# Double-Check Soft Teacher for Semi-Supervised Object Detection

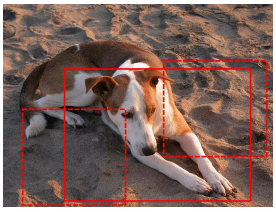
**Kuo Wang**[1] , **Yuxiang Nie**[1] , **Chaowei Fang**[2] , **Chengzhi Han**[3] , **Xuewen Wu**[3] , **Xiaohui Wang Wang**[3] , **Liang Lin**[1] , **Fan Zhou**[1] and **Guanbin Li**[1*]
[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2]School of Artificial Intelligence, Xidian University, Xi'an, China
[3]Huawei Technologies Co. China
{wangk229, nieyx3}@mail2.sysu.edu.cn, chaoweifang@outlook.com, {hanchengzhi, wuxuewen1, ross.xhwang}@huawei.com, linliang@ieee.org, {isszf,liguanbin}@mail.sysu.edu.cn

## Abstract

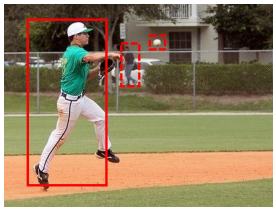
In the semi-supervised object detection task, due to the scarcity of labeled data and the diversity and complexity of objects to be detected, the quality of pseudo-labels generated by existing methods for unlabeled data is relatively low, which severely restricts the performance of semi-supervised object detection. In this paper, we revisit the pseudo-labeling based Teacher-Student mutual learning framework for semi-supervised object detection and identify that the inconsistency of the location and feature of the candidate object proposals between the Teacher and the Student branches are the fatal cause of the low quality of the pseudo labels. To address this issue, we propose a simple yet effective technique within the mainstream teacher-student framework, called Double Check Soft Teacher, to overcome the harm caused by insufficient quality of pseudo labels. Specifically, our proposed method leverages teacher model to generate pseudo labels for the student model. Especially, the candidate boxes generated by the student model based on the pseudo label will be sent to the teacher model for "double check", and then the teacher model will output probabilistic soft label with background class for those candidate boxes, which will be used to train the student model. Together with a pseudo labeling mechanism based on the sum of the TOP-K prediction score, which improves the recall rate of pseudo labels, Double Check Soft Teacher consistently surpasses state-of-the-art methods by significant margins on the MS-COCO benchmark, pushing the new state-of-the-art. Source codes are available at https://github.com/wkfdb/DCST.

## 1 Introduction

With the development in computing power and the availability of large-scale labeled datasets, deep learning has achieved great breakthroughs in various tasks. However, these advances are accompanied by a large amount of time-consuming and labor-intensive data annotation efforts, especially for object detection tasks. To make better use of unlabeled data, semi-supervised learning has received increasing attention. However, so far the research progress of semi-supervised learning is mainly concentrated in the field of image classification [Berthelot *et al.*, 2019; Sohn *et al.*, 2020a], and there are still many unsolved problems in its application to object detection where the data labeling work is more tedious and is thus more worth investigating. In this paper, a simple but effective solution is proposed by deeply revisiting the essential reasons that the pseudo-label-based semi-supervised object detection framework may have excessive noise and low recall rate that seriously restrict the model performance (refer to Figure 1).



(a) Object bounding box pseudo-labels are full of noises, e.g., the two dashed boxes in the image that mainly contain background regions will be mistaken for the "dog" category.



(b) The low recall rate of the pseudo-labeled harms the model training, e.g., small objects (ball and person in the image) will be ignored and misclassified as background.

Figure 1: Low-quality pseudo-labels and their possible detrimental analysis. The solid boxes are the pseudo-labels adopted by the model during training, and the dashed boxes are the candidate object bounding box generated during training.

