MixTeacher: Mining Promising Labels with Mixed Scale Teacher for Semi-Supervised Object Detection

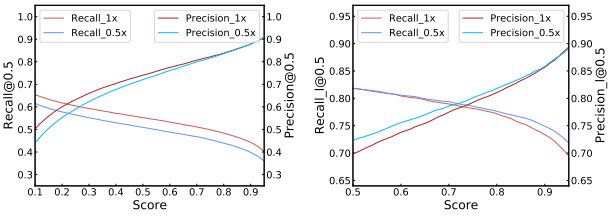
Liang Liu[1], Boshen Zhang[1], Jiangning Zhang[1], Wuhao Zhang[1], Zhenye Gan[1]
Guanzhong Tian[3], Wenbing Zhu[4], Yabiao Wang[1†], Chengjie Wang[1,2†],
[1]Youtu Lab, Tencent [2]Shanghai Jiao Tong University
[3]Ningbo Research Institute, Zhejiang University, [4]Rongcheer Co., Ltd
{leoneliu, boshenzhang, vtzhang, wuhaozhang, wingzygan}@tencent.com;
gztian@zju.edu.cn; louis.zhu@rongcheer.com; {caseywang, jasoncjwang}@tencent.com

Abstract

Scale variation across object instances remains a key challenge in object detection task. Despite the remarkable progress made by modern detection models, this challenge is particularly evident in the semi-supervised case. While existing semi-supervised object detection methods rely on strict conditions to filter high-quality pseudo labels from network predictions, we observe that objects with extreme scale tend to have low confidence, resulting in a lack of positive supervision for these objects. In this paper, we propose a novel framework that addresses the scale variation problem by introducing a mixed scale teacher to improve pseudo label generation and scale-invariant learning. Additionally, we propose mining pseudo labels using score promotion of predictions across scales, which benefits from better predictions from mixed scale features. Our extensive experiments on MS COCO and PASCAL VOC benchmarks under various semi-supervised settings demonstrate that our method achieves new state-of-the-art performance. The code and models are available at https://github.com/lliuz/MixTeacher.

1. Introduction

The remarkable performance of deep learning on various tasks can largely be attributed to large-scale datasets with accurate annotations. However, collecting a large amount of high-quality annotations is infeasible as it is labor-intensive and time-consuming, especially for tasks with complicated annotations such as object detection [23, 30] and segmentation [5, 6]. To reduce reliance on manual labeling, semi-supervised learning (SSL) has gained much attention. SSL aims to train models on a small amount of labeled images and a large amount of easily accessible unlabeled data.



(a) Precision and recall for all objects (b) Precision and recall for large objects

(c) Detection results with different input scales

Figure 1. Detection results with input of regular 1× scale and 0.5× down-sampled scale images. We plot the precision and recall under different score thresholds for (a) all objects and (b) large objects in COCO `val2017` with the same model but different input scales. Two examples of unlabeled images are given in (c). Large scale inputs have clear advantages in overall metrics, but down-sampled images are more suitable for large objects.

Following extensive pioneering studies on semi-supervised image classification [2, 14, 32], several methods on semi-supervised object detection have emerged.

Most early studies on semi-supervised object detection [13, 24, 33] can be considered as a direct extension of SSL methods designed for image classification, using a teacher-student training paradigm [2, 32, 35]. In these methods, a teacher model generates pseudo bounding boxes and corresponding class predictions on unlabeled images, and the pseudo labels are used to train a student model. Despite the performance improvement from using a large amount of unlabeled data, these methods overlooked the characteristics of object detection to some extent, resulting in a huge

† Corresponding Authors.

gap from the fully supervised counterpart.

Compared to image classification, object instances in detection tasks can vary in a wider range of scales. To address this challenge of detecting and localizing multiple objects across scales and locations, numerous works in object detection have been proposed, such as FPN [21], Trident [20], and SNIP [31]. However, the large scale variation brings new challenges in the semi-supervised context. In order to guarantee high precision, most existing semi-supervised object detection methods adopt strict conditions (*e.g.* score $>$ 0.9) to filter out highly confident pseudo labels. Although this ensures the quality of pseudo labels, many objects with low confidence are wrongly assigned as background, especially for those with extreme scales. As shown in Figure 1 (c), inappropriate scales will lead to false negatives, which can mislead the network in semi-supervised learning. We further observe the influence of the test scale of the images. Consistent with common sense, large-scale inputs have clear advantages in overall metrics, as shown in Figure 1 (a). However, down-sampled images show a superiority for large objects, as shown in Figure 1 (b). This provides a new view to handle the scale variation issue.

It is worth mentioning that recent works have paid attention to the scale variation issue in semi-supervised object detection. As shown in Figure 2 (a) and (b), previous methods have introduced an additional down-sampled view to encourage the model to make scale-invariant predictions. Specifically, SED [10] proposes to distill predictions of class probability from the regular scale to the down-sampled scale and constrain consistent predictions of localization for all proposals in two scales. PseCo [17] adopts the same pseudo labels generated from the regular scale for both scales. However, these methods mainly focus on the consistency of predictions across scales, which indirectly improves the models with regularization. Moreover, they highly rely on the pseudo labels generated from the regular scale in the teacher network. The false negatives caused by inappropriate scales still remain in these methods.

