Adapting Object Size Variance and Class Imbalance for Semi-supervised Object Detection

Yuxiang Nie[1]*, Chaowei Fang[2]*, Lechao Cheng[3], Liang Lin[1], Guanbin Li[1,4]†

[1] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2] School of Artificial Intelligence, Xidian University, Xi'an, China
[3] Zhejiang Lab, Hangzhou, China
[4] Research Institute, Sun Yat-sen University, Shenzhen, China
nieyx3@mail2.sysu.edu.cn, chaoweifang@outlook.com, chenglc@zhejianglab.com, linliang@ieee.org, liguanbin@mail.sysu.edu.cn

Abstract

Semi-supervised object detection (SSOD) attracts extensive research interest due to its great significance in reducing the data annotation effort. Collecting high-quality and category-balanced pseudo labels for unlabeled images is critical to addressing the SSOD problem. However, most of the existing pseudo-labeling-based methods depend on a large and fixed threshold to select high-quality pseudo labels from the predictions of a teacher model. Considering different object classes usually have different detection difficulty levels due to scale variance and data distribution imbalance, conventional pseudo-labeling-based methods are arduous to explore the value of unlabeled data sufficiently. To address these issues, we propose an adaptive pseudo labeling strategy, which can assign thresholds to classes with respect to their "hardness". This is beneficial for ensuring the high quality of easier classes and increasing the quantity of harder classes simultaneously. Besides, label refinement modules are set up based on box jittering for guaranteeing the localization quality of pseudo labels. To further improve the algorithm's robustness against scale variance and make the most of pseudo labels, we devise a joint feature-level and prediction-level consistency learning pipeline for transferring the information of the teacher model to the student model. Extensive experiments on COCO and VOC datasets indicate that our method achieves state-of-the-art performance. Especially, it brings mean average precision gains of 2.08 and 1.28 on MS-COCO dataset with 5% and 10% labeled images, respectively.

Introduction

Large-scale training data has promoted significant progress in object detection based on convolutional neural networks (CNN). However, collecting annotations for a large number of images is costly and time-consuming. Therefore, increasing attention is drawn to semi-supervised object detection (SSOD) which can take advantage of a large number of unlabeled images during the training stage.

Existing SSOD methods mainly rely on pseudo labeling algorithms to involve unlabeled images during training. (Sohn et al. 2020) utilize a pre-trained model to produce fixed pseudo labels. The other kind of methods (Jeong et al. 2019, 2021; Tang et al. 2021a) employ the object detection model to generate supervision signals on unlabeled images for itself. (Jeong et al. 2019; Tang et al. 2021a) are implemented based on the consistency constraints between predictions inferred from differently augmented images. (Jeong et al. 2021) relies on the image mixup operation to construct the consistency constraint. However, such a kind of method may be easily tracked into local minima and interfered by noisy predictions. Another kind of methods, such as (Xu et al. 2021; Wang et al. 2022a), depend on the mean teacher model (Tarvainen and Valpola 2017) to generate pseudo labels for unlabeled images. As shown in Fig. 1, we can observe that imbalanced data distribution and size variance are two key factors for limiting the overall detection performance, due to the high difficulty in detecting few-shot or small objects. Previous SSOD methods based on a large fixed threshold can only guarantee the high quality of pseudo-labels for easy classes, but disregard objects of relatively hard classes. Li et al. (2022b) assign dynamic

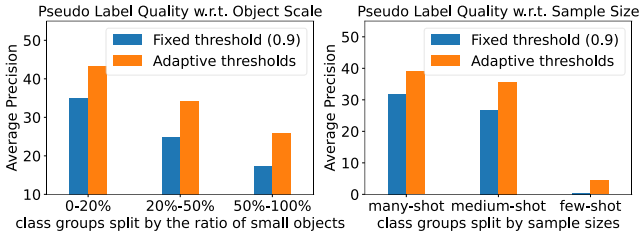


Figure 1: The pseudo label quality of models which leverage pseudo labeling strategy based on a fixed threshold (0.9) or our devised class-wise adaptive thresholds to explore unlabeled images on MS-COCO dataset with 10% labeled images. Here, an object is regarded as a small object if its area is smaller than 32×32. The sample number setting for many-shot, medium-shot, and few-shot classes are more than 4,000, 100–4,000, and less than 100, respectively. Classes dominated by small objects or having few samples severely affect pseudo label quality. Our proposed labeling method can greatly alleviate this issue.

*The first two authors contribute equally.
†The corresponding author is Guanbin Li.

thresholds to classes based on their classification and box regression confidences. However, this method tends to decrease thresholds for all classes and may be corrupted by incorrect pseudo labels. For improving the robustness against objects' scale variance, Guo et al. (2022) propose to regularize the consistency between predictions derived from images at different scales.

In this paper, we propose a novel adaptive pseudo labeling algorithm which is capable of increasing the quantity of hard classes without affecting the quality of easy classes. We devise an online hardness measure through accumulating the class-wise prediction confidences temporally, which is referred to set adaptive thresholds for screening pseudo labels of different classes. To increase the tolerance against noises in pseudo labels, we collect prediction confidences of top-K classes during the calculation of the hardness measure, instead of only considering the most confident class as with in previous adaptive threshold estimation methods (Li et al. 2022b; Chen et al. 2022a). Considering the localization quality can not be guaranteed by the labeling algorithm based on thresholding classification confidences, we further refine the positions of pseudo labels based on box jittering. The effectiveness of our devised pseudo labeling algorithm in improving pseudo labels can be observed in Fig. 1.