Pseudo-label-based semi-supervised methods [Liu *et al.*, 2021; Xu *et al.*, 2021] consist the mainstream solution for semi-supervised object detection. Among them, the quality of pseudo-labels is a key factor affecting performance. Current works have carried out a lot of exploration on "generating high-quality pseudo-labels" [Yang *et al.*, 2021; Tang *et al.*, 2021b; Zhou *et al.*, 2021]. However, for object detection tasks, high quality pseudo-label requires both accurate classification, precise location, and high recall rate, which makes "generating high quality pseudo-labels" ex-

*Corresponding author.

tremely formidable. In contrast to existing works, we identify the essential reasons for the adverse effects of low-quality pseudo-labels and propose a method to "combat low-quality pseudo-labels". Specifically, conventional pseudo-labels are tailored for classification tasks, and applying them directly to object detection leads to the following two critical problems:

i) The bounding box position corresponding to the pseudo-label is not accurate enough, which may lead to inconsistent semantic content in the model training although the candidate bounding boxes and pseudo-labels have enough overlap, as shown in Figure 1(a).

ii) The low recall rate of pseudo-labels will cause a large number of candidate boxes to be mistakenly regarded as background category due to insufficient matching of pseudo-labels to mislead model training, as shown in Figure 1(b).

To solve the above issues, we propose Double-Check Soft Teacher (DCST), an approach that corrects the candidate box labeling errors caused by the low quality pseudo labels, basing on the mainstream Teacher-Student architecture [Sohn *et al.*, 2020a], where the Teacher generates pseudo labels on weakly augmented images to train the Student with strongly augmented images. In particular, our method sends the candidate boxes generated by the student model with reference to the pseudo-labels into the teacher model for further verification, and then assigns probabilistic soft labels (including background) to each candidate box based on features extracted from weakly augmented images. We call this process as "Double Check". This mechanism enables the model to identify the mislabeled candidate boxes and re-label them correctly, which can be seen as the teacher checking the students' homework, correcting the errors while retaining the correct results. Accompanying the "Double Check", candidate boxes with low confidence may appear, which are not suitable to be assigned one-hot hard label, and the perfect soft label hence comes into being. In this way, incorrect background boxes still have a chance to be corrected as foreground (certain prediction distribution of foreground categories) to participate in training, although they are not sure of their exact class.

On the other hand, considering the negative impact of the low recall rate of pseudo-labels on the model performance, we introduce "TOP-K pseudo labeling", which leverages the sum of the cumulative Top-K probability predictions greater than a certain threshold as the selection basis for pseudo-labels. In fact, many objects of different categories share similar appearances, such as "sofa" and "chair", "TV" and "laptop". Under the condition of semi-supervised learning with limited labels, it is arduous for the model to generate sufficiently high-confident predictions for a specific category, which makes it easy for the model to miss a large number of high-quality hard foreground objects in the process of selecting pseudo-labels, resulting in an excessively low recall rate. TOP-K pseudo labeling enables the model to detect similar objects and provide more candidate object boxes for the double check processing, which would further improve the low classification accuracy.

To sum up, we highlight the contributions of this paper as follows:

- We identify the unreliable positions and the low recall rate consists of the essential reasons for the adverse effects of low-quality pseudo-labels.
- The novel Double-Check Soft Teacher with TOP-K pseudo labeling approach is proposed to prevent the detrimental effects due to noisy pseudo-labels.
- On the MS-COCO semi-supervised object detection benchmark, our model consistently performs favorably against the state-of-the-art methods by significant margins.

## 2 Related Work

### 2.1 Semi-supervised Learning

Semi-supervised learning (SSL) aims to make full use of unlabeled data to facilitate the model learning process with limited labeled data. Most of the semi-supervised learning methods aim at image classification tasks. There are two mainstream paradigms of these works, consistency based methods and pseudo label based methods. Consistency based SSL methods introduce a regularization loss to make the model produce similar predictions for the same image under different disturbance. There are many different ways to implement the disturbance, such as applying different kinds of data augmentations for the same image [Laine and Aila, 2017; Tarvainen and Valpola, 2017], or following the idea of adversarial training to perturb the input images along the adversarial direction [Miyato *et al.*, 2018]. Pseudo-labeling based (self-training based) methods [Xie *et al.*, 2020b; Li *et al.*, 2019] refer to generating pseudo labels for unlabeled images and then using them for supervised training. Recently, data augmentations have proven to be an effective paradigm for boosting SSL on image classification [Berthelot *et al.*, 2019; Xie *et al.*, 2020a; Sohn *et al.*, 2020a]. Our work also adopts different data augmentations to achieve consistent regularization, and follows the idea of pseudo labeling to deal with unlabeled data, but we focus on the consistency regularization of bounding-box level prediction.

