weight of ambiguous samples, the previous method (Xu et al., 2021) alleviates the influence of foreground–background confusion, but it also reduces the discriminability for hard negative samples.

To address these challenges, we propose a novel Size-unbiased Semi-Supervised Object Detection (S^{3}OD) pipeline tailored for label-efficient object detection in aerial images. S^{3}OD introduces three newly designed strategies: Size-aware Adaptive Thresholding (SAT), Size-rebalanced Label Assignment (SLA), and Teacher-guided Negative Learning (TNL). Specifically, SAT decouples the threshold of pseudo-label selection for different-sized objects, alleviating the quantity imbalance between large and small pseudo-labels. SLA employs a distribution-based re-sampling strategy (Xu et al., 2022b) in the pseudo-label assignment, mitigating the positive sample imbalance between small and large objects. Additionally, SLA incorporates a size-aware re-weighting strategy, further reducing the adverse influence of small objects' lack of supervision. TNL double-checks preliminary negative samples obtained by the student model with the mature teacher model. By forwarding the student's negative samples to the teacher, we filter out proposals with a high probability of being foreground and uncover hard negative samples, enhancing the network's discrimination ability. Our contributions can be summarized as follows:

- We identify three types of bias induced by small objects that impede SSOD on aerial imagery, namely the pseudo-label imbalance, label assignment imbalance, and negative learning imbalance.
- We propose the size-unbiased pipeline S^{3}OD, consisting of three modules: Size-aware Adaptive Thresholding, Size-rebalanced Label Assignment, and Teacher-guided Negative Learning, designed to address the aforementioned imbalance issues.
- S^{3}OD achieves state-of-the-art performance across various semi-supervised settings on two representative aerial image datasets: DOTA-v1.5 and SODA-A. The compelling improvement in small object classes demonstrates our method's effectiveness in tackling small object-induced issues.

The rest of this paper is organized as follows. First, we introduce related work about aerial object detection and SSOD in Section 2. Then, we experimentally support the motivation of this paper in Section 3. Following this, Section 4 provides the details of S^{3}OD. After that, extensive experiments on two datasets are performed to verify the effectiveness of S^{3}OD especially on small objects in Section 5. In addition, insights and limitations of this work are discussed in Section 6. Finally, we conclude in Section 7.

2. Related work

In this section, we provide a brief review of works relevant to this work, from the aspects of object detection in aerial images and semi-supervised object detection.

2.1. Object detection in aerial images

Different from images captured from the natural scene, aerial images are characterized by their arbitrarily oriented objects and a substantial number of small objects. Correspondingly, we survey object detection methods designed for aerial images from these two aspects.

Oriented Object Detection. The development of oriented object detection follows the trend of exploring efficient and simpler architecture and training strategies. At the early stage, Rotated RPN (Ma et al., 2018) tackles rotated object detection by employing additional rotated anchor boxes. RoI Transformer (Ding et al., 2019) gets rid of rotated anchors via learning to convert RPN-generated horizontal proposals into rotated ones. Furthermore, Oriented R-CNN (Xie et al., 2021) directly generates oriented proposals at the RPN stage. Based on a one-stage model, S2ANet (Han et al., 2021) utilizes the DCN to explicitly align features with anchors. Additionally, several works (Xie et al., 2021; Li et al., 2022c, 2021) tailor anchor-free detectors for rotated detection. Recently, there are also some works (Qiao et al., 2023; Ming et al., 2023) focusing on the optimization of rotation angle regression to achieve more accurate rotated object detection.

Small Object Detection. The high proportion of small objects brings severe challenges for existing aerial object detectors. Super-resolution, as a straightforward and effective idea, is incorporated into small object detection in several works (Shermeyer and Van Etten, 2019; Courtrai et al., 2020; Bashir and Wang, 2021). Besides, the SCRDet (Yang et al., 2019) and SCRDet++ (Yang et al., 2022) optimize the feature extraction process and emphasize richer feature information, to enhance small objects. Moreover, some works have found that the label assignment strategy in the existing detectors is

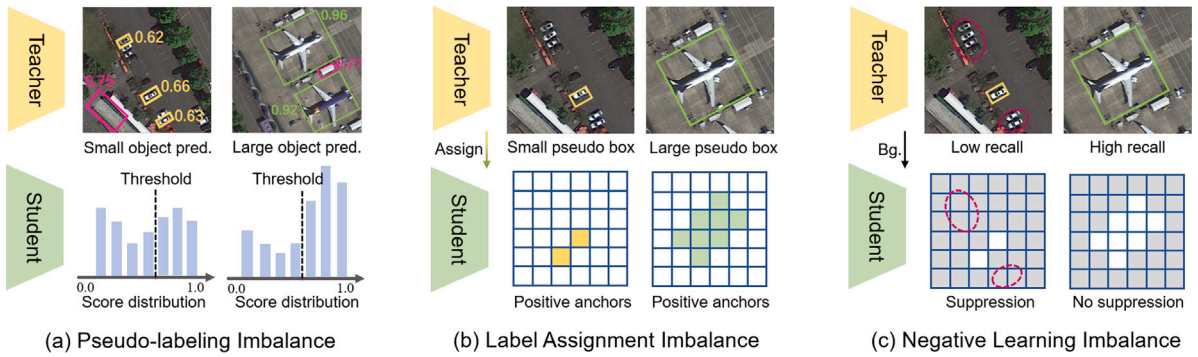

Fig. 2. Schematic diagrams of size-induced bias issues in the SSOD pipeline. (a) pseudo-labels between small and large objects are unbalanced; (b) the positive samples assigned to each small pseudo box are much fewer than the large one; (c) small objects are more easily suppressed as background regions.

extremely unfriendly to small objects, and proposed some new label assignment strategies like NWD (Xu et al., 2022a), RFLA (Xu et al., 2022b). DCFL (Xu et al., 2023b) also explores dynamic priors and balanced learning for rotated small object detection.

The above methods effectively optimize the detector scheme based on the characteristics of objects in aerial images, thereby enhancing detection performance. However, these fully-supervised learning methods usually require plenty of training data with expensive annotation costs. In contrast, we explore the semi-supervised detection algorithms for aerial images to effectively alleviate the demand for annotations.

2.2. Semi-supervised object detection

SSOD, a branch of semi-supervised learning (Sohn et al., 2020a; Zhang et al., 2021; Kim et al., 2021; Wang et al., 2022b; Zheng et al., 2022; Chen et al., 2023) in the domain of object detection, revolves around two core ideas: consistency regularization (Jeong et al., 2019, 2021; Wang et al., 2021) and pseudo-label learning (Lee et al., 2013; Li et al., 2022b). The former aims to ensure the network generates consistent predictions across data with diverse augmentations, thereby enhancing semantic comprehension. Meanwhile, the latter uses self-training to acquire pseudo-labels for unlabeled data through pre-detector training using labeled data. The main research line of SSOD focuses on generic objects, while some attention has also been turned to aerial objects recently.

Semi-supervised Generic Object Detection. Currently, the predominant line of SSOD (Sohn et al., 2020b; Zhou et al., 2021; Yang et al., 2021a; Zhang et al., 2022b) works integrates both of the two strategies. Early SSOD (Sohn et al., 2020b; Zoph et al., 2020) methods often employ multi-stage training, necessitating recurrent pseudo-label updates through iterative training phases, including Unbiased teacher (Liu et al., 2021b), Humble teachers (Tang et al., 2021). Then, Soft-teacher (Xu et al., 2021) adopts the teacher–student network from MeanTeacher (Tarvainen and Valpola, 2017), enabling end-to-end training and presenting a new paradigm for SSOD. Based on this paradigm, subsequent SSOD research mainly focuses on refining pseudo-label learning and enhancing consistency regularization. From the perspective of pseudo-label refinement, a considerable portion of research is dedicated to formulating more discerning criteria for evaluating pseudo-labels and developing effective filtering strategies (Xu et al., 2021; Liu et al., 2022b; Choi et al., 2022; Zhang et al., 2022d; Vandeghen et al., 2022; Wang et al., 2023). Apart from pseudo-label filtering, another facet of research focuses on judiciously leveraging those inaccurate pseudo-labels (Xu et al., 2021; Li et al., 2022a; Wang et al., 2023; Liu et al., 2023a; Xu et al., 2023a; Chen et al., 2022a,b; Zhou et al., 2022a). From the perspective of consistency regularization, some methods propose innovative augmentation strategies to increase the input perturbation and enhance the network's learning of semantic features (Liu et al., 2023b; Guo et al., 2022; Kim et al., 2022). In addition, some works optimize their frameworks by combining specific tasks or configurations. For instance, the optimization of sample selection methods addresses class imbalance challenges (Zhang et al., 2022a; Mi et al., 2022), and the integration of Transformer detectors is incorporated into the design of SSOD frameworks (Wang et al., 2022c; Zhang et al., 2023).

Semi-supervised Aerial Object Detection. Recently, there have been several studies focusing on SSOD in aerial images. Chen et al. (2018) utilizes a GAN network to generate adversarial negative samples for training the network's classification. However, it has only been tested on sparse small datasets and lacks universality. With the introduction of the Teacher–Student network into the SSOD pipeline, significant advancements (Hua et al., 2023; Shen et al., 2023; Liu et al., 2024) have been made in SSOD for aerial images, including recent progress in open-set SSOD (Liu et al., 2024). Among these, the most relevant to our task is SOOD (Hua et al., 2023). SOOD adopts the current mainstream semi-supervised detection paradigm and incorporates the orientations of targets in aerial images into the semi-supervised detection algorithm, achieving promising improvement. This method strengthens supervision based on the rotation consistency of densely arranged targets in aerial images. However, it still does not effectively address the challenges posed by small objects, which remains an urgent problem to be tackled.

By contrast, we are the first study that investigates the influence of aerial images' small objects on different components of the SSOD pipeline. We tackle the issue of imbalance caused by object size by selecting appropriate pseudo-labels, designing balanced pseudo-label assignment strategies, and customizing negative sample learning schemes.

3. Motivation

In this section, we investigate the behavior of small objects *vs.* large objects in the SSOD pipeline with a popular teacher–student architecture (Wang et al., 2023; Chen et al., 2022b; Li et al., 2022a). Through statistical analysis, we reveal two gaps between small and large objects: predicted confidence and recall, which further leads to three biases in the SSOD pipeline.

Limited training images and labels will enlarge the confidence gap between small and large objects. As illustrated in Fig. 3, we calculate the average prediction scores for small and large objects under (1) fully supervised training with different ratios of images (*i.e*, 100%, 10%, 5%, and 1%), (2) semi-supervised training with 1% training labels. Several noteworthy observations emerge from Fig. 3: First, the prediction confidence scores exhibit a direct correlation with the volume of training data. In scenarios with limited training data, the confidence scores experience a commensurate decline. Second, the confidence discrepancy between small and large objects gradually widens as the decrease of training labels. Third, the SSOD training

Fig. 3. The average predicted confidence scores for large and small objects under different training settings. "Sup" denotes "Supervised Setting".