For the purpose of increasing the utilization of pseudo labels and enhancing the robustness against objects' scale variance, we devise a joint feature-level and prediction-level consistency learning framework. First, increasing the intra-class compactness in the feature space can alleviate the dependence on the quantity of annotations and achieve the effect of automatic label correction. Hence, we propose a consistency learning strategy to pull close features extracted by the teacher and student models on different augmentations of training images. Then, the prediction-level consistency learning is implemented via leveraging pseudo labels to regularize classification and regression predictions of the student model. The above consistency learning involves both intra-scale and inter-scale consistency constraints for exploring pseudo labels sufficiently and strengthening the capacity in coping with small objects. Extensive experiments are conducted on MS-COCO (Lin et al. 2014) and PASCAL VOC (Everingham et al. 2010) datasets, indicating that our method achieves state-of-the-art performance. Comparisons of our method against the baseline trained with labeled images only and existing methods (Sohn et al. 2020; Xu et al. 2021; Chen et al. 2022b) are provided in Fig. 2.

Our main contributions are summarized as follows:

- We propose an adaptive pseudo labeling algorithm based on class-wise and noise-tolerant adaptive confidence thresholds and box position refinement.
- We set up a joint feature-level and prediction-level consistency learning framework for increasing the utilization of pseudo labels and enhancing the detection robustness against small objects.
- We conduct extensive experiments on MS-COCO and PASCAL VOC datasets, which verify that our method outperforms existing methods significantly.

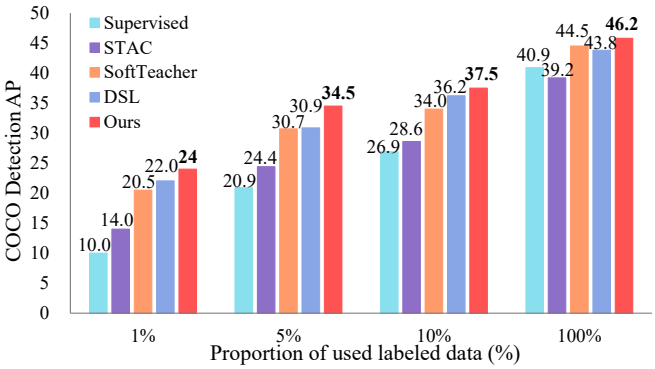


Figure 2: Comparisons of our method against the baseline model supervised with labeled images only and existing methods, STAC (Sohn et al. 2020), SoftTeacher (Xu et al. 2021) and DSL (Chen et al. 2022b) on MS-COCO dataset. The average precision (AP) is used for evaluation.

Related Work

Object Detection

Object detection (Girshick et al. 2014; Girshick 2015; Ren et al. 2015; Redmon et al. 2016; Tian et al. 2019) is a fundamental task in the realm of computer vision. Single-stage techniques (Redmon et al. 2016; Lin et al. 2017b; Tian et al. 2019; Zhang et al. 2020) generate predictions based on anchors or a grid of potential object centers. Two-stage detectors (Girshick 2015; Ren et al. 2015; Cai and Vasconcelos 2019) make predictions based on proposals. However, the complicate post-processing procedures like non-maximum suppression or anchors generation leads to expensive computational burden. Carion et al. (2020) propose an efficient end-to-end detection framework termed DETR, which can explicitly encode prior knowledge based on vision transformer. Fang et al. (2022) incorporate the masked image modeling (MIM) to pre-train vision transformer models for object detection, yielding promising performance gains.

Semi-supervised Object Detection

Semi-supervised object detection (SSOD) approaches have attracted increasing attention since they can decrease the demand for labor-consuming annotations. STAC (Sohn et al. 2020) employs regularization on strongly and weakly augmented images for SSOD. However, this kind of strategy is unable to update pseudo labels dynamically during the network training procedure. Inspired by MeanTeacher (Tarvainen and Valpola 2017), many SSOD approaches (Xu et al. 2021; Tang et al. 2021b; Zhou et al. 2021; Liu et al. 2021; Li et al. 2022a; Li, Yuan, and Li 2022; Zhang, Pan, and Wang 2022) leverage the exponential moving average (EMA) strategy to construct the teacher model which can evolve as the optimization of the student model. Nevertheless, these methods can not produce pseudo labels with sufficiently high quality, especially for relatively hard classes. We propose an adaptive pseudo labeling algorithm which can ensure the high quality for easy classes while increasing the recall of hard classes. PseCo (Li et al. 2022a) and

SED (Guo et al. 2022) share a similar spirit to regularize the prediction consistency across different views of training images. In this paper, we introduce a consistency learning framework which can directly pull close features extracted from different views of training images, which is more effective in reducing the intra-class variation of features.

Method

This paper is targeted at tackling the semi-supervised object detection task. Namely, only a small proportion of labeled data $\mathbb{D}_l$ and a large amount of unlabeled data $\mathbb{D}_u$ are available for training models. We assume $\mathbb{D}_l = \{x_i^l, y_i^l\}_{i=1}^{N_l}$ and $\mathbb{D}_u = \{x_i^u\}_{i=1}^{N_u}$. x_i^l and x_i^u denote the i-th labeled and unlabeled image, respectively. y_i^l consists of bounding box annotations and their category labels in x_i^l. N_l and N_u are the number of labeled images and unlabeled images, respectively. We assume the number of target object classes is C. The goal is to leverage unlabeled images to boost the performance of detection models.

Overview

The pipeline of our approach is illustrated in Fig. 3. Our method is built upon the mean teacher framework. The teacher model is formed by accumulating weights of the student model temporally. It produces pseudo labels on unlabeled images which are used to guide the training of the student model. Most of existing methods select pseudo labels by thresholding their confidence scores with a fixed constant. However, such a labeling mechanism is difficult to balance the precision and recall for all classes. A small threshold leads to noisy pseudo labels while a large threshold causes severe miss inspection to some classes. To overcome this issue, we propose an adaptive pseudo labeling algorithm with the help of category-adaptive label selection and bounding box refinement. Based on the obtained pseudo labels, we devise a multi-scale feature-level and prediction-level consistency learning strategy to train the student model. The feature-level consistency learning helps to generate features with small intra-class variations, which can relieve the dependence on the quantity of annotations and achieve the effect of automatic label propagation. The prediction-level consistency learning benefits in exploring pseudo labels sufficiently.

