

---

# Rethinking Pseudo Labels for Semi-Supervised Object Detection

---

Hengduo Li<sup>1</sup>, Zuxuan Wu<sup>2</sup>, Abhinav Shrivastava<sup>1</sup>, Larry S. Davis<sup>1</sup>

<sup>1</sup> University of Maryland, <sup>2</sup> Fudan University

{hdli, abhinav, lsd}@cs.umd.edu zzwu@fudan.edu.cn

## Abstract

Recent advances in semi-supervised object detection (SSOD) are largely driven by consistency-based pseudo-labeling methods for image classification tasks, producing pseudo labels as supervisory signals. However, when using pseudo labels, there is a lack of consideration in localization precision and amplified class imbalance, both of which are critical for detection tasks. In this paper, we introduce certainty-aware pseudo labels tailored for object detection, which can effectively estimate the classification and localization quality of derived pseudo labels. This is achieved by converting conventional localization as a classification task followed by refinement. Conditioned on classification and localization quality scores, we dynamically adjust the thresholds used to generate pseudo labels and reweight loss functions for each category to alleviate the class imbalance problem. Extensive experiments demonstrate that our method improves state-of-the-art SSOD performance by 1-2% and 4-6% AP on COCO and PASCAL VOC, respectively. In the limited-annotation regime, our approach improves supervised baselines by up to 10% AP using only 1–10% labeled data from COCO.

## 1 Introduction

The astounding performance of deep neural networks on various computer vision tasks can be largely attributed to the availability of large-scale datasets that are manually labeled. However, collecting human annotations is labor-intensive and time-consuming, particularly for visual understanding tasks, like object detection [26, 20] and semantic segmentation [11, 5]. To remedy this, there is an ever-growing interest in semi-supervised learning (SSL), which learns feature representations with limited supervision by exploring the massive amount of unlabeled images that are readily available. While extensive studies have been conducted on SSL for image classification tasks [2, 43, 34, 2, 44, 21, 29, 1], relatively limited effort has been made to address object detection, for which annotations are more expensive to obtain.

Most recent semi-supervised object detection (SSOD) approaches [35, 28, 17, 49] are direct extensions of SSL methods designed for image classification using a teacher-student training paradigm [36, 34, 2]. In particular, the teacher model is first trained in a supervised manner with a limited number of labeled samples. Then, given an unlabeled image, the teacher model produces *pseudo* bounding boxes together with their corresponding class predictions, which are further used as ground-truth labels for the student model. To ensure effective distillation, the teacher and the student models typically operate on two augmented views of the same image [34, 35, 28, 49].

The use of a teacher-student model at its core aims to produce reliable pseudo labels in lieu of human annotations. While effective, we argue that pseudo labels, in the form of bounding boxes associated with class predictions, are sub-optimal for SSOD. The reasons are twofold: (1) In image classification, prediction scores naturally represent the likelihood of an object appearing in an image, and thus setting a threshold to select highly confident predictions is reasonable. However, as detection

requires localizing and classifying objects using two separate branches through regression and classification, the resulting classification scores of pseudo boxes are *unaware* of the localization quality. Therefore, while widely adopted, filtering out boxes based on class predictions on top of non-maximum suppression is not appropriate; (2) Pseudo labels produced by the teacher model amplifies class imbalance which results from the long-tailed nature in detection tasks. For example, there are only 9 toasters but 12,343 persons in 5% of the COCO [26] training set even though they are both *common*<sup>1</sup> classes! As a result, lower-confidence predictions from underrepresented classes are oftentimes filtered out with a threshold that works well for top-performing classes.

To mitigate these issues, we propose certainty-aware pseudo labels together with dynamic thresholding and reweighting mechanisms tailored for SSOD. In particular, the certainty-aware pseudo labels are designed to reflect localization quality and classification confidence at the same time. Conditioned on these certainty measurements, we dynamically adjust the thresholds used to produce pseudo labels and reweight loss functions on a per-category basis to combat class imbalance. While conceptually appealing, it is challenging to have an in-vitro metric in mainstream detection frameworks that reflects localization quality to complement classification accuracy due to the design that performs localization with regression.

Motivated by a few recent studies that replace regression with classification for better localization [23, 31, 40], we formulate localization as a classification problem to obtain an estimation of localization quality. More specifically, for each side of a candidate box, we introduce a line segment that is perpendicular to it. The line is split into consecutive intervals, each of which is associated with a prediction score through classification, indicating the probability of the side intersects with the interval. We then average the maximal classification scores from all four sides of a candidate box as its localization quality metric. To ensure accurate localization, we further refine locations within intervals. The pseudo labels are now certainty-aware, measuring both localization precision and classification confidence, and can be readily used to generate better labels. In particular, for each category, conditioned on the localization and classification confidence, we dynamically determine a threshold to generate pseudo labels and reweight loss functions such that underrepresented classes are emphasized during training to mitigate class imbalance.

We conduct extensive experiments on COCO [26] and PASCAL VOC [11] under common semi-supervised settings, and demonstrate that our method improves SOTA performance by 1-2% and 4-6% AP on COCO and PASCAL VOC respectively using various training recipes, and improves the supervised baseline by up to 10% AP when using only 1/2/5/10 % annotations of COCO. We further show that our method is complementary to existing approaches resorting to orthogonal techniques like co-teaching [14] and model ensemble. Extensive ablation experiments are conducted to validate the effectiveness of different components of our method, and demonstrate that our approach is relatively robust to hyper-parameter selections.