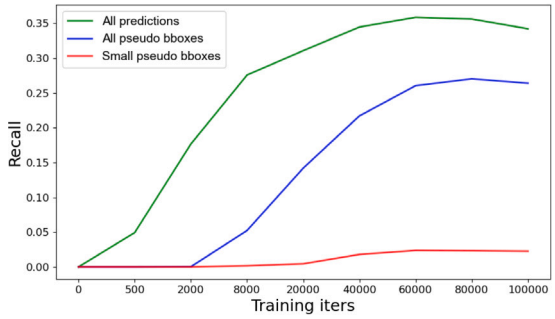

Fig. 4. The recalls of all predictions, all selected pseudo-labels, and small objects within the pseudo-labels during the training of Vanilla SSOD.

pipeline even enlarges the confidence gap between small and large objects when using the same number of labels as the supervised method. Vanilla SSOD predominantly relies on high-quality pseudo-labels to facilitate supervised training for unlabeled data. Considering the aforementioned observations, the network exhibits a proclivity towards selecting high-quality predictions associated with large-sized objects as pseudo-labels. This gives rise to the first bias issue: *unbalanced pseudo-labels between small and large objects*. In addition, previous works (Xu et al., 2022b, 2023b, 2022a) point out that the mean number of positive samples assigned to each small object is much smaller than that of the larger object. The pseudo-label imbalance issue further exacerbates this dilemma, worsening the second bias: *small objects are facing severe positive sample insufficiency issues when assigning labels between pseudo boxes and anchors.*

The recall achieved with pseudo labels for small objects significantly lags behind that of large objects. We perform statistics regarding the recalls from all predictions, all pseudo-labels, and small objects' pseudo-labels relative to the ground truth during training, as shown in Fig. 4. It can be observed that as the network undergoes training, the recall of predictions gradually ascends, exhibiting a growth trend similar to that observed in the recall of selected pseudo-labels. However, the recall rate from small objects' pseudo-labels remains consistently low, underscoring a pronounced deficiency in supervision for small objects during the training of unlabeled data. The crux of the matter lies in the fact that the bulk of the introduced unlabeled data primarily contributes to the training of large objects. The absence of supervision signal for unlabeled small objects not only hinders positive enhancement but also wrongly suppresses the undiscovered small objects as background, indicating the third type of bias: *negative sample learning is inclined to mistakenly suppress foreground small object.*

These observations motivate us to further explore SSOD frameworks tailored for aerial images, where the corresponding solutions are presented in Section 4.

4. Method

This section presents the details of our proposed method S^{3}OD. First, we will introduce the overall architecture of the basic SSOD method that we adopt, and then we will introduce our proposed methods for SSOD in aerial images, including SAT, SLA, and TNL.

4.1. Teacher–student basic pipeline

We follow the mainstream paradigm of pseudo-label learning for the overall framework (Xu et al., 2021; Liu et al., 2021b; Zhou et al., 2022a), as shown in Fig. 5. We take the common teacher–student model as the basic pipeline, where the student model is updated with the normal back-propagation, and the teacher model is the Exponential Moving Average (EMA) of the student model. Given the labeled data set $D_l = \{I_l^i, G_l^i\}_{i=1}^{N_l}$ and the unlabeled data set $D_u = \{I_u^i\}_{i=1}^{N_u}$, I_l^i and G_l^i represent the i-th labeled images and the corresponding GTs, N_l is the number of labeled images. I_u^i represents the i-th unlabeled images, N_u is the numbers of unlabeled images. In each training iteration, labeled and unlabeled data are randomly sampled from D_l and D_u, respectively, to train the student branch.

For the labeled data, a standard supervised training pipeline is employed. Note that our method can be applied to two-stage detectors that consist of Region Proposal Network (RPN) and Region-CNN (R-CNN) (Ren et al., 2016). Therefore, the supervised loss $\mathcal{L}_{sup}$ is calculated by:

$$\mathcal{L}_{\text{sup}} = \sum_i \mathcal{L}_{\text{rpn}}\left(I_l^i\right) + \mathcal{L}_{\text{roi}}\left(I_l^i\right), \tag{1}$$

where $\mathcal{L}_{\text{rpn}}$ and $\mathcal{L}_{\text{roi}}$ are the RPN loss and ROI loss of the two-stage detector, like in Faster R-CNN (Ren et al., 2016).

For the unlabeled data, weak and strong augmentations are respectively applied to the teacher and student models to enforce consistency regularization. The purpose of this step is to encourage the network to output consistent predictions for perturbed data and enhance the extraction of semantically invariant features. Following previous works, weak augmentation typically includes random flipping and resizing, while strong augmentation involves random rotation, translation, shearing, erasing, solarizing, adjusting color, contrast, sharpening, *etc*. In this paper, we do not explore new augmentation methods, and all strong and weak augmentations used are off-the-shelf approaches. After applying strong and weak augmentations, each unlabeled image produces two views. The view with weak augmentation is input into the teacher model for inference, resulting in predicted detection results. Based on a predetermined threshold, suitable predictions are selected as pseudo-labels P_u for the unlabeled image. The view with strong augmentation, along with the pseudo-labels P_u, is fed into the student model. By leveraging the supervision signal provided by the pseudo-labels, the unsupervised loss $\mathcal{L}_{unsup}$ is calculated by:

$$\mathcal{L}_{\text{unsup}} = \sum_i \mathcal{L}^u_{\text{rpn}}\left(I_u^i\right) + \mathcal{L}^u_{\text{roi}}\left(I_u^i\right), \tag{2}$$

where $\mathcal{L}^u_{\text{rpn}}$ and $\mathcal{L}^u_{\text{roi}}$ represent the RPN loss and ROI loss supervised by pseudo-labels P_u for unlabeled images. In this paper, we re-design these two losses in the semi-supervised training process to enhance the detection performance of small objects. The details are provided in Sections 4.3 and 4.4.

The overall loss function is formulated as:

$$\mathcal{L} = \mathcal{L}_{sup} + \alpha\mathcal{L}_{unsup}, \tag{3}$$

where α represents a weight parameter utilized to balance the contribution of unlabeled data.

To enable a fair comparison with previous methods, we take the well-established rotated version of Faster R-CNN as the basic detector.

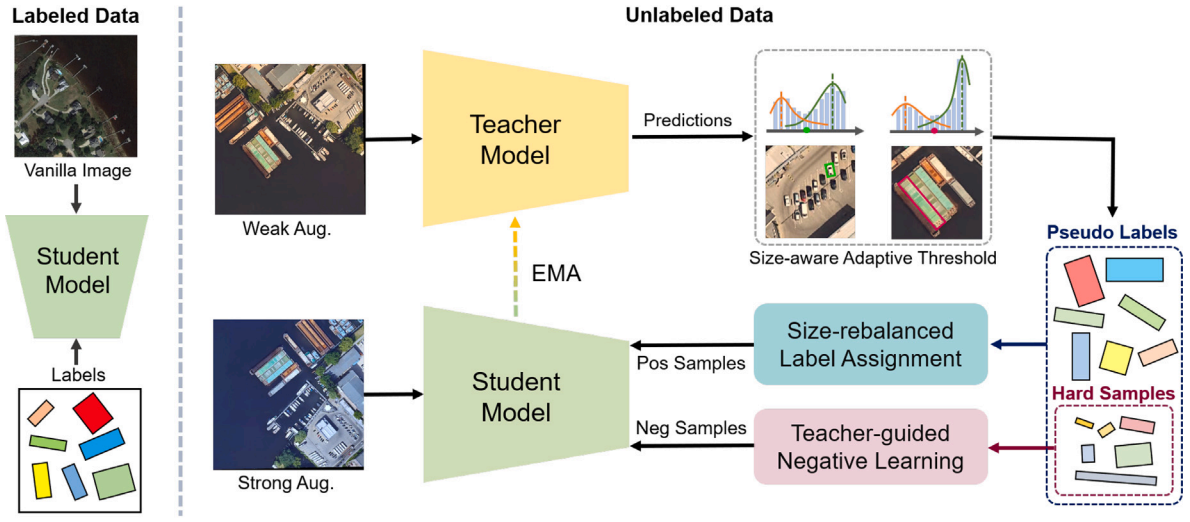

Fig. 5. The pipeline of the proposed S^{3}OD. Each training batch contains both labeled and unlabeled data. The labeled data uses Ground Truth for general supervised training. The training of unlabeled data is based on the teacher–student model, and the teacher network is updated by the EMA of the student network. The teacher network infers the unlabeled data, and then selects appropriate pseudo-labels through SAT. SLA assigns positive samples based on the pseudo-labels. TNL strengthens the negative learning based on the ambiguous predictions of the teacher network.

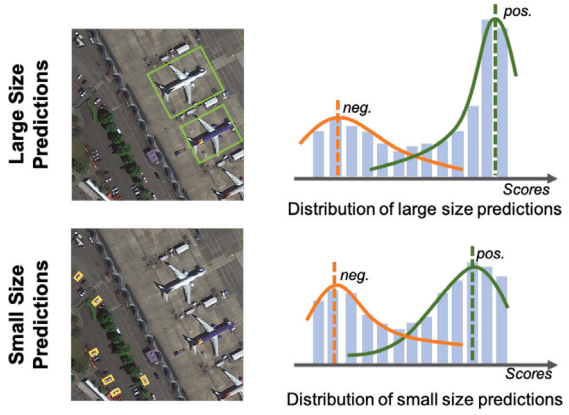

Fig. 6. Illustration of the Size-aware Adaptive Thresholding (SAT) operation. We utilize the GMM to fit the predictions of small and large size predictions respectively. Then the size-adaptive pseudo-labels are decided by the mean of negative and positive thresholds.

4.2. Size-aware adaptive thresholding

Most of the existing SSOD methods utilize a fixed threshold to filter pseudo-labels. However, when processing data with an unbalanced distribution, fixed thresholds are not always conducive to all situations. Previous works (Wang et al., 2023; Kar et al., 2023) have pointed out the class confidence imbalance problem, where the class-wise reweighing or sampling strategy was correspondingly proposed to tackle this issue. Nevertheless, existing works commonly neglect the size-confidence imbalance issue in aerial imagery. Towards this end, we propose to define pseudo-labels in a size-decoupled manner, and we name this strategy Size-aware Adaptive Thresholding (SAT).