Teacher-Student Learning Framework

The teacher-student learning framework is composed of a teacher model and a student model. We denote the parameter of the student model at the t-th training step be θ_t^{stu}. The parameter of the teacher model θ_t^{tea} is obtained by temporally accumulating θ_t^{stu},

$$\theta_t^{tea} = \lambda\theta_{t-1}^{tea} + (1-\lambda)\theta_t^{stu}, \tag{1}$$

where λ is a constant. Here, we set up the teacher and student models with FasterRCNN (Ren et al. 2015).

Given an unlabeled image x_i^u, its pseudo label defined by $\hat{y}_i^u$ can be inferred from the teacher model's predictions. For labeled images, the following training loss $\mathcal{L}_{sup}$ can be used to guide the training of the student model:

$$\begin{aligned} L_{sup} = \frac{1}{N_l}\sum_{i=1}^{N_l}[&\mathcal{L}_{cls}^{rpn}(x_i^l, y_i^l) + \mathcal{L}_{reg}^{rpn}(x_i^l, y_i^l) + \\ &\mathcal{L}_{cls}^{roi}(x_i^l, y_i^l) + \mathcal{L}_{reg}^{roi}(x_i^l, y_i^l)], \end{aligned} \tag{2}$$

$\mathcal{L}_{cls}^{rpn}(\cdot)$ and $\mathcal{L}_{reg}^{rpn}(\cdot)$ represents the classification and regression loss function respectively, which are used for constraining the outputs of the region proposal network (RPN). $\mathcal{L}_{cls}^{roi}(\cdot)$ and $\mathcal{L}_{reg}^{roi}(\cdot)$ also represents the classification and regression loss function respectively, which are used for constraining the outputs of the ROI (short for region of interest) head. Unlabeled images are explored for network training as introduced below.

Adaptive Pseudo Labeling

The key to the success of the teacher-student learning framework is generating high-quality pseudo labels for unlabeled labels. Wrong labels cause fatal interference to the learning of the student model, while a low recall rate harms the utilization rate of unlabeled images. However, most existing SSOD methods simply rely on a uniform high classification confidence threshold to filter noisy predictions for all classes, which is difficult in balancing the precision and the recall of object boxes. Besides, few of them are targeted at improving the localization quality of bounding boxes. To solve the above issues, we design an adaptive pseudo labeling mechanism, which dynamically assigns adaptive thresholds to different classes and takes the box location refinement into account as well.

1) Label Selection based on Adaptive Thresholds. We generate pseudo labels for unlabeled images by selecting confident boxes predicted by the teacher model. Due to factors such as imbalanced class distribution and object scales, the "hardness" for learning different classes is usually inconsistent. Inspired by FreeMatch (Wang et al. 2022b), we design an adaptive thresholding strategy for pseudo label selection.

At the t-th training iteration, we suppose that the teacher model produces N_t^{tea} objects from input images, i.e., $\{(\mathbf{o}_{t,j}^{tea}, \mathbf{p}_{t,j}^{tea})\}_{j=1}^{N_t^{tea}}$, where $\mathbf{o}_{t,j}^{tea} \in \mathbb{R}^4$ and $\mathbf{p}_{t,j}^{tea} \in [0,1]^C$ represent the box position and class probability predictions, respectively. We define the class-wise adaptive confidence thresholds as $\boldsymbol{\tau}_t$. To estimate $\boldsymbol{\tau}_t$, we first average the probabilities of confident object predictions according to the following formulation,

$$\bar{p}_t'[c] = \frac{\sum_{j=1}^{N_t^{tea}} \text{find}(c, \text{Top}_K(\mathbf{p}_{t,j}^{tea})) \times p_{t,j}^{tea}[c]}{\sum_{j=1}^{N_t^{tea}} \text{find}(c, \text{Top}_K(\mathbf{p}_{t,j}^{tea}))}. \tag{3}$$

Here, $p_{t,j}^{tea}[c]$ represents the c-th element in $\mathbf{p}_{t,j}^{tea}$, and $\bar{p}_t'[c]$ indicates the averaged probability value of the c-th class. $\text{Top}_K(\cdot)$ returns the set of class indices having top K largest probabilities. $\text{find}(\cdot)$ returns 1 if the first input is in the second input; otherwise, it returns 0. The confidence level of the teacher model on the c-th class is dynamically estimated via the moving average operation, i.e., $\bar{p}_t[c] = \gamma\bar{p}_{t-1}[c] + (1-\gamma)\bar{p}_t'[c]$ where $\bar{p}_0[c] = 1/C$.