## 2 Approach

Our goal is to address semi-supervised object detection where a set of labeled images with box-level annotations and a set of unlabeled images are used for training. Built upon consistency-based pseudo labeling (Sec. 2.1), our method produces certainty-aware pseudo labels for both classification and localization. This is achieved by formulating box localization as a classification problem and injecting localization confidence to guide pseudo label generation (Sec. 2.2). Conditioned on classification and localization certainty, we dynamically adjust the thresholds to generate pseudo labels and re-weight the loss function for different classes (Sec. 2.3). An overview of our method is shown in Fig. 1.

### 2.1 Preliminary

Our approach is built upon consistency-based pseudo labeling, which has proven effective for both semi-supervised image classification and object detection. Below, we briefly introduce the teacher-student training paradigm which serves as the basis for current consistency-based approaches. Overall, a teacher model is firstly trained on labeled images, and then it is used to produce pseudo labels (boxes) on unlabeled images to supervise the training of a student model.

---

<sup>1</sup>COCO: Common Objects in Context.

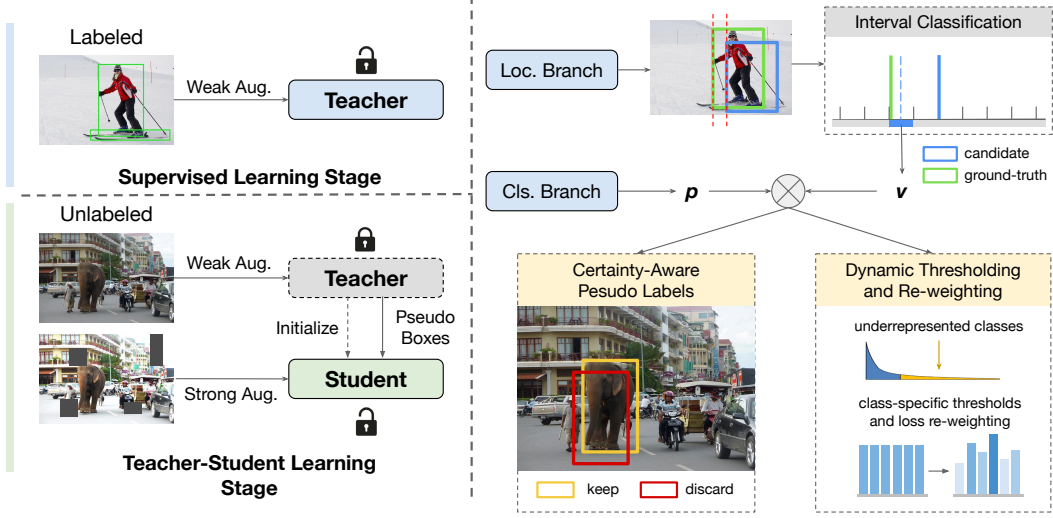


Figure 1: **A conceptual overview of our approach.** **Left:** We first train the teacher model on labeled images to generate pseudo labels (boxes) on unlabeled images. The student model is then trained with pseudo labels. **Right:** We propose to generate certainty-aware pseudo labels conditioned on both classification and localization confidence scores, for improved localization, by formulating localization as a classification problem. The scores are then used to derive dynamic thresholds and re-weight losses in a class-wise manner to mitigate class imbalance.

Formally, given a set of labeled images  $\mathcal{S}$  and a set of unlabeled images  $\mathcal{U}$ , an object detector is trained on  $\mathcal{S}$  in a standard supervised manner:

$$\begin{aligned}\mathcal{L}_s(\mathcal{I}, \mathbf{p}, \mathbf{t}, \mathbf{p}_*, \mathbf{t}_*) &= \mathbb{E}_{\mathcal{I} \sim \mathcal{S}} \mathbb{E}_{i \sim \mathcal{B}} \mathcal{L}_s(\mathcal{I}, p^i, t^i) \\ &= \mathbb{E}_{\mathcal{I} \sim \mathcal{S}} \mathbb{E}_{i \sim \mathcal{B}} [\ell_{cls}(p^i, p_*) + \ell_{loc}(t^i, t_*)]\end{aligned}\quad (1)$$

where  $\mathcal{I}$  is an input image with a set of candidate boxes  $\mathcal{B}$ , and  $p^i, t^i$  denote the prediction of class probability and bounding box coordinates for the  $i$ -th candidate box. Each candidate box is associated with a one-hot label  $p_*^i$  and a ground-truth box location  $t_*^i$  as supervisory signals, and the losses for classification and localization are often instantiated as a weighted sum of a standard cross-entropy loss and a smooth  $L_1$ .

The teacher model trained on  $\mathcal{S}$  then generates pseudo boxes on all unlabeled images in  $\mathcal{U}$  through standard inference. These pseudo boxes are further filtered by a predefined threshold  $\tau$  conditioned on the prediction confidence  $p^i$ ; the remaining boxes are used to train a student model whose weights are initialized from the teacher model:

$$\mathcal{L} = \mathcal{L}_s(\mathcal{I}_s, \mathbf{p}_*, \mathbf{t}_*) + \lambda_u \mathcal{L}_u(\mathcal{I}_u, \mathbf{p}_{u*}, \mathbf{t}_{u*}) \quad (2)$$

where  $\mathbf{p}_{u*}$  and  $\mathbf{t}_{u*}$  denote pseudo class labels and box coordinates derived from the teacher model. The loss is a weighted sum of supervised loss  $\mathcal{L}_s$  on labeled images and unsupervised loss  $\mathcal{L}_u$  on unlabeled samples controlled by  $\lambda$ . Following [35, 28, 49, 45], given an unlabeled image, when generating pseudo labels, we only use horizontal flipping as a weak augmentation; when training the student model, we use strong augmentations including color jitter, Gaussian blur and Cutout [9] for the same image.