### 2.2 Semi-supervised Object Detection

Similar to semi-supervised image classification, semi-supervised object detection also follows the two aforementioned algorithm paradigms, including consistency [Jeong *et al.*, 2019; Tang *et al.*, 2021a] based method and pseudo labeling based models [Sohn *et al.*, 2020b; Wang *et al.*, 2021; Xu *et al.*, 2021; Liu *et al.*, 2021; Yang *et al.*, 2021; Tang *et al.*, 2021b; Zhou *et al.*, 2021]. Recently, pseudo labeling based methods have developed rapidly. Among them, STAC [Sohn *et al.*, 2020b] first proposed a paradigm of generating pseudo labels using weakly augmented images and training with strongly augmented images. This method has become a classic benchmark recently. For pseudo labeling based method, the quality of pseudo labels is the key factor affecting the performance. The improvements achieved by the works after STAC [Wang *et al.*, 2021; Xu *et al.*, 2021; Liu *et al.*, 2021; Yang *et al.*, 2021; Tang *et al.*, 2021b; Zhou *et al.*, 2021] are basically to improve the quality of

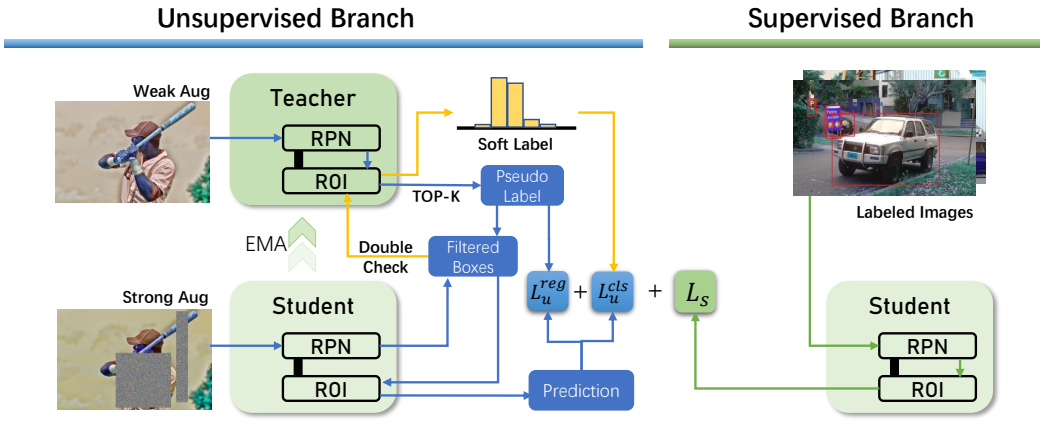



Figure 2: The overall framework of double check soft teacher, which includes a supervised branch and an unsupervised branch. In the supervised branch, the student model calculates the supervised loss directly from the labeled images. In the unsupervised branch, teacher model first generates pseudo labels based on weakly augmented images. After the student generates the filtered boxes with reference to the pseudo-labels, the teacher model updates the predictions of the filtered boxes and generate soft labels for them according to the features of the weakly augmented image version.

pseudo labels. However, most of their proposed methods can not well address the inaccurate location and low recall rate of pseudo-labels. Note that [Wang *et al.*, 2021] proposed a data uncertainty based method to detect uncertain candidate boxes and reduce their weights. [Xu *et al.*, 2021] used the background class score generated by the Teacher model as the weight of the background candidate boxes to reduce the ill effects of incorrect background predictions. However, those methods failed to fundamentally solve the mislabeling problem caused by the noisy labels. Our proposed DCST can detect and correct errors in all candidate boxes, fundamentally addressed the harmful effects of insufficient pseudo-label quality.

# 3 Method

### Problem Definition

In the semi-supervised object detection task, a set of labeled images $\boldsymbol{D}_s = \{\boldsymbol{x}_i^s, \boldsymbol{y}_i^s\}_{i=1}^{N_s}$ and a set of unlabeled images $\boldsymbol{D}_u = \{\boldsymbol{x}_i^u\}_{i=1}^{N_u}$ are available for training. $N_s$ and $N_u$ denote the number of labeled data and unlabeled data, respectively. For each image $\boldsymbol{x}_i$, the corresponding label $\boldsymbol{y}_i$ contains the position and category information of object instances contained in the image. The goal is to make use of both labeled and unlabeled data to learn the object detection model.