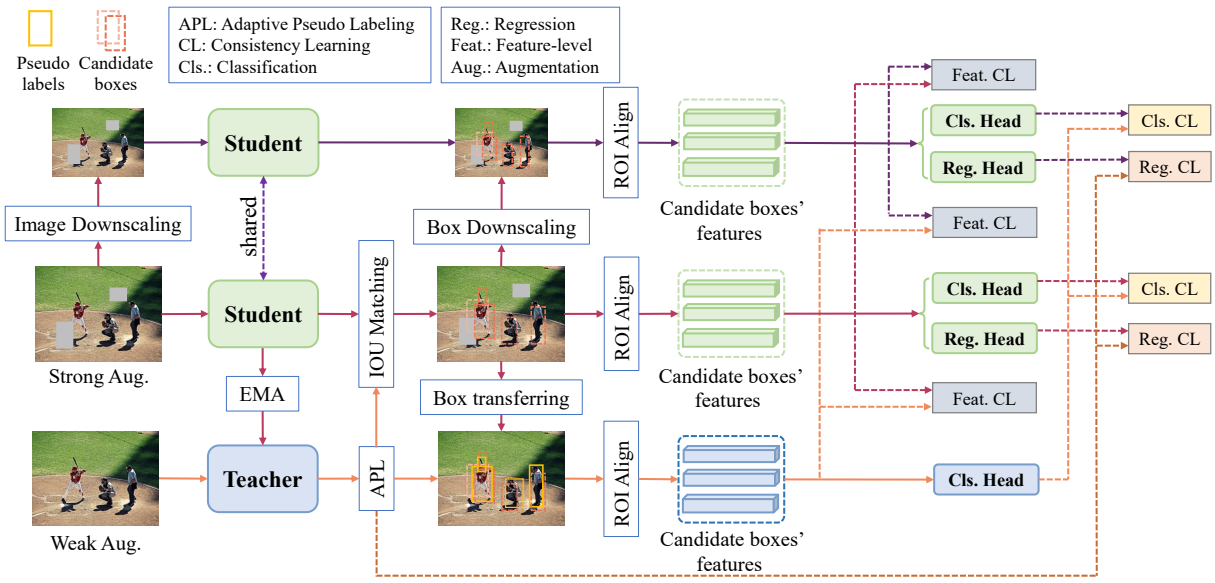


Figure 3: The overall structure of the method. Pseudo labels are generated on weakly augmented images by the teacher model according to adaptive pseudo labeling (APL). The student model is guided by the teacher with feature-level and prediction-level consistency learning. Note that the vanilla supervised learning branch uses supervised data only, to train the student model, which is not plotted in the figure.

The confidence levels can reflect the hardness of classes. Objects of classes with lower confidence levels are more easily ignored during the thresholding-based label selection. Hence, we propose to adaptively assign thresholds to different classes. Those less confident classes need to lower thresholds to ensure a higher recall rate. On the other hand, a large threshold should be preserved for highly confident classes to guarantee the quality of their pseudo labels. To achieve these goals, we calculate the thresholding value $\tau_k[c]$ for the c-th class via the following formulation,

$$\tau_k[c] = \frac{\bar{p}_k[c]}{\max_{c' \in \mathbb{N}_C} \bar{p}_k[c']} \tau_0, \tag{4}$$

where τ_0 is a constant, and $\mathbb{N}_C = \{1, 2, \cdots, C\}$. The pseudo labels for unlabeled images are collected by checking whether the class probability of each predicted box is larger than the corresponding threshold.

2) Box Refinement. The object localization quality can not be guaranteed by the thresholding operation based on classification confidences, since deviated boxes may receive high confidences as well (Xu et al. 2021). Especially, the adaptive thresholds allow part of boxes with low confidence to be used as pseudo labels. This may bring in more pseudo labels with low localization quality. To address these issues, we propose a box refinement algorithm based on box jittering. Practically, for each object pseudo label $(\mathbf{o}, \mathbf{p})$, we first perturb the box by randomly shifting its boundaries horizontally or vertically with the jitter scale r. The perturbation process is repeated by M times. We denote the m-th deviated box be $\mathbf{o}^{(m)}$, which is fed into the ROI head, deriving a rectified box $\hat{\mathbf{o}}^{(m)}$ accompanied with the classification probability vector $\hat{\mathbf{p}}^{(m)}$. The final location of the box is formed by summing up the rectified boxes with weights of classification confidences,

$$\hat{\mathbf{o}} = \frac{\sum_{m=1}^{M} \max_{c \in \mathbb{N}_C} \hat{p}^{(m)}[c] \times \hat{\mathbf{o}}^{(m)}}{\sum_{m=1}^{M} \max_{c \in \mathbb{N}_C} \hat{p}^{(m)}[c]}. \tag{5}$$

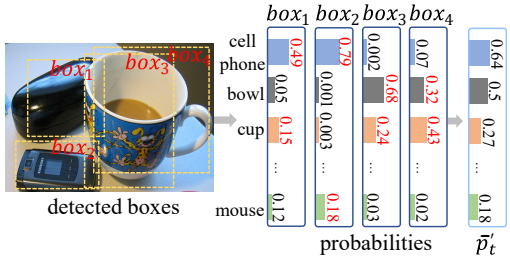


Figure 4: Illustration of adaptive dynamic thresholding. Red bold number is the probability of confident objects (TOP 2). $\bar{p}'_t$ is the mean probability of confident objects.

With help of the above pseudo labeling strategy, we can collect pseudo labels with high recall and precise object localization for unlabeled images. This can greatly mitigate the phenomenon that objects of relatively hard classes are neglected during pseudo labeling.

Multi-scale Consistency Learning

1) Feature-level consistency learning. Learning features with high generalization ability helps to relieve the dependence on large amount of annotations. Hence, we devise a self-supervised feature-level consistency learning strategy.

It is beneficial for reducing intra-class variations of features and enhancing the robustness against appearance variations.

Given an unlabeled image x_i, we first input its weakly augmented variant into the teacher model, resulting in an intermediate feature map $\mathbf{f}_i^{tea}$ produced by the backbone model. Meanwhile, a set of pseudo labels can be generated with the adaptive pseudo labeling algorithm, i.e., $\hat{y}_i = \{(\mathbf{o}_{i,j}^{tea}, \mathbf{p}_{i,j}^{tea})\}_{j=1}^{N_i^{tea}}$ in which N_i^{tea} represents the number of pseudo labels (the training iteration index t is neglected for conciseness). Then, we create other two variants of $\mathbf{x}_i$: $\mathcal{A}_1^{str}(\mathbf{x}_i)$ is a strongly augmented variant of $\mathbf{x}_i$ and has the same size of $\mathbf{x}_i$; $\mathcal{A}_2^{str}(\mathbf{x}_i)$ is created by downsampling the spatial dimensions of $\mathcal{A}_1^{str}(\mathbf{x}_i)$ to halves. $\mathcal{A}_1^{str}(\mathbf{x}_i)$ and $\mathcal{A}_2^{str}(\mathbf{x}_i)$ are fed into the backbone of the student model, deriving two feature maps $\mathbf{f}_i^1$ and $\mathbf{f}_i^2$, respectively. We further input $\mathbf{f}_i^1$ into the RPN module and apply the IoU matching operation to select out N_i^{can} candiadte boxes $\{\mathbf{o}_{i,j}^{can} | j = 1, \cdots, N_i^{can}\}$ from the object proposals inferred by RPN. Afterwards, we can extract three feature vectors from different views for $\mathbf{o}_{i,j}^{can}$, namely,

$$\mathbf{f}_{i,j}^{tea} = \text{MLP}(\text{RoIAlign}(\mathbf{f}_i^{tea}, \mathcal{M}(\mathbf{o}_{i,j}^{can}))), \tag{6}$$

$$\mathbf{f}_{i,j}^{1} = \text{MLP}(\text{RoIAlign}(\mathbf{f}_i^{1}, \mathbf{o}_{i,j}^{can})), \tag{7}$$

$$\mathbf{f}_{i,j}^{2} = \text{MLP}(\text{RoIAlign}(\mathbf{f}_i^{2}, \mathcal{T}_{\downarrow_2}(\mathbf{o}_{i,j}^{can}))). \tag{8}$$