## 2.2 Certainty-aware Pseudo Labels

Recall that existing approaches typically form bounding boxes through coordinate regression, and then predict the objects within boxes through classification. To generate pseudo boxes used as ground-truth by the student model, it is a common practice to apply a threshold  $\tau$  to filter out boxes with low classification scores. While straightforward, such a localization-agnostic strategy fails to model how well boxes are localized. To address this issue, we formulate localization as classification, producing certainty-aware boxes, such that the quality of both localization and classification are explicitly considered to guide the generation of pseudo labels.

Formally, given an *unlocalized* candidate box  $(x_1, y_1, x_2, y_2)$  with its top left corner at  $(x_1, y_1)$  and its bottom right corner at  $(x_2, y_2)$ , its corresponding ground-truth locations are denoted as  $(x_{g1}, y_{g1}, x_{g2}, y_{g2})$ . Each side of the candidate is independently localized to the corresponding side of ground-truth through *classification*. Taking the left side of the candidate box as an example, we first obtain a line segment  $l$  which is perpendicular to the side, then split  $l$  evenly into  $K$  consecutive intervals and predict which interval the *unlocalized* side belongs to according to the ground-truth position  $x_{g1}$  through a  $K$ -way classification. In particular, if the left side of the ground-truth box is perpendicular to the  $k$ -th interval, we mark that the side belongs to the  $k$ -th interval for training (see Figure 1 for an illustration). Then the loss function for localization given an image can be written as:

$$\ell_{\text{seg}}(\mathbf{T}, \mathbf{Y}) = \mathbb{E}_{i \sim \mathcal{B}} \sum_{s=1}^{s=4} \sum_{k=1}^{k=K} -y_{s,k}^i \log(\text{sigmoid}(t_{s,k}^i)) \quad (3)$$

where the superscript  $i$  denotes the  $i$ -th candidate box sampled from the box set, and  $t_{k,s}^i$  is the unnormalized prediction score from the  $k$ -th interval in the  $s$ -th side, and the label  $y_{k,s}^i = 1$  if the side belongs to the  $k$ -th interval otherwise  $y_{k,s}^i = 0$ . To measure the localization quality for the  $i$ -th box, we first obtain the maximal class score along each side and then compute the mean of these scores:

$$v^i = \frac{1}{4} \sum_{s=1}^{s=4} \max_{1 \leq k \leq K} (t_{s,k}^i). \quad (4)$$

The localization quality score  $v^i$ , indicating how well boxes are localized, together with the classification confidence  $p^i$  are two complementary metrics measuring the certainty of localization and classification, respectively. They are further used for post-processing like non-maximum suppression and pseudo label generation, which will be described below.

Thus far we have formulated box localization in a classification manner to obtain quality measurement, yet the localization performance could be largely hindered by discretizing the problem of deriving continuous bounding box coordinates. In particular, the membership of an interval is a rough estimation of location particularly when the interval size is large. To obtain the precise location of a side within the interval, we further perform regression from the center line  $x_k$  of the  $k$ -th interval to the ground-truth line  $x_g$  for finer localization. We use a smooth  $L_1$  loss for fine regression, and the overall localization loss becomes:

$$\begin{aligned} \ell_{\text{loc}} &= \ell_{\text{seg}}(\mathbf{T}, \mathbf{Y}) + \ell_{\text{reg}}(\mathbf{Y}, \mathbf{X}_g) \\ &= \mathbb{E}_{i \sim \mathcal{B}} \sum_k^K y_k [-\log(\text{sigmoid}(t_{k,s}^i)) + \text{SmoothL1}(x_k, x_g)] \end{aligned} \quad (5)$$

Finally, we replace the localization loss in Eqn. 1 used by both the teacher model and the student model with Eqn. 5. Consequently, the trained teacher model produces pseudo labels that are aware of both localization and classification quality.

### 2.3 Dynamic Thresholding and Re-weighting

As discussed above, class imbalance exists in object detection especially when annotations are scarce. The imbalance is further enlarged in semi-supervised settings since the teacher model produces relatively lower confidence scores for underrepresented classes [8], which hardly survive the often large threshold  $\tau$ . On the other hand, simply lowering  $\tau$  introduces more noisy pseudo labels in common classes. With this in mind, we propose to dynamically adjust the threshold and re-weight losses in a class-wise manner conditioned on classification and localization confidence scores for each category.