Pseudo labeling is one of the most widely adopted algorithms to involve unlabeled data during training. Traditional pseudo-label learning mechanisms designed for image classification cannot ensure both the accuracy of the generated bounding box localization and the high recall rate of box-level pseudo-labels. In order to overcome the hazards caused by low-quality pseudo labels, we propose Double-Check Soft-Teacher (DCST), an approach that corrects the candidate box labeling errors caused by the low quality pseudo labels, base on the mainstream Teacher-Student architecture. Furthermore, a new sample filtering scheme called "TOP-K pseudo labeling" which leverages the sum of the cumulative Top-K probability predictions greater than a certain threshold as the selection basis for pseudo-labels is introduced for improving the recall rate of pseudo labels. The overall pipeline of our method is shown in Figure 2.

## 3.1 End2End Teacher-Student

The Teacher-Student Mutual Learning framework [Liu *et al.*, 2021] is employed to generate pseudo labels for unlabeled images. During the network optimization procedure, every training batch is composed of several randomly sampled labeled and unlabeled images. Labeled images can be directly used for training in the supervised manner. The weakly and strongly augmented counterparts of unlabeled images are regarded as inputs of the teacher and student model respectively. The outputs of the teacher model are leveraged to construct pseudo labels for unlabeled images, which are subsequently applied for training the student model. We build up the student model with the two-stage object detection network Faster-RCNN, consisting a region proposal subnetwork (RPN) for inferring object proposals and a ROI head for predicting class probabilities and adjusting bounding boxes positions. Denote the parameters of the student model be $\theta_s$. The teacher model shares the same architecture with the student model. Suppose its parameters be $\theta_t$, then it is updated via the exponential moving average (EMA),

$$\theta_t = \alpha\theta_t + (1-\alpha)\theta_s, \tag{1}$$

where $\alpha$ is a constant indicating the ensemble ratio between the teacher and the student.

The student model is trained with the supervised loss on labeled images and the unsupervised loss on unlabeled images respectively. For labeled images, the supervised loss $\mathcal{L}_{sup}$ is consisting of classification and regression losses calculated on outputs of the RPN and ROI head,

$$\mathcal{L}_{sup} = \sum_i \mathcal{L}_{cls}^{rpn}(\boldsymbol{x}_i^s, \boldsymbol{y}_i^s) + \mathcal{L}_{reg}^{rpn}(\boldsymbol{x}_i^s, \boldsymbol{y}_i^s) + \mathcal{L}_{cls}^{roi}(\boldsymbol{x}_i^s, \boldsymbol{y}_i^s) + \mathcal{L}_{reg}^{roi}(\boldsymbol{x}_i^s, \boldsymbol{y}_i^s) \tag{2}$$

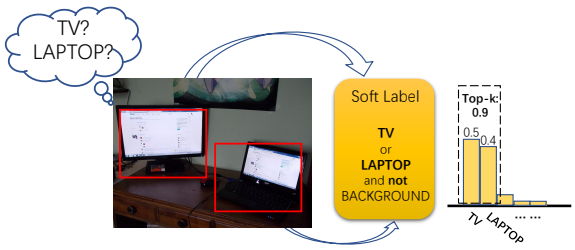



Figure 3: It is arduous for the model to generate sufficiently high-confident predictions for some specific categories, which results in an excessively low recall rate. With the TOP-K pseudo-labeling, more objects can be mined from the unlabeled data for training.

For the unlabeled image $\boldsymbol{x}_i^u$, the detected objects having inference confidence larger than a threshold $\sigma$ in the teacher model are regarded as pseudo labels $\widehat{\boldsymbol{y}}_i^u$. Then the unsupervised loss is calculated with the pseudo labels $\widehat{\boldsymbol{y}}_i^u$ ,

$$\mathcal{L}_{unsup} = \sum_i \mathcal{L}_{cls}^{rpn}\left(\boldsymbol{x}_i^u, \widehat{\boldsymbol{y}}_i^u\right) + \mathcal{L}_{reg}^{rpn}\left(\boldsymbol{x}_i^u, \widehat{\boldsymbol{y}}_i^u\right) + \mathcal{L}_{cls}^{roi}\left(\boldsymbol{x}_i^u, \widehat{\boldsymbol{y}}_i^u\right) + \mathcal{L}_{reg}^{roi}\left(\boldsymbol{x}_i^u, \widehat{\boldsymbol{y}}_i^u\right), \tag{3}$$

Finally, the overall objective function for optimizing parameters of the student model is defined as the weighted sum of the supervised loss and the unsupervised loss:

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda \mathcal{L}_{unsup}, \tag{4}$$

where $\lambda$ is a constant. The SGD optimizer is adopted to update parameters of the student model. After every optimization step, the teacher model is updated according to Eq. 1.