In SAT, as shown in Fig. 6, we compute the statistics of predicted results from the teacher model and decouple these results into two distributions according to the size of their bounding box: one for large objects and the other for small objects. Inspired by Wang et al. (2023) and our size-related observations, we hypothesize that the predicted result s in each distribution follows a Gaussian Mixture Model (GMM) $\mathcal{P}(s)$ of two modes, the positive and negative predictions.

$$\mathcal{P}(s) = w_{neg}\mathcal{N}_{neg}\left(\mu_{neg}, {\sigma_{neg}}^2\right) + w_{pos}\mathcal{N}_{pos}\left(\mu_{pos}, {\sigma_{pos}}^2\right), \tag{4}$$

where $\mathcal{N}\left(\mu, \sigma^2\right)$ denotes the Gaussian distribution, w, μ, σ denote the weight, mean, and standard deviation of the Gaussian distribution, respectively. The distributions from positive and negative predictions are distinguished by the subscripts pos and neg. Subsequently, the Expectation-Maximization (EM) algorithm is applied to fit the distribution $\mathcal{P}(s)$. The adaptive pseudo-label threshold τ is then determined by taking the mean of μ_{pos} and μ_{neg}.

During each training iteration, we separately sample a sufficient number of large and small size predictions to fit GMM: $\mathcal{P}_s(s)$ and $\mathcal{P}_l(s)$. Specifically, we respectively sample top G_s and G_l predictions, which are calculated by the summation of small and large size predictions' confidence in a batch (Wang et al., 2023), for fitting $\mathcal{P}_s(s)$ and $\mathcal{P}_l(s)$. Then we get μ^s_{pos}, μ^s_{neg} for $\mathcal{P}_s(s)$ and μ^l_{pos}, μ^l_{neg} for $\mathcal{P}_l(s)$. Finally, the corresponding adaptive thresholds τ_s and τ_l for small and large objects are calculated by:

$$\tau_s = (\mu^s_{pos} + \mu^s_{neg})/2, \quad \tau_l = (\mu^l_{pos} + \mu^l_{neg})/2. \tag{5}$$

4.3. Size-rebalanced label assignment

After obtaining a series of pseudo-labels, we need to assign positive or negative (*pos/neg*) labels to predefined anchors based on the obtained pseudo-labels to train the object detector. Meanwhile, the sample assignment between pseudo-labels and anchors plays quite a significant role in the performance of SSOD (Li et al., 2022a; Wang et al., 2023). During this process, we observe a significant problem for aerial images, *i.e.,* the small-sized positive sample insufficiency issue as discussed in Section 1.

To address this issue, we propose a Size-rebalanced Label Assignment (SLA) strategy, which explores size-balanced learning through two key aspects: distribution-based re-sampling and size-aware re-weighting. In the distribution-based re-sampling, we model the rotated box (cx, cy, w, h, θ) into the two-dimensional Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (Yang et al., 2021b,c), in which the box's geometry center $\boldsymbol{\mu} = (cx, cy)$ serves as the Gaussian's mean vector. And $\boldsymbol{\Sigma}$ is the covariance matrix of the Gaussian distribution, which can be computed by:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \frac{w^2}{4} & 0 \\ 0 & \frac{h^2}{4} \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}. \tag{6}$$

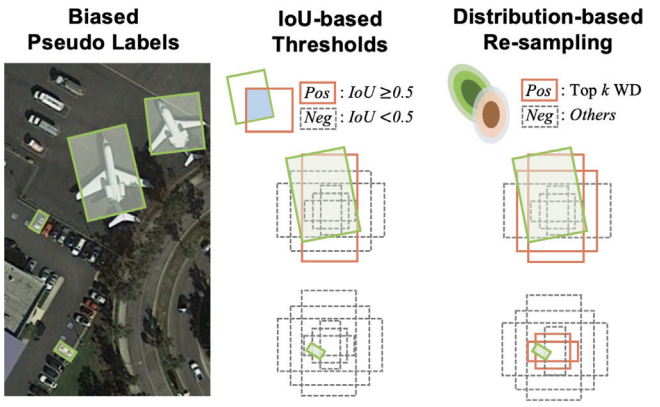

Fig. 7. Comparison of the IoU threshold-based assignment strategy and the distribution-based re-sampling used in SLA. The green boxes represent pseudo boxes, the orange boxes represent positive anchors assigned to pseudo boxes, and the black dashed boxes represent negative anchors. With the IoU threshold-based strategy, the number of anchors that small objects can match tends to be fewer. Distribution-based re-sampling proves effective in mitigating this issue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Based on the distribution-based modeling, we use the Wasserstein distance to measure the similarity between pseudo-boxes and anchors since the Wasserstein distance is demonstrated conducive to tiny objects (Xu et al., 2022b,a; Yang et al., 2021b) in label assignment, mainly owing to its ability to measure non-overlapping boxes. Specifically, the Wasserstein distance between the Gaussian pseudo-box $\mathcal{N}_p(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ and the Gaussian anchor $\mathcal{N}_a(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a)$ has a closed form solution, which can be simplified as:

$$W_2^2(\mathcal{N}_p, \mathcal{N}_a) = \left\|\boldsymbol{\mu}_p - \boldsymbol{\mu}_a\right\|_2^2 + \mathrm{Tr}\left(\boldsymbol{\Sigma}_p + \boldsymbol{\Sigma}_a - 2\left(\boldsymbol{\Sigma}_p^{1/2}\boldsymbol{\Sigma}_a\boldsymbol{\Sigma}_p^{1/2}\right)^{1/2}\right). \tag{7}$$

For brevity, we call this metric WD. With WD, which can measure the similarity between non-overlapping boxes, we can now calculate the similarity between all preset anchors and pseudo-boxes. For each pseudo-box, we assign the top K anchors that yield the highest similarity with the pseudo-box as positive samples. In general, the strategy of WD with top K sampling can alleviate the imbalance among the positive anchor sampling process to some extent. As shown in Fig. 7, compared to previous IoU threshold-based pseudo-label assignment strategies, the WD combined with the top K can robustly address label assignment in situations where anchors have extremely low IoU scores with the small-sized pseudo boxes. By doing so, this strategy can compensate more positive samples for those easily ignored small objects, enhancing small objects' supervision signal and solving the problem that large objects are easier to be over-optimized with more positive anchors (Xu et al., 2022b).

In addition to the distribution-based re-sampling, we design a novel size-aware re-weighting method. During the training process, we count the quantity of large and small positive samples in each batch. When calculating loss, we re-weight the loss of large and small positive samples to further balance the contribution of large and small objects in training. Specifically, concerning the RPN loss for unlabeled images, L^u_{rpn}, we solely adjust the weights of the loss for all positive samples, designated as L^{pos}_{rpn}, aiming to rebalance the loss for large-scale/small-scale proposals among positive samples. Let A_{pos} represent the set of positive proposals engaged in the training batch for RPN, with a total count of N_{pos}. Subsequently, based on the size of proposals, A_{pos} can be categorized into A_s and A_l. Here, A_s denotes the set of small-sized proposals, with a total count of N_s, and their corresponding labels are denoted as T_s. A_l is the set of proposals with large sizes, totaling N_l, where the corresponding label set is T_l. The re-weighted loss of all positive samples can be calculated by:

$$\mathcal{L}^{pos}_{\text{rpn}} = \frac{N_{pos}}{2N_s}\sum_{i\in A_s}\mathcal{L}_{\text{rpn}}(A_s^i, T_s^i) + \frac{N_{pos}}{2N_l}\sum_{j\in A_l}\mathcal{L}_{\text{rpn}}(A_l^j, T_l^j). \tag{8}$$

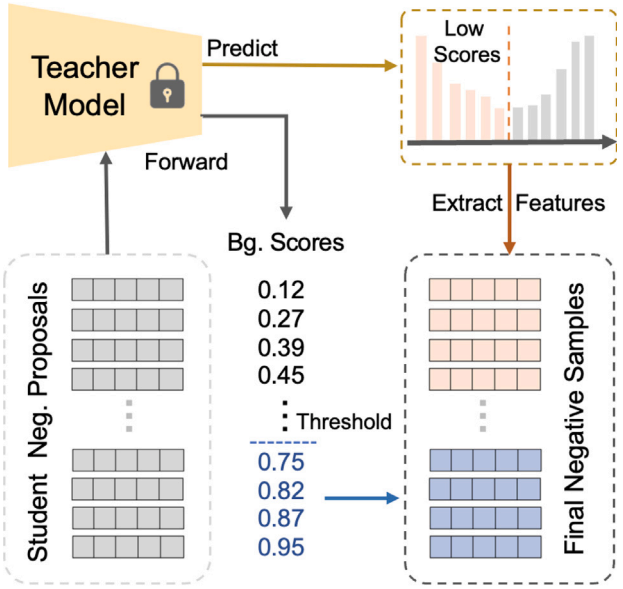

Fig. 8. Illustration of Teacher-guided Negative Learning (TNL). The final negative samples comprise two parts: the predictions of the teacher model and teacher-checked student's negative proposals. Note that "Bg." denotes "Background".

4.4. Teacher-guided negative learning

To overcome the concern that negative sample learning will mistakenly suppress undiscovered small objects, we introduce the Teacher-guided Negative Learning (TNL) approach for unbiased negative sample learning.

Object detectors' predictions are not always *black and white*. More practically, there exist numerous ambiguous predictions that cannot certainly be labeled as foreground or background via a simple threshold. Due to limited pixels and information, this ambiguity is more severe for small object prediction. Existing SSOD methods usually discard these ambiguous predictions, avoiding misleading gradients while sacrificing their potential contribution to training. Fortunately, some studies on other semi-supervised learning tasks (e.g., U^2PL Wang et al., 2022d, ANL Chen et al., 2023) suggests that the ambiguous results predicted by the network can be excavated to aid learning. However, their methods of defining ambiguous results and incorporating them into training are tailored for different tasks and are infeasible to SSOD directly. Thus, we cast a further look at ambiguous predictions in the SSOD task for ODAI.

As shown in Fig. 8, in the proposed TNL, we design two steps to make the detector rethink the utility of these ambiguous samples. In the first step, the regions of the student's negative proposals are input into the teacher model to obtain classification confidence scores of background. Subsequently, we select negative proposals with higher background scores (larger than 0.7) for negative sample supervision, aiming to effectively eliminate False Negatives (FN). However, this approach gives rise to a new issue: the selected negative samples often turn out to be overly simplistic, discarding the learning of hard samples. This limitation impacts the network's discriminative capacity, leading to a significant increase in False Positives (FP) at the later stages of training. Hence, to enhance the network's discrimination ability for hard samples, we incorporate the network's predictions with confidence scores lower than τ_h as soft hard samples into the negative sample

Table 1
Quantitative comparison between the proposed S^{3}OD and SOTA SSOD methods on DOTA-v1.5. Experiments are performed on the validation set of DOTA-v1.5 under the training with different labeling rates. All results are reported as the form $mean \pm std$ across the five-fold experiments. The best results are in bold.