For each category  $m$ , the classification and localization confidence score  $p_m^j$  and  $v_m^j$  for each foreground candidate box (indexed by  $j$ ) are accumulated online to produce an unnormalized frequency score  $c = \sum_j p_m^j v_m^j$ , which not only approximates the detector’s current overall confidence level for the category but also counts the number of foreground instances. The class-specific threshold  $\tau_m$  and re-weighting coefficient  $\alpha_m$  are then derived as follows:

$$\tau_m = \left( \frac{\sum_j p_m^j v_m^j}{\mathbb{E}_{m \sim M} \sum_j \mathbb{1}} \right)^{\gamma_1} \tau, \quad \alpha_m = \left( \frac{\mathbb{E}_{m \sim M} \sum_j \mathbb{1}}{\sum_j p_m^j v_m^j} \right)^{\gamma_2} \quad (6)$$

where  $\mathbb{E}_{m \sim M} \sum_j \mathbb{1}$  denotes the average number of foreground instances from all categories and  $\tau$  is the original manually chosen threshold. The class-specific  $\tau_m$  is then applied to filter pseudo labels, and  $\alpha_m$  is multiplied to losses (Eqn. 2) of all foreground instances in each category. Two factors  $\gamma_1$  and  $\gamma_2$  control the degree of focus on underrepresented classes; when set to 0, dynamic thresholding and re-weighting are disabled.

By keeping more pseudo labels for underrepresented classes, as well as promoting their importance during training through re-weighting the losses, the bias towards head classes is mitigated. It is worth pointing out that  $\tau_m$  needs to be bounded as it is applied on predicted probabilities, and we find clipping it into  $[0.4, 0.9]$  works well empirically.

### 3 Experiments

#### 3.1 Experimental Setup

**Datasets.** We evaluate our method on two standard object detection datasets, COCO [26] and PASCAL VOC [11], under semi-supervised settings following [17, 35, 28, 49, 45]. In particular, three settings are used: (1) *COCO-full*: the COCO train2017 set containing  $\sim 118k$  images is used as the labeled set, and the additional  $\sim 123k$  unlabeled images are used as unlabeled set; (2) *COCO-partial*: we follow [35, 28, 49] and randomly sample 1%/2%/5%/10% images from COCO train2017 set as the labeled set, and use the remaining images in train2017 as the unlabeled set; (3) *PASCAL VOC*: the VOC07 trainval set containing  $\sim 5k$  images is used as labeled set and the VOC12 trainval is used as unlabeled set. For evaluation, the val2017 set of COCO and the VOC07 test set of PASCAL VOC are used.

**Training and Testing Configuration.** Since existing methods for SSOD use various different setups for training and testing, we evaluate our method under multiple settings for fair comparison. In all settings, the teacher model is firstly trained on the labeled set, and the student model is trained on the combination of labeled and unlabeled images. We report mean Average Precision (mAP) at different IoU thresholds (e.g. AP<sub>50</sub>, AP<sub>75</sub> and AP<sub>50:95</sub> which is denoted as AP) to measure the effectiveness.

The training details are summarized as below: (1) *COCO-full*: we report results under  $1\times$  and  $3\times$  training schedules, which are roughly equivalent to 12 and 36 epochs respectively. The teacher models are trained for 180k/540k iterations, and the student models are trained using the same schedule. (2) *COCO-partial*: for 1%/2%/5%/10% settings, we train teacher models for 6k/12k/30k/60k iterations, and then train student models for 180k iterations. During testing, we report results under two score thresholds, 0.05 and 0.001, that are applied on final detection predictions. A lower threshold generally improves the recall through keeping more predicted boxes and thus results in slightly better performance. (3) *PASCAL VOC*: we train teacher models for 10k iterations, then train student models for 90k iterations. Both single-scale training and multi-scale training results are reported. For ablation study and analysis, we use the 2% COCO setting and a shorter  $0.5\times$  schedule due to the limited computational resources, if not mentioned otherwise.

**Implementation Details.** Most of our implementation details follow existing approaches for fair comparison. We use Faster-RCNN [33] with FPN [24] as our detector using a ResNet-50 [15] as its backbone network. The backbone is initialized from ImageNet pre-trained weights. We set  $\lambda_u = 1.0$  and  $\tau = 0.7$ . For the localization branch, we set  $K = 30$ . For dynamic thresholding and re-weighting, we set  $\gamma_1 = 0.05$  and  $\gamma_2 = 0.6$ .

We train the models with 4 Nvidia 1080 Ti GPUs, using a total batch size of 8. We use SGD with an initial learning rate of 0.01, a weight decay of  $1e-4$ , a momentum of 0.9. Learning rate is divided by 10 at the 120k/160k iteration for the 180k schedule, and likewise for other schedules. For single-scale training, the short side of image is resized to 600 for PASCAL VOC and 800 for COCO; for multi-scale training, the short side size is sampled from (640, 800). The long side is kept no more than 1,333 after resizing. Other details are the same as in Detectron2 [42], which is used for our implementation.

#### 3.2 Main Results

We first report the results on three settings, i.e. *COCO-full*, *COCO-partial* and *PASCAL VOC* and compare with supervised baselines as well as various state-of-the-art approaches for semi-supervised



Table 1: **Comparison with state-of-the-art approaches on COCO-full and PASCAL VOC.**

\* denotes the use of longer training schedule ( $3\times$ ).  $\S$  denotes multi-scale training.

(a) COCO-full		(b) PASCAL VOC.		
Method	AP	Method	AP <sub>50</sub>	AP <sub>75</sub> AP
Supervised	37.9	Supervised	76.3	47.5 45.3
CSD* [17]	38.8	CSD [17]	74.7	- -
STAC* [35]	39.2	STAC [35]	77.4	- 44.6
ISMT [45]	39.6	ISMT [45]	77.2	- 46.2
Instant-Teaching [49]	39.6	Instant-Teaching [49]	78.3	52.0 48.7
Multi Phase* [41]	40.1	Multi Phase [41]	77.4	- -
Unbiased Teacher* [28]	41.3	Unbiased Teacher $\S$ [28]	77.4	- 48.7
Ours	41.0	Ours	76.9	57.9 52.4
Ours*	<b>43.3</b>	Ours $\S$	<b>79.0</b>	<b>59.4</b> <b>54.6</b>

Table 2: **Results (AP) on COCO-partial.**  $\dagger$  denotes using a lower final score threshold to improve recall as in [49].  $\ddagger$  denotes using model ensemble.