### 3.2 Double-Check Mechanism

In the training process of the object detection task, the two-stage object detector first generates candidate boxes with RPN. Then, the candidate boxes are assigned with labels according to the provided annotations. For each candidate box, the best matched annotation is regarded as the ground-truth, if the IoU between the candidate box and the annotation is larger than a certain threshold; otherwise, the candidate is marked as background. In the semi-supervised scenario, due to the inaccurate location and low recall rate of pseudo labels, the assigned labels of candidate boxes usually contain a lot of errors, as shown in Figure 1. In order to correct these errors, we propose the double-check mechanism.

The double-check mechanism is ingenious in that the teacher model reloads the position of filtered boxes to extract features from the weakly augmented images through ROI Pooling. For each filtered box, a new reliable category label is generated from the prediction result inferred by the ROI head of the teacher model, as shown by the path constituted by yellow arrows in Figure 2.

If the original category label of the filtered box is correct, the prediction distribution tends to be consistent with the original category label; otherwise, the original category label is corrected by the new reliable category probabilities.

### 3.3 Soft Teacher

Assuming that there are $n$ types of objects in the data set, the Teacher will generate a prediction distribution of length $n+1$ for each filtered box in the double-check mechanism. This prediction distribution represents the probability that the filtered box belongs to each class, including the background. We directly take this distribution as the category soft label to train the student model.

There are two benefits to this approach. First, soft label enhances the double-check correction capability. This enables false background boxes to be corrected as foreground boxes, even if the model cannot determine the specific foreground class. Secondly, soft labels can be utilized for exploring the underlying similarity cues among classes. For example, as shown in Figure 3, the model may not be able to confirm whether the object in the detected box is a TV or a laptop, it is sure that the object does not belong to any other categories. Such kind of boxes can be leveraged in the form of soft labels for training the student model.

With the help of soft labels, the classification loss of ROI in unsupervised loss is calculated with the soft cross-entropy, as shown below:

$$\mathcal{L}_{cls}^{roi}\left(\boldsymbol{x}_i^u, \widehat{\boldsymbol{y}}_i^u\right) = \frac{1}{N_i}\sum_j^{N_i} -\log\_\operatorname{softmax}\left(\boldsymbol{p}_j\right) * \widehat{\boldsymbol{y}}_j^{cls}, \tag{5}$$

where $N_i$ is the number of filtered boxes for training in image $\boldsymbol{x}_i^u$, $\boldsymbol{p}_j$ is the prediction logits of the j-th filtered box, and $\widehat{\boldsymbol{y}}_j^{cls}$ is its corresponding soft label.

### 3.4 TOP-K Pseudo Labeling

The employment of the double-check mechanism relaxes the requirement for pseudo labels, which makes high precision on category no longer important. Meanwhile, the soft label mechanism enhances the error correction ability and enables the model to handle similar-looking objects. Based on these two characteristics, we further propose a TOP-K pseudo-labeling strategy to promote the recall rate of pseudo labels.

The conventional pseudo-label screening scheme uses the top-1 score of the foreground classes in the prediction result as the confidence of the detection box. This approach easily overlooks hard samples as shown in Figure 1, leading to low recall rate in pseudo labels. To tackle this problem, we propose a TOP-K pseudo labeling scheme, through thresholding the accumulated value of the top-k scores. Such pseudo labeling strategy is beneficial for preserving similar-looking objects as pseudo-labels, thereby improving the recall rate of pseudo-labels. TOP-K pseudo labeling aggravates the class uncertainty of pseudo-labels, but the double-check mechanism eliminates the negative influence of unreliable pseudo category labels. Meanwhile, the additional look-alike objects brought by TOP-K pseudo labeling can further benefit the learning of the student model under the soft label mechanism.