Methods	Partially labeled			Fully labeled
	1%	5%	10%	100%
Supervised:				
Rotated Faster R-CNN	$29.30_{\pm 1.62}$	$47.08_{\pm 0.97}$	$54.36_{\pm 0.98}$	63.6
Oriented R-CNN (Xie et al., 2021)	$33.42_{\pm 1.41}$	$50.01_{\pm 1.11}$	$56.54_{\pm 0.78}$	67.2 (+3.6)
Semi-supervised:				
SSOD baseline	$33.30_{\pm 2.54}$	$51.36_{\pm 0.82}$	$57.08_{\pm 0.74}$	64.9 (+1.3)
STAC (Sohn et al., 2020b)	$33.98_{\pm 2.13}$	$50.86_{\pm 1.63}$	$56.72_{\pm 0.75}$	63.8 (+0.2)
Unbiased Teacher (Liu et al., 2021b)	$33.18_{\pm 2.26}$	$48.20_{\pm 1.14}$	$54.74_{\pm 0.84}$	64.8 (+1.2)
Soft Teacher (Xu et al., 2021)	$37.52_{\pm 2.44}$	$52.90_{\pm 1.24}$	$57.78_{\pm 0.32}$	65.1 (+1.5)
PseCo (Li et al., 2022a)	$37.60_{\pm 2.46}$	$52.98_{\pm 1.41}$	$58.26_{\pm 0.85}$	65.5 (+1.9)
SOOD[a] (Hua et al., 2023)	$35.94_{\pm 1.48}$	$52.93_{\pm 1.30}$	$57.90_{\pm 0.49}$	67.7 (+4.1)
S^{3}OD (Ours)	$\mathbf{40.80}_{\pm 1.56}$	$\mathbf{56.10}_{\pm 1.43}$	$\mathbf{60.42}_{\pm 0.64}$	**68.6 (+5.0)**

[a] Means using Rotated FCOS as the base detector.

proposals ($\tau_h = 0.5$ in the implementation). Based on the original confidence predictions, we re-weight the loss of these hard samples: the lower the confidence score, the higher the weight, thus granting them greater significance during the training of negative samples. Following the previous work (Shrivastava et al., 2016), we add the hard-negative samples in the R-CNN stage and compute the ROI loss $\mathcal{L}_{\text{roi}}^{neg}$ for unlabeled images. Specifically, we solely adjust the loss of the negative sample in R-CNN stage, defined as $\mathcal{L}_{\text{roi}}^{neg}$, aiming to add the hard-negative samples from teacher prediction. Let B_{neg} denote the set of negative sample proposals involved in the training of the R-CNN for unlabeled images. This set comprises both hard negative sample set B_{hard} generated by the Teacher network's low score predictions, and the negative sample set B_n, sampled from the student proposals. The negative sample loss of the R-CNN head $\mathcal{L}_{\text{roi}}^{neg}$ is calculated by:

$$\mathcal{L}_{\text{roi}}^{neg} = \sum_{i \in B_{hard}} w(s_i)\mathcal{L}_{\text{roi}}^{i}(B_{hard}^{i}, \mathbb{1}_{bg}) + \sum_{j \in B_n} \mathcal{L}_{\text{roi}}^{j}(B_n^{j}, \mathbb{1}_{bg}) \tag{9}$$

where $w(\cdot)$ is the weight function for hard negative samples, calculated based on the confidence score s of hard negative samples, we use $w(s) = 2(1 - s^2)$ in our implementation. $\mathbb{1}_{bg}$ represents the label for negative samples, for background.

5. Experiment

5.1. Dataset and evaluation protocol

We conduct experiments on two large-scale datasets specifically tailored for object detection in aerial images, *i.e.*, DOTA-v1.5 (Ding et al., 2022) and SODA-A (Cheng et al., 2023).

DOTA-v1.5 (Ding et al., 2022) is updated based on DOTA-v1.0 (Xia et al., 2018). Compared to v1.0, the images in v1.5 remain unchanged, but there are additional annotations for small objects. These enriched annotations of small objects allow the dataset to better reflect the characteristics of real-world aerial imagery objects. The DOTA-v1.5 comprises 2,806 large-scale aerial images and 402,089 annotations. It is 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 without released annotations.

SODA-A (Cheng et al., 2023) is specially constructed for the detection of a large number of small objects in aerial images. A total of 2513 high-resolution aerial images and contains 872,069 instances over nine classes are included. SODA-A is similarly divided into three sets. The training set contains 1067 images, the validation set contains 576 images, and the test set contains 870 images.

Following the common setting of SSOD (Li et al., 2022a; Xu et al., 2021; Liu et al., 2021b), we conducted two sets of experiments equally on both datasets: **Partially Labeled Data** and **Fully Labeled Data**.

In the **Partially Labeled Data** experiments, 1%, 5%, and 10% of the training set are randomly sampled as labeled data, while the remaining data serves as unlabeled data. To mitigate the impact of random sampling, we perform 5-fold experiments for each sampling rate and report the mean and std of the results. In the **Fully Labeled Data** experiments, we utilize the test set without released annotations as unlabeled data, while using the entire training set as labeled data. Note that aerial images are typically large-size and are often cropped before training. Moreover, sampling from the entire large images at a 1% sampling rate would make it difficult to cover all the categories adequately. Sampling from the cropped smaller images allows for a better representation of the overall data distribution. Therefore, in our experiments, we first perform cropping on all the images before sampling.

In all experiments, we report the evaluation results on the validation set using the commonly used mean Average Precision (mAP) under the IoU threshold of 0.5, following the MS COCO evaluation metric.

5.2. Implementation details

We use Rotated Faster RCNN as the base rotated object detector and choose ResNet50 with FPN as the backbone. The implementation of the base detector is based on the MMRotate framework (Zhou et al., 2022b). Following the settings in aerial image object detection, all large-scale images are cropped to 1024×1024 pixels with a 200-pixel overlap.

The supervised baseline is trained using stochastic gradient descent (SGD) optimizer. We set the learning rate (LR) to 0.005 and the batch size (BS) to 4. In the absence of additional comments, all SSOD methods are also trained using SGD optimizer with an LR of 0.001 and a BS of 5 (4 unlabeled images and 1 labeled image). The beta value is set to 4.0 to control the contributions of unlabeled data.

All experiments are conducted on 2 RTX4090 GPUs and 4 TITAN V100 GPUs. For partially labeled data, we train 100,000 iterations for the 1% labeled data, 120,000 iterations for the 5% labeled data, and 160,000 iterations for the 10% labeled data. For fully labeled data, we train the model for 320,000 iterations.

5.3. Main results

We compare our proposed S^{3}OD with existing SOTA SSOD methods, including STAC (Sohn et al., 2020b), Unbiased Teacher (Liu et al., 2021b), Soft-Teacher (Xu et al., 2021), PseCo (Li et al., 2022a), and SOOD (Hua et al., 2023). Additionally, we use a vanilla end-to-end pseudo-labeling framework (teacher–student network) without the three modules proposed in our method as the SSOD baseline. For the

Table 2
The performance of objects with different scales in the validation set of DOTA-v1.5. We randomly sample 1% of the labels five times for training. The results are reported as the form $mean \pm std$ across the five experiments. The best results are in bold and sub-optimal results are underlined.

Methods	AP_s	AP_m	AP_l	AR^{2000}	AR_s^{2000}
Supervised:					
Rotated Faster R-CNN	$14.58_{\pm 1.01}$	$33.96_{\pm 2.04}$	$27.14_{\pm 4.18}$	$34.56_{\pm 1.93}$	$19.50_{\pm 2.15}$
Oriented R-CNN (Xie et al., 2021)	$17.96_{\pm 2.07}$	$37.88_{\pm 1.99}$	$31.10_{\pm 4.49}$	$39.00_{\pm 2.25}$	$22.74_{\pm 2.56}$
Semi-supervised:					
SSOD baseline	$20.64_{\pm 2.48}$	$36.12_{\pm 3.34}$	$35.82_{\pm 5.55}$	$38.16_{\pm 2.48}$	$27.34_{\pm 3.28}$
STAC (Sohn et al., 2020b)	$19.68_{\pm 1.00}$	$39.68_{\pm 1.58}$	$30.56_{\pm 4.08}$	$39.20_{\pm 1.79}$	$24.86_{\pm 1.31}$
Unbiased Teacher (Liu et al., 2021b)	$18.78_{\pm 2.53}$	$36.18_{\pm 1.92}$	$34.90_{\pm 5.00}$	$42.34_{\pm 2.01}$	$27.80_{\pm 2.86}$
Soft Teacher (Xu et al., 2021)	$\underline{25.62}_{\pm 2.30}$	$41.88_{\pm 2.40}$	$\underline{36.88}_{\pm 5.02}$	$46.18_{\pm 2.72}$	$36.74_{\pm 3.88}$
PseCo (Li et al., 2022a)	$24.12_{\pm 2.35}$	$\mathbf{43.66}_{\pm 3.77}$	$34.18_{\pm 2.88}$	$\mathbf{50.20}_{\pm 2.83}$	$\underline{37.64}_{\pm 3.00}$
SOOD[a] (Hua et al., 2023)	$20.86_{\pm 1.95}$	$39.72_{\pm 2.12}$	$31.26_{\pm 2.98}$	$47.62_{\pm 1.51}$	$32.50_{\pm 3.94}$
S^{3}OD (Ours)	$\mathbf{28.90}_{\pm 2.35}$	$\underline{42.94}_{\pm 1.99}$	$\mathbf{39.60}_{\pm 3.07}$	$\underline{49.24}_{\pm 1.68}$	$\mathbf{39.14}_{\pm 1.53}$

[a] Means using Rotated FCOS as the base detector.

Table 3
The performance of the proposed S^{3}OD and SOTA SSOD methods on several representative categories in the validation set of DOTA-v1.5. We randomly sample 1% of the labels five times for training. The results are reported as the form $mean \pm std$ across the five experiments. The best results are in bold and sub-optimal results are underlined.