$\mathcal{M}(\cdot)$ transforms the boxes from strongly augmented image to weakly augmented image. $\mathcal{T}_{\downarrow_2}(\cdot)$ is a function that downscales the width and height of the input box to half. RoIAlign$(\cdot)$ is a function which extracts a feature vector from the input feature map for the input box (He et al. 2017). MLP$(\cdot)$ represents a two-layer perceptron.

Finally, inspired by (Chen and He 2021), we devise a multi-scale feature-level consistency learning strategy. First, a projection head composed of a three-layer perceptron is adopted to encode $\mathbf{f}_{i,j}^{tea}$, $\mathbf{f}_{i,j}^{1}$, and $\mathbf{f}_{i,j}^{2}$ into latent feature vectors $\mathbf{e}_{i,j}^{tea}$, $\mathbf{e}_{i,j}^{1}$, and $\mathbf{e}_{i,j}^{2}$, respectively. Then, a prediction head composed of a two-layer perceptron is employed to transfer these latent feature vectors into centroid features $\mathbf{z}_{i,j}^{tea}$, $\mathbf{z}_{i,j}^{1}$, and $\mathbf{z}_{i,j}^{2}$, respectively. A training loss aiming at pulling close latent and centroid features is devised for optimizing network parameters, which is formulated as below,

$$\begin{aligned} L_{i,j}^{rep} = &\frac{1}{6}\{\sum_{l=1}^{2}[\ell_{cos}(\mathbf{z}_{i,j}^{tea}, \mathcal{S}_g(\mathbf{e}_{i,j}^{l})) + \ell_{cos}(\mathbf{z}_{i,j}^{l}, \mathcal{S}_g(\mathbf{e}_{i,j}^{tea}))] \\ &+ \ell_{cos}(\mathbf{z}_{i,j}^{1}, \mathcal{S}_g(\mathbf{e}_{i,j}^{2})) + \ell_{cos}(\mathbf{z}_{i,j}^{2}, \mathcal{S}_g(\mathbf{e}_{i,j}^{1}))\}, \end{aligned} \tag{9}$$

where $\mathcal{S}_g(\cdot)$ is the stop-gradient function; $\ell_{cos}(\mathbf{z}, \mathbf{e})$ measures the cosine distance between the two input feature vectors $\mathbf{z}$ and $\mathbf{e}$, i.e.,

$$\ell_{cos}(\mathbf{z}, \mathbf{e}) = -\frac{\mathbf{z} \cdot \mathbf{e}}{\|\mathbf{z}\|_2 \times \|\mathbf{e}\|_2}, \tag{10}$$

where $\cdot$ is the inner product operation. The final loss of the multi-scale feature-level consistency learning is as below,

$$L_{rep} = \frac{1}{N_u}\sum_{i=1}^{N_u}\frac{1}{N_i^{can}}\sum_{j=1}^{N_i^{can}} L_{i,j}^{rep}. \tag{11}$$

The above consistency learning strategy helps to condense objects' features inside each class. However, the semantic meaning of features can not be ensured by it. Hence, we further leverage the pseudo labels to construct multi-scale prediction-level consistency constraints.

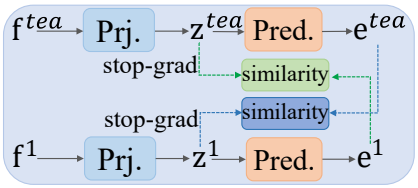


Figure 5: Illustration of feature-level consistency learning by siamese representation learning.

2) Prediction-level consistency learning. In this section, we set up the multi-scale prediction-level consistency constraints. To make full use of teacher guidance, we employ the double-check mechanism (Wang et al. 2022a) to guide the student model. First, the feature vector $\mathbf{f}_{i,j}^{tea}$ in Eq. (6) is fed into the classification head of the teacher model, deriving the probability target vector $\mathbf{p}_{i,j}^{tgt}$ for $\mathbf{o}_{i,j}^{can}$. Then, We feed the feature vector $\mathbf{f}_{i,j}^{1}$ in Eq. (7) into the box regression head and the classification head of the student model, which produce $\mathbf{o}_{i,j}^{1}$ and $\mathbf{p}_{i,j}^{1}$, respectively. Similarly, $\mathbf{f}_{i,j}^{2}$ in Eq. (8) induces $\mathbf{o}_{i,j}^{2}$ and $\mathbf{p}_{i,j}^{2}$.

For the consistency constraint on classification probability, we sharpen the values of probability target vectors which have relatively high confidence. We estimate the confidence level of a probability target vector by summing up the probability values of top S classes. Denote the confidence level of $\mathbf{p}_{i,j}^{tgt}$ be $p_{i,j}^{con}$, which is calculated as below,

$$p_{i,j}^{con} = \sum_{c' \in \text{Top}_S(\mathbf{p}_{i,j}^{tgt})} p_{i,j}^{tgt}[c']. \tag{12}$$

When $p_{i,j}^{con}$ is larger than a constant κ, we sharpen the scores in $\mathbf{p}_{i,j}^{tgt}$, according to the following rules,

$$p_{i,j}^{tgt}[c] = \begin{cases} \dfrac{p_{i,j}^{tgt}[c]}{p_{i,j}^{con}}, & c \in \text{Top}_S(\mathbf{p}_{i,j}^{tgt}), \\ 0, & \text{otherwise}. \end{cases} \tag{13}$$