## 4 Experiments

**Dataset.** Following existing works, we test the performance of our method on MS-COCO benchmark [Lin *et al.*, 2014]. Subsets `train2017` and `unlabeled2017` are used for

| Methods | 1%COCO | 5%COCO | 10%COCO | 100%COCO |
|---|---|---|---|---|
| Supervised | 9.05±0.16 | 18.47±0.22 | 23.86±0.81 | 37.63 |
| Supervised* [Sohn *et al.*, 2020b] | 9.83±0.23 | 21.18±0.20 | 26.18±0.12 | 39.48 |
| Supervised† | 10.00±0.26 | 20.92±0.15 | 26.94±0.11 | 40.90 |
| STAC [Sohn *et al.*, 2020b] | 13.97±0.35 | 24.38±0.12 | 28.64±0.21 | 39.20 |
| ISMT [Yang *et al.*, 2021] | 18.88±0.74 | 26.37±0.24 | 30.53±0.52 | 39.64 |
| Instant-Teaching [Zhou *et al.*, 2021] | 18.05±0.15 | 26.75±0.05 | 30.40±0.05 | 40.20 |
| Humble-Teacher [Tang *et al.*, 2021b] | 16.96±0.38 | 27.70±0.15 | 31.61±0.28 | 42.37 |
| Unbiased-Teacher [Liu *et al.*, 2021] | 20.75±0.12 | 28.27±0.11 | 31.50±0.10 | 41.30 |
| Soft-Teacher [Xu *et al.*, 2021] | 20.46±0.39 | 30.74±0.08 | 34.04±0.14 | 44.50 |
| DCST(Ours) | **23.02**±0.23 | **32.10**±0.15 | **35.20**±0.20 | **44.60** |

Table 1: Comparison with state-of-the-art methods using different percents of labeled data. 'Supervised' means only labeled images are employed during training. In 'Supervised*', the same argumentation scheme as in STAC are applied for expanding labeled images. In 'Supervised†', the weak augmentation strategies adopted by our method is applied for expanding labeled images.

training, and subset `val2017` is used for testing. Our experiments are divided into two protocols: partially labeled data and fully labeled data. In partially labeled data protocol, we respectively sample 1%, 5% and 10% of `train2017` as labeled data, and the remaining is used as unlabeled data. For every setting, five trials are conducted using different random seeds for random data sampling, and we report the averaged evaluation metrics. In the fully labeled data protocol, all images in `train2017` are regarded as labeled data, and all images in `unlabelled2017` are regarded as unlabeled data. The standard mean average precision (mAP) on `val2017` is used as the evaluation metric.

**Implementation Details.** For fair comparisons, we use Fast-RCNN [Ren *et al.*, 2015] built with ResNet50 [He *et al.*, 2016] and feature pyramid networks [Lin *et al.*, 2017] as the object detector. The code is implemented under the framework of MMdetection [Chen *et al.*, 2019]. Top-3 classes are considered for generating pseudo labels from predictions of the teacher model, and the threshold $\sigma$ is set to 0.7. Following existing works [Liu *et al.*, 2021; Xu *et al.*, 2021], we set $\alpha = 0.999$ in the EMA procedure and $\lambda = 4$ for the weight of unsupervised loss in Eq. 4. The same weak and strong data augmentation schemes in [Xu *et al.*, 2021] are chosen. For partially labeled data, the model is trained for 180,000 iterations on 4 GPUs. Each GPU processes ten images. The learning rate is initialized to 0.1 and is divided by 10 at the 120,000-th and 160,000-th iteration. For fully labeled data, the model is trained for 720,000 iterations on 8 GPUs. Each GPU processes eight images. $\lambda$ is set to 2. The learning rate is initialized as 0.1 and is divided by 10 at the 480,000-th and 640,000-th iteration. We follow the default settings in MMdetection to set other hyper-parameters.