Methods	SP	PL	SV	LV	BR	ST	HB	TC	SL
Supervised:									
Rotated Faster R-CNN	$56.8_{\pm 4.9}$	$68.0_{\pm 3.3}$	$22.4_{\pm 0.6}$	$39.6_{\pm 4.4}$	$13.8_{\pm 2.4}$	$31.8_{\pm 12.0}$	$23.7_{\pm 4.5}$	$68.4_{\pm 12.4}$	$31.7_{\pm 9.2}$
Oriented R-CNN (Xie et al., 2021)	$68.6_{\pm 3.5}$	$69.2_{\pm 1.6}$	$23.5_{\pm 0.6}$	$\underline{52.2}_{\pm 3.1}$	$14.9_{\pm 2.8}$	$36.3_{\pm 14.0}$	$31.1_{\pm 4.6}$	$73.6_{\pm 11.1}$	$34.9_{\pm 8.7}$
Semi-supervised:									
SSOD baseline	$41.1_{\pm 4.9}$	$83.0_{\pm 7.5}$	$12.3_{\pm 3.4}$	$34.8_{\pm 10.8}$	$19.1_{\pm 5.9}$	$35.1_{\pm 18.7}$	$24.4_{\pm 7.4}$	$73.6_{\pm 20.2}$	$47.6_{\pm 17.3}$
STAC (Sohn et al., 2020b)	$64.7_{\pm 7.2}$	$77.1_{\pm 3.7}$	$23.7_{\pm 0.7}$	$45.6_{\pm 6.2}$	$16.6_{\pm 4.1}$	$37.1_{\pm 11.2}$	$26.3_{\pm 6.6}$	$75.0_{\pm 11.1}$	$41.3_{\pm 9.6}$
Unbiased Teacher (Liu et al., 2021b)	$46.1_{\pm 11.6}$	$\underline{87.8}_{\pm 0.6}$	$17.4_{\pm 3.0}$	$32.4_{\pm 6.6}$	$\underline{21.2}_{\pm 4.0}$	$27.6_{\pm 12.8}$	$24.0_{\pm 3.6}$	$78.5_{\pm 14.3}$	$43.7_{\pm 11.5}$
Soft Teacher (Xu et al., 2021)	$67.3_{\pm 3.9}$	$86.8_{\pm 3.1}$	$23.7_{\pm 0.3}$	$37.9_{\pm 11.6}$	$19.5_{\pm 3.5}$	$43.2_{\pm 20.4}$	$23.8_{\pm 2.2}$	$\underline{87.5}_{\pm 2.0}$	$\underline{51.8}_{\pm 9.8}$
PseCo (Li et al., 2022a)	$67.5_{\pm 2.2}$	$84.5_{\pm 2.3}$	$29.6_{\pm 0.8}$	$48.4_{\pm 8.0}$	$14.1_{\pm 2.6}$	$\underline{49.5}_{\pm 12.4}$	$\underline{31.8}_{\pm 8.0}$	$84.8_{\pm 3.1}$	$48.6_{\pm 9.9}$
SOOD[a] (Hua et al., 2023)	$\mathbf{74.8}_{\pm 1.4}$	$77.9_{\pm 0.4}$	$\underline{30.3}_{\pm 1.1}$	$\mathbf{55.0}_{\pm 3.0}$	$17.8_{\pm 3.5}$	$36.5_{\pm 14.0}$	$\mathbf{34.3}_{\pm 5.1}$	$75.5_{\pm 9.7}$	$44.8_{\pm 6.0}$
S^{3}OD (Ours)	$\underline{71.1}_{\pm 3.4}$	$\mathbf{88.1}_{\pm 0.4}$	$\mathbf{41.5}_{\pm 2.0}$	$48.2_{\pm 6.9}$	$\mathbf{23.3}_{\pm 5.3}$	$\mathbf{56.1}_{\pm 23.5}$	$27.0_{\pm 7.2}$	$\mathbf{88.8}_{\pm 1.2}$	$\mathbf{58.1}_{\pm 4.8}$

[a] Means using Rotated FCOS as the base detector.

first few SOTA SSOD methods, we re-implement them based on the Rotated Faster-RCNN for rotated object detection and adjust the appropriate hyperparameters to better adapt SSOD for aerial images. To further illustrate the effectiveness of our proposed SSOD method compared to supervised methods, we also include a SOTA supervised method, Oriented R-CNN (OR) (Xie et al., 2021), for comparison. The same weak/strong augmentation strategy following Soft-teacher (Xu et al., 2021) is used in these methods for fair comparison. For SOOD (Hua et al., 2023), it uses the Rotated FCOS as the basic Detector. Moreover, owing to the specific designs of this method, both the teacher network and the student network must have inputs of the same size. Therefore, we only incorporate consistent multi-scale augmentation into its original strong augmentation.

Results on DOTA-v1.5. As shown in Table 1, for partially labeled data, our proposed S^{3}OD method achieves the best performance under the training of 1%, 5%, and 10% labeling rate, achieving 40.80 mAP, 56.10 mAP, and 60.42 mAP respectively. This outperforms the supervised baseline by 11.50 points, 9.02 points, and 6.06 points, respectively. Similarly, our method also surpasses the SOTA method PseCo (Li et al., 2022a) by 3.20 points, 3.12 points, and 2.16 points at different labeling rates. For fully labeled data, S^{3}OD also achieves optimal performance, exceeding the supervised baseline by 5.0 points to 68.6 mAP. Compared with the SOTA supervised method, S^{3}OD has excellent results when using only Rotated Faster R-CNN as the base detector. Considering the above results, our method outperforms the SOTAs by a large margin at all labeling rates on aerial images. The outstanding performance also demonstrates the superiority of our method in semi-supervised object detection tasks for aerial images.

To further demonstrate our method's effectiveness on small objects, we perform a scale-aware performance evaluation: AP_s, AP_m, AP_l, utilizing the COCO style metrics under the representative 1% labeled data. Herein, AP_s is mAP for small objects within $0 \sim 32^2$ pixels, AP_m accounts for medium-sized objects within $32^2 \sim 96^2$ pixels, and AP_l pertains to large objects exceeding 96^2 in size. The results are shown in Table 2, which reports the average results over 5-fold cross-validations. From the table, it can be observed that our method achieves the best detection performance for small objects, while also yielding very competitive results for medium and large object detection. The significant improvement in AP_s highlights the outstanding performance of our method on small objects. Additionally, we report AR^{2000}, AR_s^{2000} (following COCO style), representing the average recall for all objects and small objects, respectively, with 2000 predictions retained per image. Our method exhibits the highest recall for small objects and achieves the suboptimal result for overall recall as well.

Moreover, focusing on the detection performance of specific categories, we select several representative categories from aerial images, including "*ship*" (SP), "*plane*" (PL), "*small-vehicle*" (SV), "*large-vehicle*" (LV), "*bridge*" (BR), "*storage-tank*" (ST), "*harbor*" (HB), '*tennis-court*' (TC), and "*swimming-pool*" (SL)'. To thoroughly validate the improvement of semi-supervised learning with limited labeled data, we present the results of S^{3}OD and other SOTA methods for each category at the labeling rate of 1%, as shown in Table 3. (We report the average results of 5-fold experiments). From the results, our method shows significant improvements compared to supervised detection results in all categories. Among the categories with the highest proportions in the entire dataset, including "*ship*", "*small-vehicle*", and "*large-vehicle*", our method achieves improvements of 14.3, 19.1, and 8.6 points in mAP respectively. Focusing on small objects, "*small-vehicle*", and "*storage-tank*" are mostly in small object sizes. Our method outperforms other SSOD methods in these categories. Notably, "*small-vehicle*" is the category with the most instances and the highest number of small objects in aerial images. Our method shows significant improvement in this category. In contrast, other SSOD methods show minimal improvement in "*small-vehicle*" detection, and the performance of the baseline SSOD is even lower than that of supervised learning. This confirms our previous observation that generic SSOD methods are inclined to focus on large objects while neglecting the learning of small objects, thus affecting the ability to detect small objects (see Table 4).

Table 4
Quantitative comparison between the proposed S^{3}OD and SOTA SSOD methods on SODA-A. Experiments are performed on the validation set of SODA-A under the training with different labeling rates. All results are reported as the form *mean* ± *std* across the five-fold experiments. The best results are in bold.

Methods	Partially labeled			Fully labeled
	1%	5%	10%	100%
Supervised:				
Rotated Faster R-CNN	$40.83_{\pm 1.43}$	$50.92_{\pm 0.65}$	$57.07_{\pm 0.48}$	61.6
Oriented R-CNN (Xie et al., 2021)	$43.77_{\pm 1.36}$	$54.91_{\pm 0.75}$	$58.94_{\pm 0.41}$	62.9 (+1.3)
Semi-supervised:				
SSOD baseline	$34.08_{\pm 1.71}$	$53.38_{\pm 0.88}$	$56.66_{\pm 0.74}$	61.4 (−0.2)
STAC (Sohn et al., 2020b)	$45.52_{\pm 0.79}$	$54.06_{\pm 0.92}$	$57.96_{\pm 0.33}$	63.2 (+1.6)
Unbiased Teacher (Liu et al., 2021b)	$41.88_{\pm 2.08}$	$53.40_{\pm 0.81}$	$57.33_{\pm 0.44}$	60.4 (−1.2)
Soft Teacher (Xu et al., 2021)	$47.99_{\pm 1.80}$	$58.67_{\pm 0.71}$	$60.05_{\pm 0.31}$	62.8 (+1.2)
PseCo (Li et al., 2022a)	$48.73_{\pm 1.86}$	$58.87_{\pm 0.49}$	$61.34_{\pm 0.55}$	62.2 (+0.6)
SOOD[a] (Hua et al., 2023)	$50.03_{\pm 1.08}$	$60.44_{\pm 0.42}$	$63.11_{\pm 0.51}$	67.7 (+6.1)
S^{3}OD (Ours)	$\mathbf{54.69}_{\pm 2.31}$	$\mathbf{63.77}_{\pm 0.51}$	$\mathbf{65.98}_{\pm 0.34}$	**69.3 (+7.7)**

[a] Means using Rotated FCOS as the base detector.

Table 5
The performance of the proposed S^{3}OD and SOTA SSOD methods on each category in the validation set of SODA-A. We randomly sample 1% of the labels five times for training. The results are reported as the form *mean* ± *std* across the five experiments. The best results are in bold and sub-optimal results are underlined.