Here, we empirically set $S = 3$ and $\kappa = 0.8$. Then, the soft cross entropy function $\ell_{cls}(\cdot)$ is used for calculating the consistency loss on classification probabilities,

$$L_{i,j}^{cls} = \ell_{cls}(\mathbf{p}_{i,j}^{1}, \mathbf{p}_{i,j}^{tgt}) + \ell_{cls}(\mathbf{p}_{i,j}^{2}, \mathbf{p}_{i,j}^{tgt}). \tag{14}$$

For consistency constraint on box regression, we regard the box of the matched pseudo label as the target for each candidate box. Denote the target box for $\mathbf{o}_{i,j}^{can}$ be $\mathbf{o}_{i,j}^{tgt}$. The following training loss is utilized to constrain multi-scale box regression of the student model,

$$L_{i,j}^{reg} = \ell_{reg}(\mathbf{o}_{i,j}^{1}, \mathbf{o}_{i,j}^{tgt}) + \ell_{reg}(\mathbf{o}_{i,j}^{2}, T_{\downarrow_2}(\mathbf{o}_{i,j}^{tgt})), \tag{15}$$

where $\ell_{reg}(\cdot)$ is the box regression loss function based on smoothed L_1 norm. The final multi-scale prediction consistency loss can be summarized as below,

$$L_{con} = \frac{1}{N_u}\sum_{i=1}^{N_u}\frac{1}{N_i^{can}}\sum_{j=1}^{N_i^{can}}(L_{i,j}^{cls} + L_{i,j}^{reg}). \tag{16}$$

Methods	1%	5%	10%	100%
Supervised	10.00±0.26	20.92±0.15	26.94±0.11	40.90
STAC	13.97±0.35	24.38±0.12	28.64±0.21	39.20
ISMT	18.88±0.74	26.37±0.24	30.53±0.52	39.64
InsantTeaching	18.05±0.15	26.75±0.05	30.40±0.05	40.20
UbTeacher	20.75±0.12	28.27±0.11	31.50±0.10	41.30
HumbleTeacher	16.96±0.38	27.70±0.15	31.61±0.28	42.37
SoftTeacher	20.46±0.39	30.74±0.08	34.04±0.14	44.50
MUM	21.88±0.12	28.52±0.09	31.87±0.30	42.11
DCST	23.02±0.23	32.10±0.15	35.20±0.20	44.60
Rethinking	19.02±0.25	28.40±0.15	32.23±0.14	43.30
MA-GCP	21.30±0.28	31.67±0.16	35.02±0.26	45.92
SED	-	29.01	34.02	41.50
PseCo	22.43±0.36	32.50±0.08	36.06±0.24	46.10
DSL	22.03±0.28	30.87±0.24	36.22±0.18	43.80
Ours	**24.04**±0.69	**34.58**±0.23	**37.50**±0.14	**46.20**

Table 1: Comparison with other SOTA methods on MS-COCO dataset.

Methods	Model	mAP
Supervised (Li et al. 2022b)	FRCNN	76.30
STAC (Sohn et al. 2020)	FRCNN	77.45
ISMT (Yang et al. 2021)	FRCNN	77.23
InstantTeaching (Zhou et al. 2021)	FRCNN	79.20
HumbleTeacher (Tang et al. 2021b)	FRCNN	80.94
UbTeacher (Liu et al. 2021)	FRCNN	77.37
MUM (Kim et al. 2022)	FRCNN	78.94
Rethiking (Li et al. 2022b)	FRCNN	79.00
SED (Guo et al. 2022)	FRCNN	80.60
MA-GCP (Li, Yuan, and Li 2022)	FRCNN	81.72
DSL (Chen et al. 2022b)	FCOS	80.70
Ours	FRCNN	**84.70**

Table 2: Comparison with other methods on PASCAL VOC.

Training Protocols. The overall training loss is formulated as below:

$$L = L_{sup} + \alpha L_{con} + \beta L_{rep} \tag{17}$$

where α and β are constants. The SGD algorithm is chosen for minimizing the above training loss, where the initial learning rate is set to 0.015, the weight decay is set to 0.0001, and the momentum is set to 0.9.

Experiments

Experimental Setup

Dataset and Evaluation Protocol. We evaluate the SSOD methods on the MS-COCO (Lin et al. 2014) and PASCAL VOC (Everingham et al. 2010) datasets.

• **MS-COCO dataset.** We use *train2017* subset which contains 118k labeled images and *unlabeled2017* subset which contains 123k unlabeled images to train the network. We test the performance of SSOD methods on *val2017* subset which consists of 5k images. The standard mean average precision (mAP) is adopted as the evaluation metric. Following (Sohn et al. 2020), two settings are used for testing the performance: (1) We randomly sample 1%, 5%, or 10% images of *train2017* as labeled data, and the remained images of *train2017* are regarded as unlabeled data. For each ratio of labeled data, there are 5 random sampling data splits. We report the average performance of 5 runs. (2) All images of *train2017* are regarded as labeled data, and images of *unlabeled2017* are regarded as unlabeled data. This setting aims to validate how the large-scale unlabeled data benefits to fully supervised models.

• **PASCAL VOC dataset.** The *trainval* set of VOC07 which contains 5k images is used as labeled data, and the *trainval* set of VOC12 which contains 11k images is used as unlabeled data. We report average precision (AP) with IoU threshold 0.5 on the test set of VOC07.