Based on the above methods, which are equipped with an additional down-sampled view of unlabeled images, we propose to explicitly improve the quality of pseudo labels to handle the scale variation of objects. As shown in Figure 2 (c), we introduce a mixed-scale feature pyramid, which is built from the large-scale feature pyramid in the regular view and the small-scale feature pyramid in the down-sampled view. The mixed-scale feature pyramid is supposed to be capable of adaptively fusing features across scales, thus making better predictions in the teacher network. Furthermore, to avoid object instances missing in the pseudo labels due to low confidence scores, we propose to leverage the improvement of score as an indicator for mining pseudo labels from low confidence predictions. In summary, the main contributions are as follows:

- We propose a semi-supervised object detection framework MixTeacher, in which high-quality pseudo labels are generated from a mixed scale feature pyramid.
- We propose a method for pseudo labels mining, which leverages the improvement of predictions as the indicator to mining the promising pseudo labels.
- Our method achieves state-of-the-art performance on MS COCO and Pascal VOC benchmarks under various semi-supervised settings.

2. Related works

Semi-supervised Learning aims to train a model using a small amount of labeled data and a large amount of unlabeled data. Early studies mainly focused on image classification task [14, 32] and have gradually been generalized to various tasks [27, 33, 40]. Pseudo labeling [15] is one of the most popular paradigms in semi-supervised image classification, where labels of unlabeled data are generated by a pre-trained teacher network. Under this paradigm, Mean Teacher [35] proposes to maintain the teacher model as an exponential moving average (EMA) of the student model, thus generating pseudo labels end-to-end. Meanwhile, some works encourage models to make consistent predictions with perturbations [1, 26, 29]. Typical approaches such as MixMatch [2] and UDA [37] enforce consistent predictions across image views with different augmentations. Our work follows the pseudo labeling paradigm and consistency regularization, but focuses on object detection, where some task-specific challenges are relatively under-explored.

Semi-supervised Object Detection (SSOD) methods are mainly derived from semi-supervised image classification. As early attempts, CSD [13] encourages consistent predictions for horizontally flipped image pairs, whereas STAC [33] transfers the weak and strong augmentations from FixMatch [32] to SSOD. After that, [34, 38, 39, 42] simplify the trivial multi-stage training with the idea of EMA from Mean Teacher [35], realizing the end-to-end training. Considering the characteristic of object detection, some task-specific improvements are proposed. Soft Teacher [38] adopts separate and strict conditions to filter out high-quality pseudo labels for classification and regression and reduces the classification weights of negative proposals to suppress the influence of missing objects in pseudo labels. Unbiased Teacher [24] replaces the cross-entropy loss with Focal loss [22] to alleviate the numerous negative pseudo labels problem. On the basis of these work, our work focus on another challenge in SSOD, *i.e.* the large scale variation problem.

Scale Variation Challenge exists in most vision tasks. Since the scale of object instances in the detection task could vary in a wide range, numerous methods were proposed to detect objects across scales [20, 21, 31]. However,

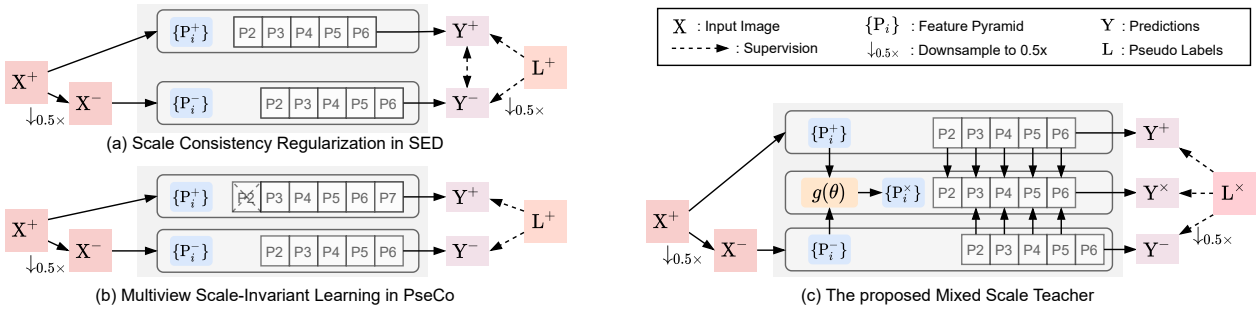


Figure 2. Comparison of multi-scale learning in semi-supervised object detection methods. Previous methods [10, 17] only focus on encouraging consistent predictions for the input image with different scales. The proposed MixTeacher explicitly introduces a mixed scale feature pyramid to adaptive fuse features from appropriate scales, which is capable to detect objects with varying sizes. The mixed scale features generate more accurate pseudo labels and help to mine promising labels, as a plug-in which can be discarded after training.

scale variation brings new challenges for semi-supervised object detection. For instance, objects with extreme scale tend to have low confidence, which makes them missed in the pseudo labels for most SSOD methods with a strict filter condition [33, 38]. Recent attempts have noticed the scale problem. [10] proposed to distill the predictions between a regular view and a down-sampled view to enhance the robustness of the model for scale variation. [17] also adopt a down-sampled view but shifts the layer of feature pyramid to reuse the same scale pseudo boxes as a regular view. These methods could be regarded as adding consistent regularization on varying scales. In contrast, our work proposes a mixed scale feature pyramid which can adaptive select scale for generating and mining promising pseudo labels.