Methods	SP	PL	SV	LV	HC	ST	CT	SL	WM
Supervised:									
Rotated Faster R-CNN	$42.4_{\pm 5.1}$	$71.7_{\pm 2.7}$	$25.6_{\pm 3.0}$	$11.2_{\pm 3.8}$	$19.9_{\pm 13.6}$	$47.5_{\pm 3.3}$	$33.1_{\pm 3.5}$	$72.3_{\pm 3.2}$	$43.8_{\pm 4.6}$
Oriented R-CNN (Xie et al., 2021)	$45.8_{\pm 5.3}$	$73.2_{\pm 2.7}$	$27.9_{\pm 2.7}$	$\underline{17.1}_{\pm 4.1}$	$\underline{24.3}_{\pm 9.1}$	$49.6_{\pm 3.5}$	$37.9_{\pm 5.9}$	$73.7_{\pm 0.8}$	$44.4_{\pm 5.1}$
Semi-supervised:									
SSOD baseline	$22.4_{\pm 9.6}$	$66.7_{\pm 15.4}$	$13.5_{\pm 4.6}$	$12.7_{\pm 6.1}$	$21.6_{\pm 12.2}$	$31.5_{\pm 11.1}$	$24.7_{\pm 5.8}$	$79.0_{\pm 10.4}$	$34.6_{\pm 8.7}$
STAC (Sohn et al., 2020b)	$49.6_{\pm 5.1}$	$79.1_{\pm 1.5}$	$28.6_{\pm 3.4}$	$16.5_{\pm 5.1}$	$2.1_{\pm 6.8}$	$51.4_{\pm 0.7}$	$33.6_{\pm 3.3}$	$81.2_{\pm 2.9}$	$47.7_{\pm 3.2}$
Unbiased Teacher (Liu et al., 2021b)	$37.5_{\pm 4.5}$	$80.5_{\pm 3.1}$	$23.0_{\pm 2.8}$	$15.4_{\pm 8.1}$	$18.8_{\pm 13.2}$	$48.8_{\pm 2.1}$	$29.3_{\pm 3.1}$	$81.5_{\pm 1.8}$	$42.3_{\pm 3.9}$
Soft Teacher (Xu et al., 2021)	$48.6_{\pm 9.7}$	$\mathbf{85.3}_{\pm 0.7}$	$28.4_{\pm 3.6}$	$15.6_{\pm 5.7}$	$19.7_{\pm 11.6}$	$57.3_{\pm 1.6}$	$43.0_{\pm 1.6}$	$\mathbf{84.2}_{\pm 0.2}$	$\underline{49.9}_{\pm 4.6}$
PseCo (Li et al., 2022a)	$\underline{57.3}_{\pm 2.9}$	$\underline{84.6}_{\pm 1.0}$	$30.4_{\pm 1.5}$	$\mathbf{19.1}_{\pm 6.2}$	$17.2_{\pm 12.2}$	$56.6_{\pm 1.3}$	$\mathbf{46.1}_{\pm 1.3}$	$\underline{82.7}_{\pm 0.3}$	$44.7_{\pm 6.0}$
SOOD[a] (Hua et al., 2023)	$56.9_{\pm 7.3}$	$76.1_{\pm 0.5}$	$\underline{40.0}_{\pm 1.0}$	$16.0_{\pm 3.7}$	$\mathbf{28.6}_{\pm 11.5}$	$\underline{59.9}_{\pm 1.8}$	$43.4_{\pm 4.6}$	$78.5_{\pm 2.3}$	$\mathbf{50.7}_{\pm 2.5}$
S^{3}OD (Ours)	$\mathbf{65.0}_{\pm 8.3}$	$\mathbf{85.3}_{\pm 0.9}$	$\mathbf{52.8}_{\pm 2.3}$	$15.3_{\pm 4.6}$	$24.1_{\pm 15.1}$	$\mathbf{74.8}_{\pm 0.6}$	$\underline{45.8}_{\pm 2.3}$	$\mathbf{84.2}_{\pm 0.5}$	$45.2_{\pm 7.1}$

[a] Means using Rotated FCOS as the base detector.

Results on SODA-A. As shown in Table 4, the proposed S^{3}OD also achieves the new state-of-the-art results in SODA-A. Specifically, for partially labeled data, S^{3}OD achieves 54.69 mAP, 63.77 mAP, and 65.98 mAP under three settings of 1%, 5%, and 10% labeling rate, respectively. These represent improvements of 13.86, 12.85, and 8.91 points over the baseline under supervised training. In the case of fully labeled data, S^{3}OD also achieves 69.3 mAP, surpassing the supervised baseline by 7.7 points. When compared to experiments on the DOTA-v1.5 dataset, our method showcases even more substantial improvements on the SODA-A dataset. This underscores our method's efficacy in precisely enhancing the performance of small objects in semi-supervised object detection for aerial images. It is worth noting that the SODA-A dataset is specially designed for small objects, with over 95% of the instances being small objects with fewer than 32^2 pixels. The excellent performance on SODA-A further proves the effectiveness of our S3OD in enhancing SSOD for small objects.

Furthermore, we follow the aforementioned configuration to present the performance results of various methods on individual categories of the SODA-A dataset with 1% labeled data. There are a total of 9 categories, namely "*airplane*" (PL), "*helicopter*" (HC), "*small-vehicle*" (SV), "*large-vehicle*" (LV), "*ship*" (SP), "*container*" (CT), "*storage-tank*" (ST), "*swimming-pool*" (SL), and "*windmill*" (WM), as shown in Table 5. Owing to the higher prevalence of small objects, S^{3}OD demonstrates noteworthy performance improvements, achieving optimal or near-optimal results in seven out of the nine categories. Notably, focusing on the categories with relatively more small objects, "*ship*", "*small-vehicle*", and "*storage-tank*", S^{3}OD exhibits significant advancements compared to other methods. The performance gains are substantial, exceeding the supervised baseline by 22.6, 27.2, and 27.3 points in mAP, respectively. Furthermore, when compared to suboptimal results, our method consistently showcases remarkable improvements, achieving gains of 7.7, 12.8, and 14.9 points in mAP, respectively.

Visualization. Fig. 9 presents the qualitative comparison between S^{3}OD and other SOTA methods in the DOTA-v1.5 dataset. We can see that, by introducing the proposed strategies, false negatives (red boxes) of small objects are remarkably reduced, indicating that small objects are more sufficiently learned. Furthermore, by focusing on the number of yellow boxes in the image, it can be noticed that compared to other SOTA SSOD methods, our approach demonstrates a stronger suppression effect on false alarms, which is the desired effect of the proposed TNL (Teacher-guided Negative Labeling) technique.

Training efficiency and complexity analysis. We test the training efficiency of SOTA SSOD methods by measuring the training time for 100K iterations. The experiments are conducted on 1 RTX 4090 GPU under the 1% labeled DOTA-v1.5 dataset, and the results are shown in Table 6. It can be observed that our method demonstrates a training speed comparable to some SOTA semi-supervised methods utilizing the same detector. Additionally, the additional training time incurred by our method over the SSOD baseline framework is within an acceptable range when training for 100k iterations.

Notably, during inference, we only need one branch in the symmetric Teacher–Student model to perform object detection, so the proposed method maintains the same computational complexity as the baseline detector. Taking Rotated Faster R-CNN as the baseline detector, the model has $41.09M$ parameters, 206.57 GFlops, and 45.1 FPS with one RTX 4090 GPU during inference.

5.4. Ablation study

In this section, we conduct detailed ablations to validate our key designs. Without losing generality, all the ablation experiments are performed on the single data fold with 1% labeled DOTA-v1.5 dataset.

(a) Supervised (b) Soft Teacher (Xu et al., 2021) (c) Pseco (Li et al., 2022a) (d) SOOD (Hua et al., 2023) (e) S^{3}OD (Ours)

Fig. 9. Visualization of the detection results of different SSOD methods on DOTA-v1.5 dataset. Correct predictions are marked with green boxes, false positive predictions are marked with yellow boxes, and missing targets are marked with red boxes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 6
The comparison of training time between different SSOD methods. All training is conducted on the 1% labeled DOTA-v1.5 dataset with 100k iterations.

Method	Time
SSOD baseline	7.5 h
Unbiased Teacher (Liu et al., 2021b)	8.1 h
Soft Teacher (Xu et al., 2021)	9.2 h
PseCo (Li et al., 2022a)	8.8 h
SOOD[a] (Hua et al., 2023)	7.5 h
S^{3}OD(Ours)	9.1 h

[a] Means using Rotated FCOS as the base detector, while others use Rotated Faster R-CNN.

5.4.1. Component analysis

To verify the effectiveness of each proposed strategy individually, we conduct experiments with all possible combinations of the proposed three strategies. As depicted in Table 7a, the baseline SSOD algorithm achieves an mAP of 36.0 when no additional strategies are employed. When we apply any of the proposed strategies, the baseline performance is consistently improved. By gradually incorporating all three strategies, the mAP shows a progressive improvement, verifying each design's effectiveness. Notably, SAT contributes significantly to the improvements. SAT enhances the baseline performance by 4.1 points because it directly impacts the supervision of small objects. These findings indirectly affirm the substantial impact of small objects on aerial image SSOD performance. Fortunately, our proposed methods effectively mitigate this impact, making them highly suitable for SSOD tasks in aerial images.

5.4.2. Comparisons of different thresholds

In this section, we compare several ways of setting the threshold for selecting pseudo-labels, including the default fixed threshold of

Table 7
Ablations. We train on DOTA-v1.5 training set with only 1% labeled data, test on the validation set, and report mAP under IoU threshold 0.5. The best results are in bold.

(a) Component analysis of the proposed method

	SAT	SLA	TNL	mAP
SSOD baseline				36.0
Different configurations	✓			40.1
		✓		37.8
			✓	38.1
	✓	✓		41.3
	✓		✓	40.8
		✓	✓	40.4
S^{3}OD (Ours)	✓	✓	✓	**42.1**

(b) Different pseudo-labels thresholding strategy

Methods	Setting	mAP
Fixed threshold	0.7	33.6
	0.8	36.6
	0.9	36.0
	Large-0.9 & Small-0.8	38.1
Adaptive threshold	Class-aware	35.2
	Size-aware (Ours)	**40.1**

(c) Component analysis of the SLA

Num	Setting	mAP
I	Baseline	36.0
II	Resampling	36.5
III	SLA	**37.8**

(d) Effects of parameters K

K	1	2
mAP	39.8	**42.1**
K	3	4
mAP	40.9	40.7

(e) Component analysis of the negative learning

Num	Setting	FA	mAP
I	T-BG	0.832	37.7
II	T-BG (Reweight)	0.759	39.5
III	T-BG (HNS)	0.747	41.2
IV	TNL (Ours)	**0.626**	**42.1**

(f) Effects of parameters τ_h

τ_h	0.3	0.4	0.5	0.6	0.7
mAP	40.7	41.3	**42.1**	41.5	40.5

(g) The comparison of different weighting function $w(\cdot)$

$w(s)$	1	2	$1-s$	$2(1-s)$	$3(1-s)$	$1-s^2$	$2(1-s^2)$	$3(1-s^2)$
mAP	40.6	40.9	40.6	41.3	40.4	41.2	**42.1**	41.1

Fig. 10. Visualization of pseudo-labels selection. (a) and (c) are under the fixed threshold of 0.9, (b) and (d) are under our SAT. Correct pseudo labels are marked with green boxes, and the missing labeled targets are marked with red boxes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

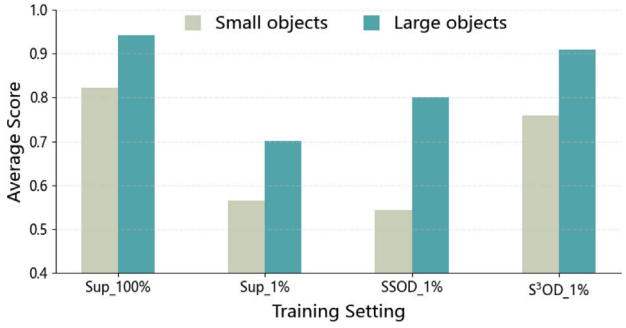

Fig. 11. Comparing the average confidence scores of large and small objects in correct predictions under the three settings of supervised by 100% labeled data, supervised, semi-supervised, and our S^{3}OD by 1% labeled data.