Implementation Details

For fair comparison, we use Faster RCNN (Ren et al. 2015) with FPN (Lin et al. 2017a) as detection model and ImageNet pre-trained RestNet50 (He et al. 2016) as backbone. Following SoftTeacher (Xu et al. 2021), we implement our method on MMDetection (Chen et al. 2019). For the partially labeled data setting of MS-COCO, the model is trained for 180k iterations on 4 GPUs with 10 images per GPU. The ratio of labeled images against unlabeled images per GPU is $1:4$. Learning rate is divided by 10 at 120k and 160k iterations, and the loss weight α is set to 4. In the fully labeled data setting of MS-COCO, the model is trained for 720k iterations with 16 images per GPU, and the ratio of labeled data to unlabeled data is $1:1$. Learning rate is divided by 10 at 480k and 640k iteration, and α is set to 2. Score threshold for testing Faster-RCNN head is set to 0.001. For the PASCAL VOC dataset, the model is trained for 60k iterations on 8 GPUs with 5 images per GPU. The ratio of labeled data to unlabeled data is $1:4$. The learning rate is initialized as 0.01, divided by 10 at 40k iteration and 50k iteration, and α is set to 4. In adaptive pseudo labeling, τ_0 is set to 0.9. In box refinement, $M = 10$ and $r = 0.05$. λ and γ are 0.999 for EMA. The loss weight β is set to 1 for the partially labeled data setting of MS-COCO and PASCAL VOC, while $\beta = 0.5$ for the fully labeled data setting of MS-COCO.

Comparison with Other Methods

We compare the proposed method with supervised baseline and state-of-the-art SSOD methods, including STAC (2020), Unbiased Teacher (2021), SoftTeacher (2021), DSL (2022b), etc. The experimental results on the MS-COCO dataset are presented in Table 1. Our proposed approach outperforms the supervised baseline in all partially labeled data settings by up to 10 mAP. Specifically, we improve the supervised baseline by +14.04 mAP and +13.66 mAP, respectively, yielding 24.04 mAP and 34.58 mAP in the rarely labeled data circumstances, i.e., 1% and 5% labeling ratio. This demonstrates the superiority of the proposed semi-supervised learning regime. Furthermore, the proposed approach significantly outperforms other approaches by a large margin. For example, our method surpasses the recently proposed methods, i.e. DCST (2022a), PseCo (2022a) and DSL (2022b), by +1.02 mAP, +2.08 mAP and +1.28 mAP under 1%, 5% and 10% ratio, respectively. In a nutshell, we provide new cutting-edge results for the task of semi-supervised

FCL	PCL	APL	sharpen	mAP	AP_{50}	AP_{75}
				27.2	44.9	28.8
✓				35.1	54.6	38.4
✓	✓			35.8	55.9	39.0
✓	✓	✓		36.9	56.3	40.2
✓	✓	✓	✓	**37.2**	**56.9**	**40.6**

Table 3: Ablation studies on key components. FCL and PCL indicates feature-level and prediction-level consistency learning, respectively. 'sharpen' represents the probability sharpening operation in PCL.

object detection under limited labeled data. In addition, our method remains robust under fully supervised conditions. Table 2 presents the results on the PASCAL VOC dataset. Our method achieves 84.7 mAP, which is superior to existing approaches as well.

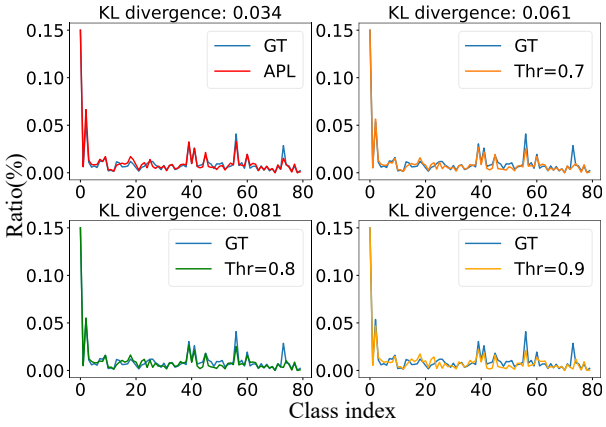


Figure 6: Class distributions of pseudo and GT labels.

Ablation Study

In this section, we validate the key components of our method. All experiments are conducted on 10% labeled data setting unless otherwise specified.

Effects of key components. The effectiveness of critical components is shown in Tab. 3. The first row in the table represents the supervised baseline, and it achieves 27.2 mAP with 10% labeled data. Applying a feature-level consistency learning strategy improves performance and reaches 35.1 mAP. We further add a consistency module to encourage prediction-level consistency learning. The performance attains +8.6 mAP gains compared to the baseline. We adopt APL to filter high-quality pseudo labels, and the performance improved by +1.1 mAP. It is worth noting that APL brings +1.2 mAP on strict evaluation metrics, i.e., AP_{75}, from 39.0 to 40.2, which demonstrates the superiority of accurate pseudo labels, especially for box regression. Finally, a simple label sharpening operation is introduced to achieve +0.3 mAP gains, contributing to a faster convergence speed.

Quality of pseudo labels. In this part, we examine the effectiveness of the pseudo labels. The F1 scores of APL

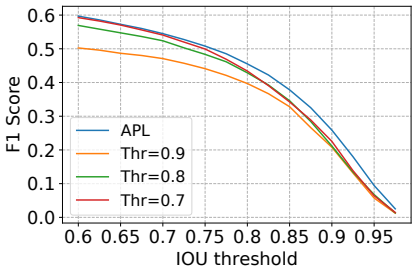
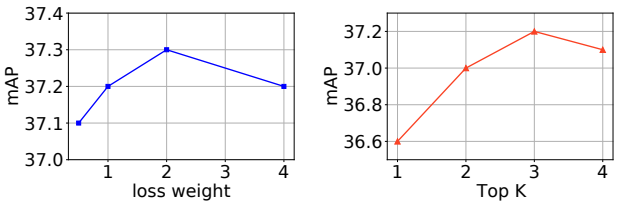


Figure 7: F1 scores of pseudo labels produced by models which are learned with the pseudo labeling strategy based on a fixed threshold (0.7, 0.8, or 0.9) and our devised APL.

and fixed threshold filtering strategy with various thresholds are presented in Fig. 7. One can see that for different IOU thresholds, the F1 score of APL is consistently higher than the baseline method in all IOU thresholds. In particular, when IOU ≥ 0.7, APL still outperforms the baseline method, which validates the high-quality localization of our method. Fig. 6 shows the pseudo label distribution in all classes. The pseudo labels filtered by APL have the closest class distribution to ground truth labels and remain a minimum KL divergence. These results indicate that the pseudo labels are accurate both in localization and class distribution.