Methods	1% COCO	2% COCO	5% COCO	10% COCO
Supervised	9.05 $\pm$ 0.16	12.70 $\pm$ 0.15	18.47 $\pm$ 0.22	23.86 $\pm$ 0.81
CSD [17]	10.20 $\pm$ 0.15 (+1.15)	13.60 $\pm$ 0.10 (+0.90)	18.90 $\pm$ 0.10 (+0.43)	24.50 $\pm$ 0.15 (+0.64)
STAC [35]	13.97 $\pm$ 0.35 (+4.92)	18.25 $\pm$ 0.25 (+5.55)	24.38 $\pm$ 0.12 (+5.91)	28.64 $\pm$ 0.21 (+4.78)
Unbiased Teacher [28]	17.84 $\pm$ 0.12 (+8.79)	21.98 $\pm$ 0.07 (+9.28)	26.30 $\pm$ 0.11 (+7.83)	29.64 $\pm$ 0.10 (+5.78)
Instant-Teaching $\dagger$ [49]	16.00 $\pm$ 0.20 (+6.95)	20.70 $\pm$ 0.30 (+8.00)	25.50 $\pm$ 0.05 (+7.03)	29.45 $\pm$ 0.15 (+5.59)
Instant-Teaching $\ddagger$ [49]	18.05 $\pm$ 0.15 (+9.00)	22.45 $\pm$ 0.15 (+9.75)	26.75 $\pm$ 0.05 (+8.28)	30.40 $\pm$ 0.05 (+6.54)
Ours	<b>18.21</b> $\pm$ 0.31 (+9.16)	<b>22.62</b> $\pm$ 0.24 (+9.92)	<b>27.78</b> $\pm$ 0.17 (+9.31)	<b>31.67</b> $\pm$ 0.18 (+7.81)
Ours $\dagger$	<b>19.02</b> $\pm$ 0.25 (+9.97)	<b>23.34</b> $\pm$ 0.18 (+10.64)	<b>28.40</b> $\pm$ 0.15 (+9.93)	<b>32.23</b> $\pm$ 0.14 (+8.37)

object detection, such as CSD [17], STAC [35], ISMT [45], Instant-Teaching [49], Multi-Phase Learning [41] and Unbiased Teacher [28]. For approaches using ensemble techniques like [49, 41], we report their single-model results for fair comparison. For Unbiased Teacher [28] which uses larger batch size and longer training schedules, we retrain it under our training schedules with their official implementation. Results are summarized in Table 1 and Table 2.

**COCO-full and PASCAL VOC.** As shown in Table 1(a-b), our approach obtains 1 – 2% and 4 – 6% AP improvement over state-of-the-art results on COCO and PASCAL VOC. For example, when trained under  $3\times$  schedule, our method obtains 43.3% AP and outperforms Unbiased Teacher [28] by 2.0%. In the short schedule ( $1\times$ ) setting, our approach obtains 41.0% AP, which outperforms methods using long schedules like CSD [17] and STAC [35]. On PASCAL VOC, we obtain 52.4% AP and 54.6% AP with single-scale training and multi-scale training respectively. Notably, large improvements are obtained by our method when precise localization is needed (e.g. AP<sub>75</sub>), indicating that our approach improves the localization quality for semi-supervised object detection.

**COCO-partial.** We then evaluate our method under the limited-annotation regime on COCO-partial. As demonstrated in Table 2, our method improves supervised baselines by up to 10%. When 10% annotations are available, our method achieves 32.23% AP and is  $\sim 2\%$  higher than Instant-Teaching [49] even though model ensemble is used in their method. With only 1%/2%/5% annotations available, our method is able to achieve state-of-the-art 19.02%, 23.34% and 28.40% APs respectively.

**Compatibility to other methods.** It is worth pointing out that our method is orthogonal to many useful techniques explored in existing approaches mentioned above. For example, when using a simplified model ensemble method from [49], a further performance improvement is observed as in Table 3. In particular, we train two teacher models separately and use the ensemble of them to generate pseudo labels, which are then used to train two student models. Finally, the ensemble of two student models is evaluated. As can be seen, AP<sub>50</sub> and overall AP are improved by 0.8% and 0.5% respectively. Other methods like Mean Teacher [36] and Co-teaching [14] have also been utilized in [28, 49, 45] but not in our method, for which we believe our work could be complementary to many current state-of-the-art methods for semi-supervised object detection.

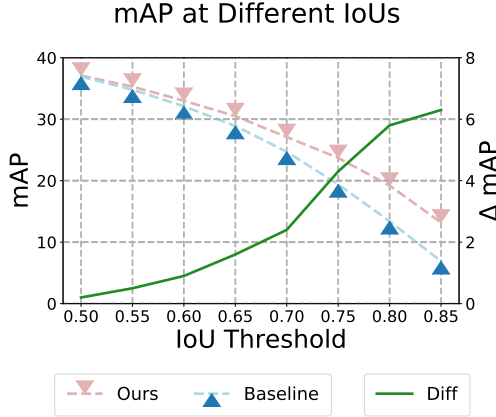


Figure 2: **Performance at different IoU criteria** under 2% COCO setting. Larger improvement is observed when more precise localization is required (larger IoU criterion), indicating our method improves localization performance over the baseline.