3. Method

In the problem of semi-supervised object detection, a model is trained with a labeled set $\mathcal{D}^l = \{(\mathrm{X}_i^s, \mathrm{Y}_i^s)|_{i=1}^{N^l}\}$ and an unlabeled image set $\mathcal{D}^u = \{\mathrm{X}_i^u|_{i=1}^{N^u}\}$, where N^l and N^u are the numbers of labeled and unlabeled data. For each labeled image X_i^s, the annotation Y_i^s is composed of a set of boxes and corresponding category labels for the instances that appeared in the image. Built upon the pseudo labeling framework, our method follows a score filtering mechanism [38] to generate pseudo labels Y^u for unlabeled image X^u. An overview of our method is shown in Figure 3. We propose a mixed scale feature pyramid for pseudo labels generation and scale consistent learning, and based on it, a promising label mining strategy is introduced. Note that although the proposed method is independent of detection models, we adopt Faster RCNN [28] with FPN [21] as the default model, following most of the previous work.

3.1. Basic Pseudo-Labeling Framework

Following the common practice in previous work [17,24, 38], we adopt the pseudo-labeling under the teacher-student paradigm as our basic training framework. Specifically, the training images are sampled from both labeled and unlabeled datasets, and the overall objective is made up of these two parts to update a student model. Due to the lack of ground truth on unlabeled images, a teacher model provides pseudo labels for students, whose weights are updated by the exponential moving average of the student model.

In every training iteration, the training objective on labeled data follows a regular manner with fully supervised by the ground truth labels. On the unlabeled data, the teacher model first generates pseudo labels on a weakly augmented view of the image, which provides supervision signals for a strongly augmented view of the image for the student model. Subsequently, the student model is updated with the objective from labeled data and a strongly augmented view of the image with pseudo labels. The overall training objective can be formulated as:

$$\mathcal{L} = \mathcal{L}_s + \alpha\mathcal{L}_u \tag{1}$$

where $\mathcal{L}_s$ and $\mathcal{L}_u$ denote supervised loss of labeled images and unsupervised loss of unlabeled images respectively, α controls contribution of unsupervised loss.

3.2. Mixed Scale Teacher

Existing works have proved that incorporating an extra down-sampled view of the unlabeled image, and regularizing the network with consistency constraints on either feature level or label level can significantly improve the performance of semi-supervised object detection [10, 17]. Based on this observation, we also leverage the down-sampled view, but resort to building a more informative feature representation, which is more suitable for pseudo labels.

Given an image, most of detectors first extract multi-scale features P_i with decreasing spatial sizes which constitute a feature pyramid $\mathbb{P}$. In the case of FPN [21], the spatial sizes of adjacent levels in the feature pyramid always differ by $2\times$, which results in $P_2 - P_6$ layers[1] with spatial sizes from $1/2^2$ to $1/2^6$ w.r.t. the size of the input image.

[1]In the original implementation of Faster RCNN, P2-P6 is used, while in some detector such as FCOS and RetinaNet, P3-P7 is used.

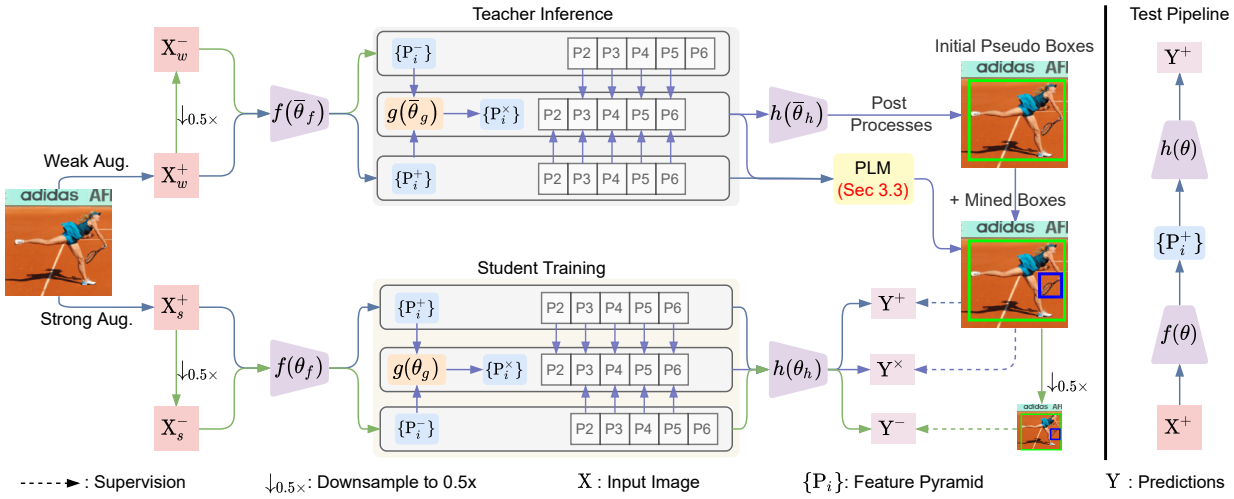


Figure 3. During training, the model first constructs two feature pyramids for a regular scale and a down-sampled scale with a feature extraction module $f(\theta_f)$, respectively. Next, an additional mixed scale feature pyramid is built by a feature fusion module $g(\theta_g)$. The student model trains on three scales with a shared detection head $h(\theta_h)$ taking the pseudo boxes generated from the mixed-scale of teacher model as supervisions. In addition, promising labels with low confidence scores are mined with a PLM strategy. The weights $\bar{\theta}$ in teacher are updated by EMA of the weights θ in student. In testing, the model with original architecture and regular input scale is used.

In this work, we first extract two feature pyramids from the regular view and the down-sampled view of the input image, respectively, which denote $\mathbb{P}^+ = \{P_2^+, ..., P_6^+\}$ and $\mathbb{P}^- = \{P_2^-, ..., P_6^-\}$, respectively. Then, we build a mixed scale feature pyramid $\mathbb{P}^\times = \{P_2^\times, ..., P_6^\times\}$ from the aforementioned two views to adaptive fuse appropriate features for input images containing objects of different scales.