## 4.1 Experimental Results

We compare our method against a series of existing methods including STAC [Sohn *et al.*, 2020b], ISMT [Yang *et al.*, 2021], Instant-Teaching [Zhou *et al.*, 2021], Humble-Teacher [Tang *et al.*, 2021b], Unbiased-Teacher [Liu *et al.*, 2021], and Soft-Teacher [Xu *et al.*, 2021], when partially or fully labeled data is leveraged for network optimization. Under scenarios of partially labeled data, our proposed method (DCST) consistently outperforms existing methods by large margins, as shown in Table 1. Owing to the capacity in refining the bounding boxes and the re-weighting mechanism for eliminating the influence of incorrect background boxes, Soft-Teacher achieves the best performance among existing methods. However, the mislabeling problem can not be fundamentally overcome by the re-weighting mechanism. Meanwhile, the impact of misclassified foreground objects is not taken into consideration. In contrast, our proposed DCST assigns reliable soft labels to the filtered candidate boxes, which simultaneously rectified the misclassified boxes both in foreground and background. Compared to Soft-Teacher, our method derives mAP gains of **2.56**, **1.36** and **1.16** on 1%, 5% and 10% labeled data, respectively. The improvement brought by our method increases as the percent of labeled data decreases. Considering less labeled data usually leads to pseudo labels with lower quality, the above phenomenon supports the argument that "combating low-quality pseudo labels" is a more effective solution in the protocol of limited labeled data.

The adoption of fully labeled data generates pseudo labels with relatively higher quality. Under this scenario, our method can still achieve the best performance, as shown by the last column of Table 1. This indicates that our method is capable of exploring unlabeled data more effectively, even when sufficient labeled data is provided.

The experimental results illustrate that the quality of pseudo labels is the key factor for promoting semi-supervised object detection. Compared to the strategy of discovering high-quality pseudo labels which is widely adopted by previous methods, "combating low quality pseudo labels" is a simpler while more effective solution. Moreover, further research is suggested to combine the two kinds of techniques for extracting reliable pseudo labels from unlabeled images more thoroughly.

## 4.2 Ablation Study

In this subsection, we conduct experiments to verify the efficacy of critical components in our method. Without specification, 10% labeled data is adopted for network optimization.

| foreground double-check | background double-check | soft label | mAP |
|---|---|---|---|
| - | - | - | 33.3 |
| - | ✓ | ✓ | 34.1 |
| ✓ | - | ✓ | 33.8 |
| ✓ | ✓ | - | 34.8 |
| ✓ | ✓ | ✓ | **35.5** |

Table 2: Performance of "double-check" and soft label mechanism.

| TOP-K | Precision(%) | Recall(%) | F-measure(%) | mAP |
|---|---|---|---|---|
| top-1 | 77.5 | 54.4 | 63.9 | 34.8 |
| top-2 | 76.9 | 56.2 | 64.9 | 35.5 |
| top-3 | 76.7 | 56.5 | 65.1 | 35.5 |
| top-4 | 76.6 | 56.6 | 65.1 | 35.3 |

Table 3: Performance of using different numbers of top classes for the pseudo labeling mechanism.

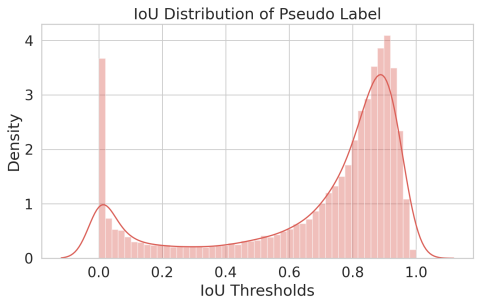



Figure 4: Distribution of the maximum IoU between pseudo labels and ground-truth labels.

| $\sigma$ | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|
| mAP | 34.9 | **35.6** | 35.5 | 35.1 | 34.7 |

Table 4: Performance of using different values for the pseudo labeling threshold $\sigma$. All results are obtained via top-3 pseudo labeling.

**Efficacy of Double-Check Soft Teacher.** Variants of our method are implemented via applying different double-check strategies. The experimental results are shown in Table 2. Employing the double-check strategy to post-processing foreground or background boxes improves the mAP metric by 0.5 or 0.8, respectively. Processing both kinds of boxes with the double-check strategy can bring more performance gain, leading to 2.2 higher mAP compared to the baseline without using the double-check strategy. We also attempt to re-implement the double-check module through directly allocating hard labels to object proposals. As shown by the second last row of Table 2, it produces results with 0.7 lower mAP, compared to the final version which assigns soft labels to object proposals. The reason is that soft labels can help to explore the underlying similarity cues among different categories and model the class probabilities more reasonably.