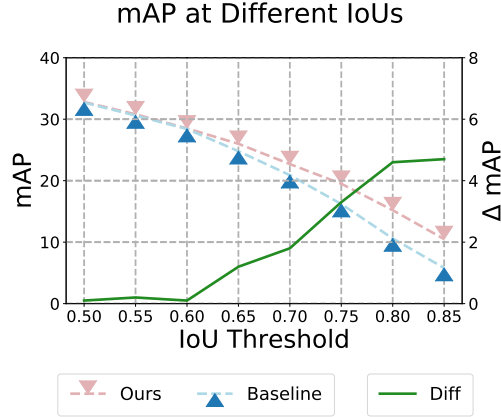


Figure 3: **Evaluating the localization precision of pseudo boxes** from teacher model under 2% COCO setting. We evaluate the quality of pseudo boxes with the withheld ground-truth annotations. Our method produces pseudo boxes with higher localization precision.

**Improvement on localization performance.** Having demonstrated the overall efficacy of our approach, we now evaluate the localization performance. We first compare our method against the baseline without the proposed components at different IoU criteria. As shown in Figure 2, our method improves baseline by a larger margin when higher localization precision is required: the performances are similar at 0.5 IoU threshold, whereas our approach obtains more than 6% higher mAP at 0.85 IoU threshold. We further evaluate the performance of teacher models on the withheld unlabeled images and see whether pseudo labels produced by our method are better localized. Similar trends in Figure 3 confirm that pseudo labels produced by our method are localized more precisely, and thus improve the detection performance for semi-supervised object detection.

**Improvements for underrepresented classes.** To validate the effectiveness of our method on improving the detection performance for underrepresented classes, we also show results (Table 4) on the rarest 10 classes in terms of number of annotations in the training set. We can see after adding dynamic pseudo label thresholding and loss re-weighting methods described in Sec. 2.3, the overall performance is improved by 0.9% AP (from 21.6% to 22.5%) and the performance on rare classes is improved by 2.1% AP (from 23.9% to 26.0%). This confirms that our method indeed promotes the performance for underrepresented classes.

**Qualitative results.** In addition to the quantitative analysis presented above, we provide some qualitative results in Figure 4. As can be observed, our method produces more precise localization results than the baseline without proposed components in Sec. 2.2 and 2.3. In particular, our method is better at localizing boundaries of irregular-shaped objects like the bear and person in Figure 4.

### 3.3 Ablation Study

**Effectiveness of different components.** We validate the effectiveness of proposed components and summarize the results in Table 5. We can see by adding the certainty-aware pseudo labels, class-specific loss re-weighting and dynamic thresholding, the performance is improved by 1.7%,

Table 3: **Performance of our method with model ensemble** (a simplified version of [49, 45]), indicating our method is complementary to existing approaches.

2% COCO	AP <sub>50</sub>	AP <sub>75</sub>	AP
Ensemble	37.1 → <b>37.9</b>	23.7 → <b>24.1</b>	22.5 → <b>23.0</b>

Table 4: **Performance improvement on the rarest 10 classes.**

2% COCO	AP
Overall	21.6 → <b>22.5</b>
Rarest 10 Classes	23.9 → <b>26.0</b>

Table 5: **Effectiveness of proposed components** including certainty-aware pseudo labels (CA), loss re-weighting (RE) and dynamic thresholding (DT).

CA	RE	DT	AP <sub>50</sub>	AP <sub>75</sub>	AP
			36.9	19.4	19.9
✓			35.3	22.9	21.6
✓	✓		36.2	23.2	22.1
✓	✓	✓	<b>37.1</b>	<b>23.7</b>	<b>22.5</b>

Table 6: **Effectiveness of different data augmentations** applied when training the student model, including color jitter (Color), Gaussian blur (Blur) and Cutout.

Color	Blur	Cutout	AP <sub>50</sub>	AP <sub>75</sub>	AP
			33.9	21.3	20.3
✓			34.9	22.4	21.1
✓	✓		35.5	22.7	21.5
✓	✓	✓	<b>37.1</b>	<b>23.7</b>	<b>22.5</b>

Table 7: Hyper-parameter sensitivity on number of intervals  $K$ .

$K$	AP <sub>50</sub>	AP <sub>75</sub>	AP
4	37.2	18.9	19.7
8	38.1	21.1	21.0
20	36.3	23.4	22.3
30	37.1	23.7	22.5
40	37.9	23.4	22.4

Table 8: Hyper-parameter sensitivity on variance controlling factors  $\gamma_1$  and  $\gamma_2$ .

$\gamma_1$	$\gamma_2$	AP <sub>50</sub>	AP <sub>75</sub>	AP
0.05	0.4	36.6	23.2	22.2
0.05	0.6	37.1	23.7	22.5
0.05	0.8	37.0	23.7	22.5
0.03	0.6	36.8	23.6	22.3
0.07	0.6	36.7	23.3	22.3

0.5% and 0.4%. When all the components are added, our approach improves the baseline by 2.6% AP and 4.3% AP<sub>75</sub>, confirming the proposed components are effective and especially useful for improving localization quality.