Notice that with a $0.5\times$ down-sample ratio, the network produces a small scale feature pyramid. Specifically, the feature in P_i^+ shares the same spatial size as P_{i-1}^-. The spatial aligned feature maps of adjacent levels in two scales greatly simplify the design of feature fusion module. To avoid introducing a complex feature space, we propose a linear combination of features from the regular and down-sampled scales for the mixed scale feature pyramid, where the weight is formulated as follows:

$$\gamma = \sigma\left(g\left(P_i^+, P_{i-1}^-|\theta_g\right)\right), \tag{2}$$

where $\sigma(.)$ denotes the sigmoid activation with temperature T, and $g(.|\theta_g)$ is inspired by the SE block [12] which involves global average pooling followed by MLPs for channel concatenation of P_i^+ and P_{i-1}^-. Thus, the feature $P_i^\times$ in the mixed scale feature pyramid is a weighted sum of large scale and small scale features, which can be derived as:

$$P_i^\times = \gamma P_i^+ + (1-\gamma)P_{i-1}^-. \tag{3}$$

Note that the first level in the mixed-scale pyramid is a direct copy from regular view, for which the corresponding level in the down-sampled view does not exist.

As the mixed scale feature pyramid contains more information, we use it to generate pseudo labels for unlabeled images in the teacher network. Consequently, the training objective for unlabeled data becomes:

$$\begin{aligned}\mathcal{L}_u = &\mathcal{L}_{det}\left(h(\mathbb{P}^\times), \hat{y}^\times\right) + \mathcal{L}_{det}\left(h(\mathbb{P}^+), \hat{y}^\times\right) \\ &+\mathcal{L}_{det}\left(h(\mathbb{P}^-), \hat{y}^\times\right),\end{aligned} \tag{4}$$

in which the overall loss is the sum of losses from all scales. Similarly, the objective of the labeled part are also extended to multiple scales. For the specific form of $\mathcal{L}_{det}$ at each scale, we followed the baseline method [38].

It is worth mentioning that after extracting the regular and down-sampled scale features, the extra computation and parameters for building the mixed scale feature pyramid are negligible. Specifically, we implement the feature fusion module using two linear layers and a ReLU activation, which is a drop in the bucket compared to the millions of parameters in modern detection models.

Once the model is trained completely, all components not belong to the original detector can be discarded, including the steps used to extract small/mixed scale features and the feature fusion module. It satisfies the method to maintain the original detector architecture and input scale for inference, ensuring fairness for comparison.

3.3. Promising Labels Mining

In most of the previous methods, a high score threshold τ_h is applied to filter out the high-quality pseudo labels from the predictions of teacher network. However, it brings false negatives due to the inappropriate scales which will mislead the student network in semi-supervised learning.

To address this problem, we propose a strategy named Promising Labels Mining (PLM), which takes the predic-

tions from the mixed scale feature pyramid as the target, and measure the improvement from the regular scale. The improvement of confidence score is adopted as the indicator to mine pseudo labels from predictions with score $\tau_l < s \leq \tau_h$, which we called low-confidence candidates.

As a common practice in label assignment methods for fully supervised object detection [7–9, 19, 36], we first construct a bag of proposals from RPN for each low-confidence candidate, which will be used to measure the quality of each candidate. To reduce training costs, proposals that do not belong to any bag of low-confidence candidates or are not assigned to any high-confidence pseudo labels are not involved in the process of mining promising labels.

Notably, in order to avoid the influence introduced by strong augmentation, the proposals and feature pyramids from the teacher network are used to measure the score improvement. Besides, only the regular view is adopted as the source view, and the mixed scale is adopted as the target view to mine promising pseudo labels from candidates. The proposals generated from the source view are shared with the target view to align the input of the two views.

For a candidate with the highest score on class c, we define the indicator Δq as the promotion of the mean score from the source to the target view, formulated as:

$$\Delta q = \frac{1}{K}\left(\sum_{i=1}^{K} h_c(b_i, \mathbb{P}^{\times}|\bar{\theta}) - \sum_{i=1}^{K} h_c(b_i, \mathbb{P}^{+}|\bar{\theta})\right) \quad (5)$$

where K is the number of proposals in the bag for this candidate. $h_c(.|\bar{\theta})$ denotes the classification head in the teacher network, which takes the feature pyramid $\mathbb{P}$ and the RoI of each proposal b as input, and predicts the confidence score of class c for each proposal in the bag.

A low-confidence candidate with a higher Δq indicates the corresponding object instance prediction could reach greater promotion in the mixed scale feature pyramid. Therefore, it may be hard to detect in the source view but relatively easy in the target view with more information, suggesting that it is more likely to be a hard instance.

Candidates with Δq higher than the promotion threshold δ are regarded as pseudo labels. These mined pseudo labels, along with the existing high-confidence pseudo labels, are used in the classification loss for all three scales.

Discussion: Previous works [7, 9, 17] mainly utilize the idea of the bag of proposals for label assignment to reduce the proposals wrongly assigned to false positives or false negatives, which cannot address the missing pseudo labels with low confidence. In this work, we focus on mining low-confidence pseudo labels, and propose to measure their quality with two views. Therefore, our method is orthogonal to those label assignment methods. Besides, we only mined pseudo labels for the classification loss, while the criterion of filtering pseudo labels for the regression loss follows a widely used baseline [38], in which an uncertainty threshold is adopted for each box prediction. Other advanced strategies for filtering localization pseudo labels [4, 17] can also be integrated with our method, which might lead to better results but are irrelevant to our mixed scale framework, thus we do not discuss them in this paper.