0.9 (baseline), other fixed thresholds (0.8 or 0.7), setting thresholds based on the object size where a threshold of 0.9 is used for large objects and 0.8 for small objects, setting thresholds based on category distribution (Similar to SAT, we use GMM to fit each class's prediction distribution and generate class-aware adaptive thresholds), and the proposed SAT method. Results are listed in Table 7b. Compared to a completely fixed threshold, setting different thresholds for large and small objects shows higher performance, indicating that the optimal threshold for pseudo-labels is different for small and large objects. The proposed SAT method still improves performance compared to other ways, demonstrating that adapting the threshold based on the GMM-guided distribution is a superior strategy.

Furthermore, we provide visualizations of the selected pseudo-labels when training under different strategies in Fig. 10. In this figure, green boxes denote the correct pseudo-labels, yellow boxes denote the false pseudo-labels, and red boxes denote objects missed to be pseudo-labeled. It can be observed that when using a fixed threshold, a large number of small objects are filtered out, while our method can retain more accurate pseudo-labels for small objects.

5.4.3. Different compositions in SLA

Here, we dissect different compositions in SLA. Results are shown in Table 7c. When the distribution-based re-sampling strategy is used to replace the original IoU threshold-based label assignment strategy, we can achieve an improvement of 0.5 mAP. The distribution-based re-sampling strategy warrants a balanced number of positive samples between different-sized pseudo labels, reconciling the model's supervision signal on small or large objects. However, this improvement is

relatively modest due to the limited supervision information for small objects in pseudo-labels. On the other side, when we incorporate the re-weighting strategy, which includes training loss for both large and small objects (SLA), we can achieve a more notable improvement of 1.8 mAP. This verifies that re-weighting can partially mitigate the issue of imbalanced training samples for different-sized objects during the training process. Additionally, we conduct tests on the optimal value of the hyper-parameter K within the complete S^{3}OD framework, as presented in Table 7d. The best performance is achieved when $K = 2$ and the obtained mAP only waves slightly around the optimal value.

5.4.4. Different strategies of negative learning

For the selection of negative samples, we also compare our proposed TNL with several correlated competitors, including:

- T-BG: Only using the teacher model to select negative samples with high background confidence scores (higher than 0.7).
- T-BG (Reweight): Following the strategy in Soft-teacher (Xu et al., 2021), we use the background confidence scores to reweight the loss of negative samples.
- T-BG (HNS): Under the setting of T-BG, we incorporate low-confidence predictions of the teacher model directly as hard negative samples.
- TNL: Under the setting of T-BG, we incorporate the weighted low-confidence predictions of the teacher model as hard negative samples.

Here, we also calculate the overall false alarm (FA) of the detection results of the validation set under each setting, which reflects the model's ability to discriminate negative samples. A high false alarm represents that the detector struggles to differentiate negative samples, indicating insufficient classification performance. We get $FA = \frac{FP}{P}$, where FP represents the total number of incorrect detections, P represents the number of detected objects and objects with $IoU < 0.5$ between the prediction and ground truth were classified as false positives. From the results of Table 7e, we can observe that directly using T-BG to select negative samples leads to a boom in the false alarm rate. The increase in false alarms accumulates more errors in the network, resulting in performance degradation. Utilizing T-BG to reweight the negative samples can partially restrain the effect of false negative samples and improve performance. Our method effectively utilizes ambiguous predictions generated by the teacher model as challenging negative samples, leading to better performance improvement. Compared to directly introducing these hard negative samples, using confidence score weighting can obtain better results. This is because, among these low-confidence predictions, a small number of positive samples may also exist. Weighting them can reduce the impact of these potential positive samples and avoid confusion between positive and negative samples by the network.

Furthermore, we conduct an empirical comparison concerning the threshold τ_h employed for the selection of hard negative samples within the TNL, as well as the weighting function $w(\cdot)$ applied to the loss attributed to hard negative samples. The outcomes of these comparisons are delineated in Tables 7f and 7g. To maximize the identification of low-quality predictions to get hard negative samples, a threshold value of 0.5 is found to yield optimal performance. Elevating this threshold could potentially incorporate an increased number of accurately identified samples as negatives, thereby diminishing performance. Moreover, considering the inevitable presence of genuine objects within the hard negative samples, a function inversely proportional to the prediction scores s can enable a softer learning approach for hard negative samples. In this study, we have explored several such functions(as shown in Table 7g) and ultimately employed $w(s) = 2(1 - s^2)$ for this purpose.

6. Discussion

Overall, extensive experiments have demonstrated the effectiveness of the proposed S^{3}OD. The achieved mAP for specific categories and the corresponding visualizations also provide sufficient evidence that the S^{3}OD can optimize the detection performance of small objects on semi-supervised object detection in aerial images. To provide straightforward evidence of S^{3}OD's alleviation of size-induced bias, we present the small objects' average predicted scores across various methods after semi-supervised training, which is in Fig. 11. It is evident that, unlike existing SSOD methods that often prioritize the enhancement of large object detection while overlooking small objects in unlabeled images, our S^{3}OD method achieves a more balanced improvement in the predicted confidence across all scales of objects. This observation validates the effectiveness of our method in mitigating the impact that small objects have on semi-supervised object detection in aerial imagery.

However, beyond the challenges posed by numerous small objects, we note that certain objects with extreme geometry characteristics also exhibit subpar performance, such as objects with extreme aspect ratios, including 'bridge' and 'harbor' in DOTA-v1.5. Analysis of the detection results for the 'Harbor' (HB) category in Table 3 reveals that the SSOD method does not yield substantial enhancements for harbors, with our method performing comparably to existing methods in this context. The underlying reason is that objects with these characteristics also face a disadvantaged position in the SSOD pipeline, compared to objects with regular shapes. Similarly, these objects may be overlooked during semi-supervised training. Addressing the detection issues associated with objects exhibiting extreme characteristics necessitates the adoption of differentiated criteria and the selection of more suitable pseudo-label supervision information. This problem also arises in categories such as 'windmill' in the SODA-A dataset. The 'windmill' typically exhibits sparse pixel distribution, and some instances of 'windmill' in aerial images have large aspect ratios. These "disadvantaged" characteristics lead to their being overlooked in the SSOD pipeline, resulting in negligible improvements. Additionally, aerial image object detection also faces an inherent class imbalance issue. Compared to natural scenes, aerial image objects often exhibit a more severe long-tail distribution (Sun et al., 2022), and the class imbalance is further amplified in the process of semi-supervised detection. Categories with a larger number of samples tend to show better detection performance, allowing for the acquisition of more reliable pseudo-labels for supervision. In contrast, categories with fewer samples are more prone to being overlooked within this virtuous cycle. Effectively mitigating these biases within the semi-supervised framework warrants further exploration in future research endeavors.

7. Conclusion

In this study, we have conducted a comprehensive analysis of the challenges impacting the effectiveness of Semi-Supervised Object Detection for aerial imagery. Our investigation reveals a notable oversight in existing SSOD approaches regarding the detection of small objects, which are prevalent in aerial scenes. To tackle this issue, we have developed a novel pipeline, S^{3}OD, specifically tailored to address the various imbalance challenges stemming from small objects. Within S^{3}OD, we introduce an adaptive methodology for selecting scale-aware thresholds, enabling the retention of more supervision information during pseudo-labeling for small objects. Additionally, we propose a label assignment strategy incorporating Gaussian-based sampling and size-aware re-weighting to mitigate imbalance issues arising from limited supervision, ensuring balanced assignments of positive samples across different object scales. Furthermore, we leverage the teacher model's predictions to enhance the negative sample learning process, filtering out possible foreground samples and excavating hard negative samples. By integrating these three novel designs, our S^{3}OD pipeline presents a significant advancement over current state-of-the-art SSOD methods for aerial imagery.

CRediT authorship contribution statement

Ruixiang Zhang: Writing – original draft, Visualization, Validation, Methodology, Investigation. **Chang Xu:** Writing – review & editing, Visualization, Methodology. **Fang Xu:** Writing – review & editing. **Wen Yang:** Writing – review & editing, Supervision, Investigation, Funding acquisition. **Guangjun He:** Writing – review & editing. **Huai Yu:** Writing – review & editing. **Gui-Song Xia:** Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62271355 and NSFC Regional Innovation and Development Joint Fund, China (No. U22A2010). In addition, the numerical calculations in this article have been done on the supercomputing system in the Supercomputing Center, Wuhan University.

References

Bashir, S.M.A., Wang, Y., 2021. Small object detection in remote sensing images with residual feature aggregation-based super-resolution and object detector network. Remote. Sens. 13 (9), 1854.

Chen, B., Chen, W., Yang, S., Xuan, Y., Song, J., Xie, D., Pu, S., Song, M., Zhuang, Y., 2022a. Label matching semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14381–14390.

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

Chen, G., Liu, L., Hu, W., Pan, Z., 2018. Semi-supervised object detection in remote sensing images using generative adversarial networks. In: IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 2503–2506.

Chen, Y., Tan, X., Zhao, B., Chen, Z., Song, R., Liang, J., Lu, X., 2023. Boosting semi-supervised learning by exploiting all unlabeled data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7548–7557.

Cheng, G., Yuan, X., Yao, X., Yan, K., Zeng, Q., Xie, X., Han, J., 2023. Towards large-scale small object detection: Survey and benchmarks. IEEE Trans. Pattern Anal. Mach. Intell. 45 (11), 13467–13488.

Choi, H., Chen, Z., Shi, X., Kim, T.-K., 2022. Semi-supervised object detection with object-wise contrastive learning and regression uncertainty. Br. Mach. Vis. Conf. 1–20.

Courtrai, L., Pham, M.-T., Lefèvre, S., 2020. Small object detection in remote sensing images based on super-resolution with auxiliary generative adversarial networks. Remote. Sens. 12 (19), 3152.