**Data augmentations.** We also study the usefulness of different data augmentation techniques. Table 6 summarizes the results. When no data augmentation is applied for training the student model, the performance degrades from 22.5% to 20.3% AP, indicating that data augmentation is critical. Adding color jittering and Gaussian blurring improves the performance by 1.2%, and applying Cutout further boosts AP by 1%. We also tried applying Dropblock [13] on intermediate features which is used in ISMT [45] but did not observe further improvements, possibly due to its similar functionality with Cutout [9].

**Hyper-parameter Sensitivity.** We experiment with different hyper-parameters and summarize the results in Table 7 and 8. For localization, our method is robust to hyper-parameter selection as long as  $K$  is large enough to produce fine-grained localization intervals. For example, with  $K$  set between (20, 40), the performance differs no larger than 0.2%. However, when  $K$  is set to be a small number like 4, the intervals are too coarse and the localization branch degenerates to a similar form of pure regression method, resulting in a degraded performance of 19.7% AP. For dynamic thresholding and loss re-weighting, larger  $\gamma_1$  and  $\gamma_2$  leads to more emphasis on infrequent classes during training, and we find using  $\gamma_1 = 0.05$  and  $\gamma_2 = 0.6$  gives the best result, as shown in Table 8. When set as 0, the corresponding method is disabled.



Figure 4: **Visualization of localization quality.** Our method (Green) localizes objects more precisely than the baseline (Blue). Best viewed in color.



## 4 Related Work

**Object Detection.** As a fundamental computer vision task, object detection has been extensively studied for decades [39, 12, 33, 32, 27, 25, 50, 37, 4]. Modern object detectors have evolved from anchor-based detectors like Faster RCNN [33], YOLO [32] and SSD [27], to anchor-free [37, 10, 50] and transformer-based detectors [4] in the past few years in pursuit of simpler formulation and stronger performance. A variety of directions have also been actively explored on improving localization precision [18, 40, 31, 16], inference efficiency [30, 38], training paradigms [47, 48, 22, 19], to name a few. While powering a wide range of applications, standard object detectors require box-level annotations for all objects-of-interest in images during training, which are time-consuming and labour-intensive to obtain. Our work follows the semi-supervised learning setting and exploits readily available unlabeled images to improve detection performance.

**Semi-Supervised Learning.** Semi-supervised learning (SSL) for visual understanding leverages unlabeled images for improved performance in various tasks [2, 2, 43, 34, 29, 1, 44, 21]. Recent advances on SSL mostly resort to consistency-based methods with data augmentation and have significantly improved performance for image classification. Specifically, the model is incentivized to produce consistent predictions across different views of an input image generated with semantics-preserving data augmentations. Typical approaches like MixMatch [2] and UDA [43] enforce a consistent prediction of class distributions across multiple views, while FixMatch [34] encourages correct predictions on strongly augmented unlabeled images given one-hot pseudo labels generated on weakly augmented ones. Data augmentations used in existing methods span conventional techniques [9, 46], learned augmentations [6, 7] and adversarially generated ones [29]. A line of work following Mean Teacher [36] also explore updating teacher model with an Exponential Moving Average (EMA) of student model during SSL training to provide better pseudo labels [34, 3]. Our work follows the consistency-based paradigm with a focus on object detection, which is relatively under-explored compared to image classification yet it requires fine grained annotations.

**Semi-Supervised Object Detection.** The expensive labeling cost of object detection has also drawn a growing attention on developing effective SSL methods. CSD [17] enforces consistent predictions on original and horizontally flipped images, whereas STAC [35] encourages consistency between weakly and strongly augmented views of images as in FixMatch [34]. On top of them, methods like Unbiased Teacher [28], ISMT [45] and Instant-Teaching [49] update the teacher model online with an evolving student model in a similar way of Mean Teacher [36]. Instant-Teaching [49] and ISMT [45] further explore training an ensemble of two model backbones/heads like Co-teaching [14] for better performance. A multi-phase learning method is also introduced in [41] to combat the noise in pseudo labels. While semi-supervised object detection performance has been steadily improved, most current approaches directly leverage recent advances on semi-supervised image classification for object detection. In contrast, we investigate and address the unique challenge of semi-supervised object detection—injecting localization precision to generate better boxes and dynamically adjusting pseudo label threshold to combat class imbalance.

## 5 Conclusion

In this paper, we rethink the use of pseudo labels for semi-supervised object detection (SSOD), and equip pseudo labels to be certainty-aware so as to address the lack of localization confidence when generating pseudo labels and the amplified class imbalance. We presented certainty-aware pseudo labeling considering both classification and localization quality by formulating box localization as a classification problem. Conditioned on the quality scores, the pseudo labels are filtered by dynamically derived thresholds and the losses are re-weighted in a class-specific manner, in pursuit of improved localization quality and balanced network learning for SSOD. Extensive experiments under multiple settings demonstrated the efficacy of our method.

## 6 Broader Impact

We investigate the problem of semi-supervised object detection, which alleviates negative impacts of the labour-intensive and expensive label annotation process. Our method promotes the performance of underrepresented classes in object detection datasets, and can hopefully mitigate the fairness issues in real-world applications as well. While the need of annotation is reduced, training the models on large-scale unlabeled datasets is still resource-consuming, which is a possible limitation of our work. The negative impacts could be that the system might be exploited by companies that attempt to use it for criminal activities.

**Acknowledgement** This work is supported by IARPA via Department of Interior/Interior Business Center (DOI/IBC) contract number D17PC00345.