4. Experimental Results

4.1. Dataset and Evaluation Protocol

We present experimental results on MS COCO [23] and Pascal VOC [6] benchmarks. Specifically, there are three common settings following previous works [10, 38]. The results are evaluated on `val2017` for the settings on COCO, and VOC17 `test` set for the setting on VOC, respectively. The specific data settings are as follows:

COCO Partially Labeled. It contains 4 splits with various labeled ratios, in which 1%, 2%, 5%, and 10% images from `train2017` are randomly sampled as labeled data, and the rest images are used as the unlabeled part for each split, respectively. Following the common practice, 5 different folds are used to evaluate each split, and the mean and standard deviation of results are reported.

COCO Additional. All 118k images in `train2017` are adopted as labeled data, and all of the additional 123k images in `unlabeled2017` are used as unlabeled data.

VOC Additional. VOC07 `trainval` is used as labeled data and VOC12 `trainval` is used as unlabeled data.

VOC Mixture. VOC07 `trainval` is used as labeled data and images from COCO containing 20 classes in VOC along with VOC12 `trainval` are taken as unlabeled data.

The evaluation metrics of this paper including AP at different IoU thresholds (*e.g.* $AP_{50:95}$ denoted as mAP, AP_{50}, AP_{75}) and different box scales (*e.g.* AP_s, AP_m, AP_l). All metrics are calculated via the COCO evaluation kit.

4.2. Implementation Details

Following the mainstream choice of the community, we adopt Faster-RCNN [28] with FPN [21] and ResNet-50 [11] as the detection model. The proposed method is implemented based on the well-known SoftTeacher [38], and we reuse its augmentation and training hyperparameters without any modification for a fair comparison. All models are trained on 8 GPUs with a base learning rate 0.01. Details of each setting are as follows: In COCO Partially Labeled setting, all models are trained for 180k steps with 1 labeled image and 4 unlabeled images per GPU in each step. The learning rate is decreased by 0.1 at 120k and 160k steps. In COCO Additional setting, the model is trained for 720k iterations with 4 labeled images and 4 unlabeled images per GPU in each step. The learning rate is decreased by 0.1 at 480k and 680k steps. In Pascal VOC settings, the model is trained for 40k steps with a constant learning rate with 2 labeled and 2 unlabeled images per GPU in each step.

	COCO Partially Labeled				COCO Additional
	1%	2%	5%	10%	100 %
Supervised Baseline	12.15±0.27	16.65±0.18	21.45±0.16	27.10±0.07	40.9
STAC [33]	13.97±0.35	18.25±0.25	24.38±0.12	28.64±0.21	39.5 $\xrightarrow{-0.3}$ 39.2
SED [10]	-	-	29.01	34.02	40.2 $\xrightarrow{+3.2}$ 43.4
Unbiased Teacher [24]*	20.75±0.12	24.30±0.07	28.27±0.11	31.50±0.10	40.2 $\xrightarrow{+1.1}$ 41.3
Soft Teacher [38]†	20.46±0.39	-	30.74±0.08	34.04±0.14	40.9 $\xrightarrow{+3.6}$ 44.5
LabelMatching [3]*	**25.81±0.28**	-	32.70±0.18	35.49±0.17	40.3 $\xrightarrow{+5.0}$ 45.3
PseCo [17]†	22.43±0.36	27.77±0.18	32.50±0.08	36.06±0.24	41.0 $\xrightarrow{+5.1}$ **46.1**
DTG-SSOD [16]†	21.27±0.12	26.84±0.25	31.90±0.08	35.92±0.26	40.9 $\xrightarrow{+4.8}$ 45.7
Unbiased Teacher v2 [25]*	25.40±0.36	28.37±0.03	31.85±0.09	35.08±0.02	40.9 $\xrightarrow{+3.9}$ 44.8
MixTeacher (Ours)†	25.16±0.26	**29.11±0.21**	**34.06±0.13**	**36.72±0.16**	40.9 $\xrightarrow{+4.8}$ 45.7

Table 1. Comparison with state-of-the-art methods on COCO benchmark. $AP_{50:95}$ on `val2017` set are reported. Under the Partially Labeled setting, results are the average of all five folds and numbers behind ± indicate the standard deviation. Under the Additional setting, numbers in front of the arrow indicate the supervised baseline. †: using labeled/unlabeled batch size 8/32, * indicates 32/32, and rest of the results using batch size 8/8.Bold indicates the best, while underline indicates the second best.

	1%	2%	5%	10%
Dense Teacher [41]	22.38	27.20	33.01	**37.13**
Unbiased Teacher v2 [25]	22.71	26.03	30.08	32.61
MixTeacher (Ours)	**23.83**	**27.88**	**33.42**	36.95

Table 2. Experimental results on the COCO Partially Labeled with FCOS [36]. PLM is not used for our method in this setting.

The weight of unlabeled loss α is set to 4.0 for the COCO Partially Labeled setting, and 2.0 for other settings considering the proportion of labeled and unlabeled images under different settings. Other hyper-parameters are the same for all settings, *i.e.* $T = 3.0$, $\tau_h = 0.9$, $\tau_l = 0.7$, $\delta = 0.1$.

4.3. Comparison with State-of-the-art

We compare the proposed MixTeacher with the supervised baseline and several recent methods on MS COCO and Pascal VOC benchmarks. For a fair comparison, we follow the experimental settings with Soft Teacher [38]. Without loss of generality, some known tricks which can improve the results in most cases are not used. For instance, Unbiased Teacher [24] and PseCo [17] adopt Focal Loss [22] to handle the class imbalance problem. Some methods [3, 24] adopt a larger batch size for unlabeled data. Besides, it is worth mention that PseCo [17] changes the output levels of FPN from P2-P6 to P3-P7 in testing.