Ablation study on hyper-parameters. We conduct ablation studies on the parameter K of Top_K in APL and the feature-level consistency learning loss weight β. The results are illustrated in Fig. 8(b) and Fig. 8(a). For the selection of K, the best choice is $K = 3$. $K = 1$ means only the probability of most confident class is considered, which will impede the threshold updating of underrepresented classes and interferes the performance. For the feature-level consistency loss weight, $\beta = 2.0$ achieves the best result.



(a) Ablation studies on β (b) Ablation studies on Top_K.

Figure 8: Ablation studies on loss weight β and Top_K.

Conclusion

In this paper, we propose a novel method to tackle the SSOD task, based on adaptive pseudo labeling and consistency learning. Our devised adaptive labeling strategy is capable of exploring more samples for hard classes while preserving the quality of easy classes. The proposed intra-scale and inter-scale consistency learning algorithm can facilitate the utilization of pseudo labels and improve the detection of small objects. Extensive experiments indicate that our method achieves state-of-the-art performance on SSOD.

Acknowledgments

This work was supported in part by the Shenzhen Science and Technology Program (NO. JCYJ20220530141211024), in part by the Guangdong Basic and Applied Basic Research Foundation (NO. 2020B1515020048), in part by the National Natural Science Foundation of China (NO. 61976250, NO. U1811463, NO. 62106235 and No. 62003256), in part by the Fundamental Research Funds for the Central Universities under Grant 22lgqb25. This work was also supported by Open Research Projects of Zhejiang Lab (NO. 2019KD0AD01/017), the Exploratory Research Project of Zhejiang Lab(NO. 2022PG0AN01) and MindSpore which is a new deep learning computing framework[1].

References

Cai, Z.; and Vasconcelos, N. 2019. Cascade R-CNN: high quality object detection and instance segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 43(5): 1483–1498.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chen, B.; Chen, W.; Yang, S.; Xuan, Y.; Song, J.; Xie, D.; Pu, S.; Song, M.; and Zhuang, Y. 2022a. Label Matching Semi-Supervised Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14381–14390.

Chen, B.; Li, P.; Chen, X.; Wang, B.; Zhang, L.; and Hua, X.-S. 2022b. Dense Learning based Semi-Supervised Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4815–4824.

Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; Zhang, Z.; Cheng, D.; Zhu, C.; Cheng, T.; Zhao, Q.; Li, B.; Lu, X.; Zhu, R.; Wu, Y.; Dai, J.; Wang, J.; Shi, J.; Ouyang, W.; Loy, C. C.; and Lin, D. 2019. MMDetection: Open MMLab Detection Toolbox and Benchmark. arXiv:1906.07155.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15750–15758.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338.

Fang, Y.; Yang, S.; Wang, S.; Ge, Y.; Shan, Y.; and Wang, X. 2022. Unleashing Vanilla Vision Transformer with Masked Image Modeling for Object Detection. *arXiv preprint arXiv:2204.02964*.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 580–587.

Guo, Q.; Mu, Y.; Chen, J.; Wang, T.; Yu, Y.; and Luo, P. 2022. Scale-Equivalent Distillation for Semi-Supervised Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14522–14531.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

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

Jeong, J.; Lee, S.; Kim, J.; and Kwak, N. 2019. Consistency-based semi-supervised learning for object detection. *Advances in neural information processing systems*, 32.

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

Kim, J.; Jang, J.; Seo, S.; Jeong, J.; Na, J.; and 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 (CVPR)*, 14512–14521.

Li, A.; Yuan, P.; and Li, Z. 2022. Semi-Supervised Object Detection via Multi-Instance Alignment With Global Class Prototypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9809–9818.

Li, G.; Li, X.; Wang, Y.; Zhang, S.; Wu, Y.; and Liang, D. 2022a. PseCo: Pseudo Labeling and Consistency Training for Semi-Supervised Object Detection. arXiv:2203.16317.

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

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

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

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

Liu, Y.-C.; Ma, C.-Y.; He, Z.; Kuo, C.-W.; Chen, K.; Zhang, P.; Wu, B.; Kira, Z.; and Vajda, P. 2021. Unbiased Teacher for Semi-Supervised Object Detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In

[1]https://www.mindspore.cn/

Proceedings of the IEEE conference on computer vision and pattern recognition, 779–788.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Sohn, K.; Zhang, Z.; Li, C.-L.; Zhang, H.; Lee, C.-Y.; and Pfister, T. 2020. A Simple Semi-Supervised Learning Framework for Object Detection. arXiv:2005.04757.

Tang, P.; Ramaiah, C.; Wang, Y.; Xu, R.; and Xiong, C. 2021a. Proposal learning for semi-supervised object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2291–2301.

Tang, Y.; Chen, W.; Luo, Y.; and Zhang, Y. 2021b. Humble Teachers Teach Better Students for Semi-Supervised Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3132–3141.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 1195–1204. Curran Associates, Inc.

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

Wang, K.; Nie, Y.; Fang, C.; Han, C.; Wu, X.; Wang, X.; Lin, L.; Zhou, F.; and Li, G. 2022a. Double-Check Soft Teacher for Semi-Supervised Object Detection. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Wang, Y.; Chen, H.; Heng, Q.; Hou, W.; Savvides, M.; Shinozaki, T.; Raj, B.; Wu, Z.; and Wang, J. 2022b. FreeMatch: Self-adaptive Thresholding for Semi-supervised Learning. arXiv:2205.07246.

Xu, M.; Zhang, Z.; Hu, H.; Wang, J.; Wang, L.; Wei, F.; Bai, X.; and Liu, Z. 2021. End-to-End Semi-Supervised Object Detection with Soft Teacher. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

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

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

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

Zhou, Q.; Yu, C.; Wang, Z.; Qian, Q.; and 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*, 4081–4090.