## References

- [1] P. Bachman, O. Alsharif, and D. Precup. Learning with pseudo-ensembles. In *NeurIPS*, 2014. 1, 9
- [2] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, 2019. 1, 9
- [3] Z. Cai, A. Ravichandran, S. Maji, C. Fowlkes, Z. Tu, and S. Soatto. Exponential moving average normalization for self-supervised and semi-supervised learning. In *CVPR*, 2021. 9
- [4] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 9
- [5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 1
- [6] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le. Autoaugment: Learning augmentation policies from data. In *CVPR*, 2019. 9
- [7] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR Workshops*, 2020. 9
- [8] A. Dave, P. Dollár, D. Ramanan, A. Kirillov, and R. Girshick. Evaluating large-vocabulary object detectors: The devil is in the details. *arXiv preprint arXiv:2102.01066*, 2021. 4
- [9] T. DeVries and G. W. Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 3, 8, 9
- [10] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian. Centernet: Keypoint triplets for object detection. In *ICCV*, 2019. 9
- [11] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 1, 2, 5
- [12] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 2009. 9
- [13] G. Ghiasi, T.-Y. Lin, and Q. V. Le. Dropblock: A regularization method for convolutional networks. In *NeurIPS*, 2018. 8
- [14] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018. 2, 6, 9
- [15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5
- [16] Y. He, C. Zhu, J. Wang, M. Savvides, and X. Zhang. Bounding box regression with uncertainty for accurate object detection. In *CVPR*, 2019. 9
- [17] J. Jeong, S. Lee, J. Kim, and N. Kwak. Consistency-based semi-supervised learning for object detection. In *NeurIPS*, 2019. 1, 5, 6, 9
- [18] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang. Acquisition of localization confidence for accurate object detection. In *ECCV*, 2018. 9
- [19] K. Kim and H. S. Lee. Probabilistic anchor assignment with iou prediction for object detection. In *ECCV*, 2020. 9
- [20] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Mallocci, A. Kolesnikov, et al. The open images dataset v4. *IJCV*, 2020. 1
- [21] S. Laine and T. Aila. Temporal ensembling for semi-supervised learning. In *ICLR*, 2017. 1, 9
- [22] H. Li, Z. Wu, C. Zhu, C. Xiong, R. Socher, and L. S. Davis. Learning from noisy anchors for one-stage object detection. In *CVPR*, 2020. 9
- [23] X. Li, W. Wang, X. Hu, J. Li, J. Tang, and J. Yang. Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection. *arXiv preprint arXiv:2011.12885*, 2020. 2
- [24] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 5
- [25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 9
- [26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 2, 5

- [27] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 9
- [28] Y.-C. Liu, C.-Y. Ma, Z. He, C.-W. Kuo, K. Chen, P. Zhang, B. Wu, Z. Kira, and P. Vajda. Unbiased teacher for semi-supervised object detection. In *ICLR*, 2021. 1, 3, 5, 6, 9
- [29] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE TPAMI*, 2018. 1, 9
- [30] M. Najibi, B. Singh, and L. S. Davis. Autofocus: Efficient multi-scale inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9745–9755, 2019. 9
- [31] H. Qiu, H. Li, Q. Wu, and H. Shi. Offset bin classification network for accurate object detection. In *CVPR*, 2020. 2, 9
- [32] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 9
- [33] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 5, 9
- [34] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020. 1, 9
- [35] K. Sohn, Z. Zhang, C.-L. Li, H. Zhang, C.-Y. Lee, and T. Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 1, 3, 5, 6, 9
- [36] A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 1, 6, 9
- [37] Z. Tian, C. Shen, H. Chen, and T. He. FCOS: Fully convolutional one-stage object detection. In *ICCV*, 2019. 9
- [38] B. Uzkent, C. Yeh, and S. Ermon. Efficient object detection in large images using deep reinforcement learning. In *WACV*, 2020. 9
- [39] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001. 9
- [40] J. Wang, W. Zhang, Y. Cao, K. Chen, J. Pang, T. Gong, J. Shi, C. C. Loy, and D. Lin. Side-aware boundary localization for more precise object detection. In *ECCV*, 2020. 2, 9
- [41] Z. Wang, Y. Li, Y. Guo, L. Fang, and S. Wang. Data-uncertainty guided multi-phase learning for semi-supervised object detection. In *CVPR*, 2021. 6, 9
- [42] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019. 5
- [43] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le. Unsupervised data augmentation for consistency training. In *NeurIPS*, 2020. 1, 9
- [44] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le. Self-training with noisy student improves imagenet classification. In *CVPR*, 2020. 1, 9
- [45] Q. Yang, X. Wei, B. Wang, X.-S. Hua, and L. Zhang. Interactive self-training with mean teachers for semi-supervised object detection. In *CVPR*, 2021. 3, 5, 6, 7, 8, 9
- [46] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 9
- [47] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *CVPR*, 2020. 9
- [48] X. Zhang, F. Wan, C. Liu, X. Ji, and Q. Ye. Learning to match anchors for visual object detection. *IEEE TPAMI*, 2021. 9
- [49] Q. Zhou, C. Yu, Z. Wang, Q. Qian, and H. Li. Instant-teaching: An end-to-end semi-supervised object detection framework. In *CVPR*, 2021. 1, 3, 5, 6, 7, 9
- [50] X. Zhou, D. Wang, and P. Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 9