MS COCO. We first evaluate the proposed method on MS COCO in Table 1. Under the Partially Labeled setting, the statistical results over five folds for four different labeled ratios are reported. Our method achieves more than 12% mAP improvements against the supervised baseline for the settings with less than 10% labeled data and demonstrates superiority in most of the labeled ratios. For the case of 2% and 5% labeling ratio, our method yields 29.11 and 34.06 mAP, which is around 1.5 mAP higher than the previous best method PseCo [17]. When the labeled data is extremely scarce, i.e., 1% labeling ratio, there are only tens labeled images for some tail categories. Our method without any class-specific process still reaches comparable performance to the previous best method LabelMatching [3], which estimates the class distribution and tunes thresholds adaptively. Under the Additional setting, our method is competitive to the state-of-the-art PseCo [17]. It demonstrates the effectiveness of our method even with adequate labeled data. Besides, we conduct an experiment to evaluate our method on the anchor-free detector FCOS [36] with ResNet-50 backbone. In this case, only the mixed scale teacher is used. The results in Table 2 show that the anchor-free detector can also benefit from the proposed mixed scale teacher.

	AP_{50}	$AP_{50:95}$
Supervised [24]	72.63	42.13
STAC [33]	77.45 (+4.82)	44.64 (+2.51)
Humble Teacher [34]	80.94 (+8.31)	53.04 (+10.91)
Rethinking Pse [18]	79.00 (+6.37)	54.60 (+12.47)
LabelMatching [3]	85.48 (+12.85)	55.11 (+12.98)
MixTeacher (Ours)	**85.85** (+13.22)	**56.25** (+14.12)

Table 3. Experimental results on the VOC Additional setting.

	AP_{50}	$AP_{50:95}$
Supervised [24]	72.63	42.13
STAC [33]	79.08 (+6.45)	46.01 (+3.88)
Humble Teacher [34]	81.29 (+8.66)	54.41 (+12.28)
Rethinking Pse [18]	79.60 (+6.79)	56.10 (+13.97)
LabelMatching† [3]	85.81 (+13.18)	55.50 (+13.37)
MixTeacher (Ours)	**86.58** (+13.95)	**56.83** (+14.70)

Table 4. Experimental results on the VOC Mixture setting.

Pascal VOC. We also evaluate our method under two VOC settings. The results are shown in Table 3 and Table 4. † denotes that we report the results for the method with their official implementation. The results demonstrate that

Feature Scales $\mathbb{P}^+$	$\mathbb{P}^-$	$\mathbb{P}^\times$	MST	PLM	mAP	AP_{50}	AP_{75}
					26.8	45.1	28.4
✓					33.9 (+7.1)	54.0	37.0
✓	✓				34.7 (+7.9)	54.7	37.8
✓	✓	✓			34.4 (+7.4)	54.2	37.2
✓	✓	✓	✓		36.2 (+9.4)	56.5	39.5
✓	✓	✓	✓	✓	**36.7 (+9.9)**	**57.0**	**39.7**

Table 5. Analysis of various components of proposed approach. MST indicates generating pseudo labels from the mixed scale feature pyramid. PLM indicates promising labels mining.

our method consistently reaches the best performance under two settings. Similar to the conclusion of COCO Additional setting, the results indicate the model can benefit from our method even if there is already sufficient labeled data.

4.4. Ablation Study

In this part, we conduct experiments under the COCO Partially Labeled setting to analyze and validate our method in detail. All the experiments in this section are conducted on a single data fold with 10% labeling ratio.

Effect of each component. We validate the effectiveness of each component step by step, and the results are shown in Table 5. The model starts from 26.8 mAP when using the labeled data only. After using the regular scale $\mathbb{P}^+$ of unlabeled data, the semi-supervised baseline reaches 33.9 mAP immediately. Furthermore, leveraging additional down-sampled scale features $\mathbb{P}^-$, and with supervision from the regular scale of teacher network, the model achieves another +0.8 mAP improvement. However, there are no gains of introducing a mixed scale features $\mathbb{P}^\times$, but still with the supervision from the regular scale makes the results worse. Instead, with the guidance of pseudo labels from mixed scale, the mAP boosts to 36.2. Finally, our method equips mixed scale pseudo labels and promising labels mining reaches the best performance, 36.7 mAP.

Comparison with other multi-view methods. Scale variation across object instances is a key challenge in semi-supervised object detection, and some pioneer works [10, 17] have introduced an additional down-sampled view to improve the model handle the scale issue. Table 6 shows the performance and the training time for each iteration of different multi-view methods integrated into the naive implement of Soft Teacher [38]. SCR indicates Scale Consistency Regularization in SED [10]. MSIL indicates Multi-view Scale-Invariant Learning in PseCo [17]. we report the results of MST‡ training with randomly dropping a path from large scale and mixed scale for the loss of student network, *i.e*. only the 0.5× scale and one of 1× scales are used to keep a comparable training time with other multi-scale methods (1.03 *vs*. 0.97 *vs*. 0.93 sec/iter). The results show that all methods with additional view improve the performance compared with the single view baseline. Our method presents a significant advantage of more than 1.3 mAP gains compared with previous consistency learning methods. Randomly dropping a path in training saves 0.19 second every iteration and reaches comparable results.