**Efficacy of TOP-K Pseudo Labeling.** Table 3 presents the performance of using different number of top classes for the pseudo labeling mechanism. The precision, recall, and F-measure evaluates the efficacy in harvesting class-agnostic annotations. A pseudo annotation is regarded as a true positive, once it hits certain ground-truth annotation with an over 0.5 IoU. As can be observed, higher recall and F-measure can be obtained after setting the number of top classes larger than 1. This indicates that the top-k pseudo labeling is capable of discovering more objects without sacrificing much precision. Consequently, the final object detection performance can also be improved. Meanwhile, we can observe that using different numbers of top classes larger than 1 has little influence on the pseudo labeling and object detection performance.

**Analysis about Quality of Pseudo Labels.** The distribution of the maximum IoU between pseudo labels and ground-truth labels is presented in Figure 4. The variation of precision/recall metrics against the "K" value for pseudo labeling is illustrated in Table 3. We can observe that, there exist enormous number of pseudo labels which severely deviate from any ground-truth labels. Rectifying those labels is critical to explore unlabeled images in semi-supervised object detection. Our proposed double-check mechanism can achieve this goal via re-allocating the class probabilities, hence producing promising object detection performance with small amount of labeled data.

**Choice for Pseudo Labeling Threshold.** Table 4 shows the performance of setting different values for the pseudo labeling threshold $\sigma$. The best performance is achieved when $\sigma$ is set to 0.6. Varying $\sigma$ to 0.7 causes little interference to the detection performance.

## 5 Conclusion

In this paper, we revisit the semi-supervised object detection task and point out two crucial problems of conventional pseudo labeling techniques: inaccurate position and low recall rate, which lead to category labeling errors during the training process. To overcome these issues, we propose Double-Check Soft Teacher, a novel algorithm to correct the errors among pseudo labels. Experiments on the COCO benchmark illustrate that our method outperforms the state-of-the-art methods under various percents of labeled data. Based on the experimental results, we conclude that "combating low-quality pseudo labels" is crucial for semi-supervised object detection.

## Acknowledgments

This work was supported in part by the Key-Area Research and Development Program of Guangdong Province under Grant No.2021B0101410003, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant No.2020B1515020048, in part by the National Natural Science Foundation of China under Grant No.61976250 and No.U1811463, and in part by the Guangzhou Science and technology project under Grant No.202102020633. This work was also supported by the Guangdong Provincial Key Laboratory of Big Data Computing, the Chinese University of Hong Kong, Shenzhen.

## References


[Berthelot *et al.*, 2019] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, page 5049–5059, 2019.

[Chen *et al.*, 2019] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Jeong *et al.*, 2019] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. *Advances in Neural Information Processing Systems*, 32:10759–10768, 2019.

[Laine and Aila, 2017] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[Li *et al.*, 2019] Xinzhe Li, Qianru Sun, Yaoyao Liu, Qin Zhou, Shibao Zheng, Tat-Seng Chua, and Bernt Schiele. Learning to self-train for semi-supervised few-shot classification. *Advances in Neural Information Processing Systems*, 32:10276–10286, 2019.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[Lin *et al.*, 2017] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[Liu *et al.*, 2021] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *ICLR*, 2021.

[Miyato *et al.*, 2018] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[Sohn *et al.*, 2020a] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 2020.

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

[Tang *et al.*, 2021a] Peng Tang, Chetan Ramaiah, Yan Wang, Ran Xu, and Caiming Xiong. Proposal learning for semi-supervised object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2291–2301, 2021.

[Tang *et al.*, 2021b] Yihe Tang, Weifeng Chen, Yijun Luo, and Yuting Zhang. Humble teachers teach better students for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3132–3141, 2021.

[Tarvainen and Valpola, 2017] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, pages 1195–1204, 2017.

[Wang *et al.*, 2021] Zhenyu Wang, Yali Li, Ye Guo, Lu Fang, and Shengjin Wang. Data-uncertainty guided multi-phase learning for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4568–4577, 2021.

[Xie *et al.*, 2020a] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 2020.

[Xie *et al.*, 2020b] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020.

[Xu *et al.*, 2021] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. *ICCV*, 2021.

[Yang *et al.*, 2021] Qize Yang, Xihan Wei, Biao Wang, Xian-Sheng Hua, and Lei Zhang. Interactive self-training with mean teachers for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5941–5950, 2021.

[Zhou *et al.*, 2021] Qiang Zhou, Chaohui Yu, Zhibin Wang, Qi Qian, and Hao Li. Instant-teaching: An end-to-end semi-supervised object detection framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4081–4090, 2021.