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

Ding, J., Xue, N., Xia, G.-S., Bai, X., Yang, W., Yang, M.Y., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L., 2022. Object detection in aerial images: A large-scale benchmark and challenges. IEEE Trans. Pattern Anal. Mach. Intell. 44 (11), 7778–7796.

Guo, Q., Mu, Y., Chen, J., Wang, T., Yu, Y., Luo, P., 2022. Scale-equivalent distillation for semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14522–14531.

Han, J., Ding, J., Li, J., Xia, G.-S., 2021. Align deep features for oriented object detection. IEEE Trans. Geosci. Remote Sens. 60, 1–11.

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

Jeong, J., Lee, S., Kim, J., Kwak, N., 2019. Consistency-based semi-supervised learning for object detection. Adv. Neural Inf. Process. Syst. 32.

Jeong, J., Verma, V., Hyun, M., Kannala, J., Kwak, N., 2021. Interpolation-based semi-supervised learning for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11602–11611.

Kar, P., Chudasama, V., Onoe, N., Wasnik, P., 2023. Revisiting class imbalance for end-to-end semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4569–4578.

Kim, B., Choo, J., Kwon, Y.-D., Joe, S., Min, S., Gwon, Y., 2021. Selfmatch: Combining contrastive self-supervision and consistency for semi-supervised learning. arXiv preprint arXiv:2101.06480.

Kim, J., Jang, J., Seo, S., Jeong, J., Na, J., Kwak, N., 2022. Mum: Mix image tiles and unmix feature tiles for semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14512–14521.

Lee, D.-H., et al., 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on Challenges in Representation Learning, ICML. Vol. 3, p. 896.

Li, W., Chen, Y., Hu, K., Zhu, J., 2022c. Oriented reppoints for aerial object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1829–1838.

Li, Z., Hou, B., Wu, Z., Jiao, L., Ren, B., Yang, C., 2021. Fcosr: A simple anchor-free rotated detector for aerial object detection. arXiv preprint arXiv:2111.10780.

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

Li, H., Wu, Z., Shrivastava, A., Davis, L.S., 2022b. Rethinking pseudo labels for semi-supervised object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36, pp. 1314–1322.

Liang, D., Geng, Q., Wei, Z., Vorontsov, D.A., Kim, E.L., Wei, M., Zhou, H., 2022. Anchor retouching via model interaction for robust object detection in aerial images. IEEE Trans. Geosci. Remote Sens. 60, 1–13.

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

Liu, N., Celik, T., Li, H.-C., 2022a. Gated ladder-shaped feature pyramid network for object detection in optical remote sensing images. IEEE Geosci. Remote. Sens. Lett. 19, 1–5.

Liu, N., Celik, T., Zhao, T., Zhang, C., Li, H.-C., 2021a. AFDet: Toward more accurate and faster object detection in remote sensing images. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 14, 12557–12568.

Liu, Y.-C., Ma, C.-Y., He, Z., Kuo, C.-W., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P., 2021b. Unbiased teacher for semi-supervised object detection. Int. Conf. Learn. Represent. 1–17.

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

Liu, N., Xu, X., Gao, Y., Zhao, Y., Li, H.-C., 2024. Semi-supervised object detection with uncurated unlabeled data for remote sensing images. Int. J. Appl. Earth Obs. Geoinf. 129, 103814.

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

Liu, L., Zhang, B., Zhang, J., Zhang, W., Gan, Z., Tian, G., Zhu, W., Wang, Y., Wang, C., 2023b. 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. pp. 7370–7379.

Ma, J., Shao, W., Ye, H., Wang, L., Wang, H., Zheng, Y., Xue, X., 2018. Arbitrary-oriented scene text detection via rotation proposals. IEEE Trans. Multimed. 20 (11), 3111–3122.

Mi, P., Lin, J., Zhou, Y., Shen, Y., Luo, G., Sun, X., Cao, L., Fu, R., Xu, Q., Ji, R., 2022. Active teacher for semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14482–14491.

Ming, Q., Miao, L., Zhou, Z., Song, J., Dong, Y., Yang, X., 2023. Task interleaving and orientation estimation for high-precision oriented object detection in aerial images. ISPRS J. Photogramm. Remote Sens. 196, 241–255.

Qiao, Y., Miao, L., Zhou, Z., Ming, Q., 2023. A novel object detector based on high-quality rotation proposal generation and adaptive angle optimization. IEEE Trans. Geosci. Remote Sens. 61, 1–15.

Ren, S., He, K., Girshick, R., Sun, J., 2016. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39 (6), 1137–1149.

Shen, J., Zhang, C., Yuan, Y., Wang, Q., 2023. Enhancing prospective consistency for semi-supervised object detection in remote sensing images. IEEE Trans. Geosci. Remote Sens. 61, 1–12.

Shermeyer, J., Van Etten, A., 2019. The effects of super-resolution on object detection performance in satellite imagery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 1–10.

Shi, F., Zhang, T., Zhang, T., 2020. Orientation-aware vehicle detection in aerial images via an anchor-free object detection approach. IEEE Trans. Geosci. Remote Sens. 59 (6), 5221–5233.

Shrivastava, A., Gupta, A., Girshick, R., 2016. Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 761–769.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.-L., 2020a. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Adv. Neural Inf. Process. Syst. 33, 596–608.

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

Sun, X., Wang, P., Yan, Z., Xu, F., Wang, R., Diao, W., Chen, J., Li, J., Feng, Y., Xu, T., et al., 2022. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. ISPRS J. Photogramm. Remote Sens. 184, 116–130.

Tang, Y., Chen, W., Luo, Y., Zhang, Y., 2021. Humble teachers teach better students for semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3132–3141.

Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Adv. Neural Inf. Process. Syst. 30.

Vandeghen, R., Louppe, G., Van Droogenbroeck, M., 2022. Adaptive self-training for object detection. arXiv preprint arXiv:2212.05911.

Wang, P., Cai, Z., Yang, H., Swaminathan, G., Vasconcelos, N., Schiele, B., Soatto, S., 2022c. Omni-DETR: Omni-supervised object detection with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9367–9376.

Wang, J., Li, F., Bi, H., 2022a. Gaussian focal loss: Learning distribution polarized angle prediction for rotated object detection in aerial images. IEEE Trans. Geosci. Remote Sens. 60, 1–13.

Wang, Z., Li, Y., Guo, Y., Fang, L., Wang, S., 2021. Data-uncertainty guided multi-phase learning for semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4568–4577.

Wang, J., Lukasiewicz, T., Massiceti, D., Hu, X., Pavlovic, V., Neophytou, A., 2022b. Np-match: When neural processes meet semi-supervised learning. In: International Conference on Machine Learning. PMLR, pp. 22919–22934.

Wang, Y., Wang, H., Shen, Y., Fei, J., Li, W., Jin, G., Wu, L., Zhao, R., Le, X., 2022d. Semi-supervised semantic segmentation using unreliable pseudo-labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4248–4257.

Wang, X., Yang, X., Zhang, S., Li, Y., Feng, L., Fang, S., Lyu, C., Chen, K., Zhang, W., 2023. 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. pp. 3240–3249.

Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L., 2018. DOTA: A large-scale dataset for object detection in aerial images. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3974–3983.

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

Xu, B., Chen, M., Guan, W., Hu, L., 2023a. Efficient teacher: Semi-supervised object detection for YOLOv5. arXiv preprint arXiv:2302.07577.

Xu, C., Ding, J., Wang, J., Yang, W., Yu, H., Yu, L., Xia, G.-S., 2023b. Dynamic coarse-to-fine learning for oriented tiny object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7318–7328.

Xu, C., Wang, J., Yang, W., Yu, H., Yu, L., Xia, G.-S., 2022a. Detecting tiny objects in aerial images: A normalized Wasserstein distance and a new benchmark. ISPRS J. Photogramm. Remote Sens. 190, 79–93.

Xu, C., Wang, J., Yang, W., Yu, H., Yu, L., Xia, G.-S., 2022b. RFLA: Gaussian receptive field based label assignment for tiny object detection. In: Proceedings of the European Conference on Computer Vision. Springer, pp. 526–543.

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

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

Yang, X., Yan, J., Liao, W., Yang, X., Tang, J., He, T., 2022. Scrdet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing. IEEE Trans. Pattern Anal. Mach. Intell. 45 (2), 2384–2399.

Yang, X., Yan, J., Ming, Q., Wang, W., Zhang, X., Tian, Q., 2021b. Rethinking rotated object detection with gaussian Wasserstein distance loss. In: International Conference on Machine Learning. PMLR, pp. 11830–11841.

Yang, X., Yang, J., Yan, J., Zhang, Y., Zhang, T., Guo, Z., Sun, X., Fu, K., 2019. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8232–8241.

Yang, X., Yang, X., Yang, J., Ming, Q., Wang, W., Tian, Q., Yan, J., 2021c. Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. Adv. Neural Inf. Process. Syst. 34, 18381–18394.

Zhang, R., Guo, H., Xu, F., Yang, W., Yu, H., Zhang, H., Xia, G.-S., 2022c. Optical-enhanced oil tank detection in high-resolution SAR images. IEEE Trans. Geosci. Remote Sens. 60, 1–12.

Zhang, J., Lin, X., Zhang, W., Wang, K., Tan, X., Han, J., 2023. Semi-DETR: Semi-supervised object detection with detection transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23809–23818.

Zhang, F., Pan, T., Wang, B., 2022a. Semi-supervised object detection with adaptive class-rebalancing self-training. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36, pp. 3252–3261.

Zhang, L., Sun, Y., Wei, W., 2022b. Mind the gap: Polishing pseudo labels for accurate semi-supervised object detection. arXiv preprint arXiv:2207.08185.

Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., Shinozaki, T., 2021. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. Adv. Neural Inf. Process. Syst. 34, 18408–18419.

Zhang, Y., Yao, X., Liu, C., Chen, F., Song, X., Xing, T., Hu, R., Chai, H., Xu, P., Zhang, G., 2022d. S4od: Semi-supervised learning for single-stage object detection. arXiv preprint arXiv:2204.04492.

Zheng, M., You, S., Huang, L., Wang, F., Qian, C., Xu, C., 2022. Simmatch: Semi-supervised learning with similarity matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14471–14481.

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

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

Zhou, Q., Yu, C., Wang, Z., Qian, Q., Li, H., 2021. Instant-teaching: An end-to-end semi-supervised object detection framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4081–4090.

Zoph, B., Ghiasi, G., Lin, T.-Y., Cui, Y., Liu, H., Cubuk, E.D., Le, Q., 2020. Rethinking pre-training and self-training. Adv. Neural Inf. Process. Syst. 33, 3833–3845.