	mAP	AP_{50}	AP_{75}	sec/iter
Baseline	33.9	54.0	37.0	0.75
SCR [10]	34.6	54.6	37.9	0.97
MSIL [17]	34.9	55.1	37.6	**0.94**
MST‡ (Ours)	36.0	56.3	**39.5**	1.03
MST (Ours)	**36.2**	**56.5**	**39.5**	1.22

Table 6. Comparison with other multi-view methods.

	mAP	AP_{50}	AP_{75}
Baseline	34.7	54.7	37.8
CONV-ADD	35.1 (+0.4)	55.0	38.5
CAT-CONV	35.2 (+0.5)	55.2	38.3
GAP-MLP (Ours)	**36.2 (+1.5)**	**56.5**	**39.5**

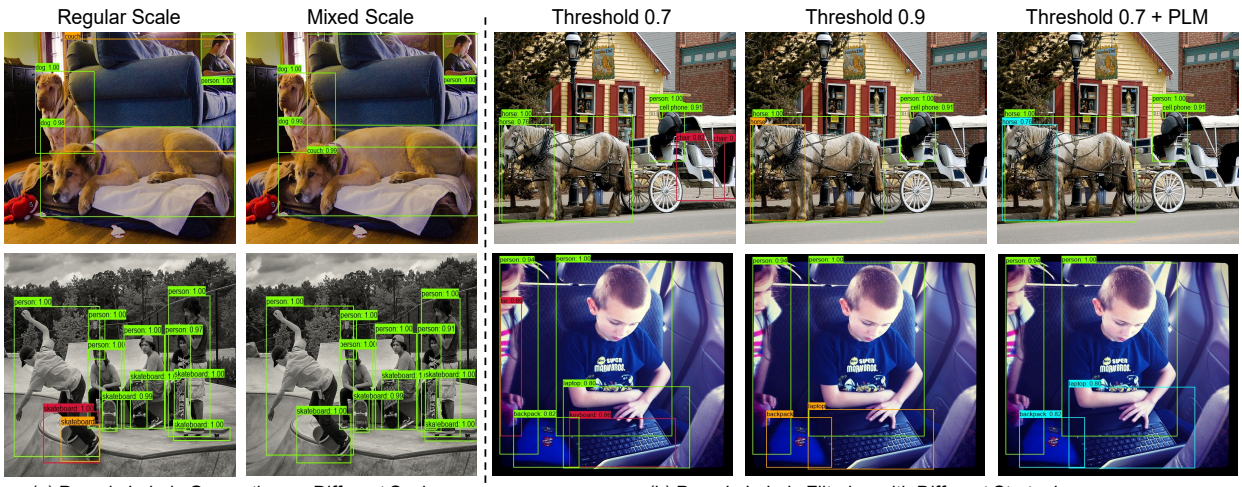
Table 7. Comparison of feature fusion approaches.

	mAP	AP_s	AP_m	AP_l	FPS
Baseline	36.5	21.8	39.2	48.6	33.4
Test on $\mathbb{P}^-$	33.2	14.6	36.1	50.0	**37.1**
Test on $\mathbb{P}^+$	36.7	**21.8**	39.2	48.6	33.4
Test on $\mathbb{P}^\times$	**37.5**	**21.8**	**40.1**	**51.1**	27.0

Table 8. Performance of different scales testing.

Comparison of feature fusion approaches. we compare the effects of different feature fusion approaches to build mixed scale features. We evaluate the performance of three simple fusion architectures, *i.e*. "CONV-ADD" denotes that employ two 3×3 convolution layer to align features for regular scale and down-sampled scale, followed with an element-wise addition, "CAT-CONV" denotes that concatenated by channel and then apply convolution to reduce channels. As shown in Table 7, compared with the baseline without fusion features, all three fusion methods obtain gains in mAP. Among them, our method that builds the mixed scale feature as a weighted summation of regular and down-sampled scales achieves the best performance.

Performance of different scales testing. In the proposed method, each feature pyramid in three scales is capable to detect objects. The inference results with different feature scales are useful to guide the design of the strategy for pseudo label generation, and can also be used as a separate detector. Table 8 shows the performance of a model with testing on different feature scales. We also report the average inference speed on a single V100 GPU. Notice that the pipeline and architecture are exactly the same as the vanilla faster R-CNN when tested on the regular or down-sampled scale. However, additional computation and parameters are required to build the feature pyramid when tested on mixed



(a) Pseudo Labels Generation on Different Scales (b) Pseudo Labels Filtering with Different Strategies

Figure 4. Qualitative visualization for the components in MixTeacher. (a) Comparison of pseudo labels generated from the regular scale and the mixed scale feature pyramids. (b) Comparison of pseudo labels under different score thresholds and our promising label mining results. The green boxes denote True Positives. The red boxes highlight the False Positives, and orange boxes denote the False Negatives. Besides, the mined labels are highlight with a cyan box.

T	0.1	0.3	1.0	3.0	10.0
mAP	35.9	35.9	36.1	**36.2**	36.1
AP_{50}	56.4	56.3	56.3	**56.5**	**56.5**
AP_{75}	39.3	39.2	39.4	**39.5**	39.2

(a) Study on temperature T.

τ_l	mAP	AP_{50}	AP_{75}
0.6	33.8	54.2	37.1
0.7	**36.7**	**57.0**	**39.8**
0.8	36.3	56.8	39.2

(b) Study on lower threshold τ_l

δ	AvgBox	mAP	AP_{50}	AP_{75}
0.0	3.87	36.2	56.5	39.5
0.1	4.15	**36.7**	**57.0**	**39.8**
0.2	3.91	36.2	56.7	39.5

(c) Study on promotion threshold δ.

Table 9. Comparison of different hyper-parameters for the proposed MixTeacher.

scale. The results show that the down-sampled scale makes the worst mAP and falls behind the regular scale one on AP_s especially, but it reaches the best AP_l for large objects. It suggests that previous methods [10, 17] generating pseudo labels from the regular scale is not appropriate for large objects. On the other hand, the mixed scale which adaptive select scales achieves competitive results on all objects and it is more suitable for pseudo labels generation.

Choice of hyper-parameters. In order to analyze the sensitivity of some key hyper-parameters, we investigate the influence of different temperature T in building mixed scale feature pyramid, the lower score threshold τ for promising labels mining, and the score promotion threshold δ between scales. As shown in Table 9, the temperature T is set to 3.0 for better results, which encourages the feature space of mixed scale to be more likely to one of the source scales. The lower threshold τ_l is set to 0.7, which controls the number of candidates for mining. The promotion threshold δ is set to 0.1, which mines 0.28 boxes for a image in average and brings 0.5 gains in mAP.

4.5. Qualitative Visualization

We show qualitative results to demonstrate the quality of pseudo labels more intuitively. As shown in Figure 4 (a), the pseudo labels generated from the proposed mixed scale features are more accurate than the regular scale. Figure 4 (b) shows that there are many false positives when threshold=0.7 and false negatives when threshold=0.9, our promising label mining method alleviates these problems.

5. Conclusion and Limitation

In this work, we delve into the scale variation problem in semi-supervised object detection, and propose a novel framework by introducing a mixed scale teacher to improve the pseudo labels generation and scale invariant learning. In addition, benefiting from better predictions from mixed scale features, we propose to mine pseudo labels with the score promotion of predictions across scales. Extensive experiments on MS COCO and Pascal VOC benchmarks under various semi-supervised settings demonstrate that our method achieves new state-of-the-art performance. While we have shown the superiority of MixTeacher, the method is built on an old-fashion detector with the simplest FPN and naive label assignment strategy. Whether the scale variation problem in SSOD can be addressed with more advanced FPN architectures or label assignment methods is unclear, which is an interesting future work.

References

[1] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27, 2014. 2

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2

[3] Binbin Chen, Weijie Chen, Shicai Yang, Yunyi Xuan, Jie Song, Di Xie, Shiliang Pu, Mingli Song, and Yueting Zhuang. Label matching semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14381–14390, 2022. 6

[4] Changrui Chen, Kurt Debattista, and Jungong Han. Semi-supervised object detection via virtual category learning. In *European Conference on Computer Vision*, 2022. 5

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 1

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, June 2010. 1, 5

[7] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3490–3499. IEEE Computer Society, 2021. 5

[8] Ziteng Gao, Limin Wang, and Gangshan Wu. Mutual supervision for dense object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3641–3650, 2021. 5

[9] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 303–312, 2021. 5

[10] Qiushan Guo, Yao Mu, Jianyu Chen, Tianqi Wang, Yizhou Yu, and Ping Luo. Scale-equivalent distillation for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 3, 5, 6, 7, 8

[11] 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. 5

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 4

[13] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. *Advances in neural information processing systems*, 32, 2019. 1, 2

[14] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017. 1, 2

[15] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896, 2013. 2

[16] Gang Li, Xiang Li, Yujie Wang, Yichao Wu, Ding Liang, and Shanshan Zhang. Dtg-ssod: Dense teacher guidance for semi-supervised object detection. In *Advances in neural information processing systems*, 2022. 6

[17] Gang Li, Xiang Li, Yujie Wang, Yichao Wu, Ding Liang, and Shanshan Zhang. PseCo: Pseudo Labeling and Consistency Training for Semi-Supervised Object Detection. In *European Conference on Computer Vision*, 2022. 2, 3, 5, 6, 7, 8

[18] Hengduo Li, Zuxuan Wu, Abhinav Shrivastava, and Larry S Davis. Rethinking pseudo labels for semi-supervised object detection. *arXiv preprint arXiv:2106.00168*, 2021. 6

[19] Shuai Li, Chenhang He, Ruihuang Li, and Lei Zhang. A dual weighting label assignment scheme for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9387–9396, 2022. 5

[20] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware trident networks for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6054–6063, 2019. 2

[21] 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. 2, 3, 5

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2, 6

[23] 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. 1, 5

[24] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*, 2021. 1, 2, 3, 6

[25] Yen-Cheng Liu, Chih-Yao Ma, and Zsolt Kira. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9819–9828, June 2022. 6

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

[27] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, pages 12674–12684, 2020. 2

[28] 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, 2015. 3, 5

[29] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29, 2016. 2

[30] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1

[31] Bharat Singh and Larry S Davis. An analysis of scale invariance in object detection snip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3578–3587, 2018. 2

[32] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020. 1, 2

[33] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 1, 2, 3, 6

[34] 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. 2, 6

[35] 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*, 30, 2017. 1, 2

[36] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 5, 6

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

[38] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3060–3069, 2021. 2, 3, 4, 5, 6, 7

[39] 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. 2

[40] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Sess: Self-ensembling semi-supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11079–11087, 2020. 2

[41] Hongyu Zhou, Zheng Ge, Songtao Liu, Weixin Mao, Zeming Li, Haiyan Yu, and Jian Sun. Dense teacher: Dense pseudo-labels for semi-supervised object detection. In *European Conference on Computer Vision*, page 35–50, Berlin, Heidelberg, 2022. Springer-Verlag. 6

[42] 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. 2