Label Matching Semi-Supervised Object Detection

Binbin Chen[2], Weijie Chen[1,2,†], Shicai Yang[2], Yunyi Xuan[2], Jie Song[1]
Di Xie[2], Shiliang Pu[1], Mingli Song[1], Yueting Zhuang[1,†]

[1]Zhejiang University, [2]Hikvision Research Institute

{chenbinbin8,chenweijie5,yangshicai,xuanyunyi,xiedi,pushiliang.hri}@hikvision.com

{sjie,songml,yzhuang}@zju.edu.cn

Abstract

Semi-supervised object detection has made significant progress with the development of mean teacher driven self-training. Despite the promising results, the label mismatch problem is not yet fully explored in the previous works, leading to severe confirmation bias during self-training. In this paper, we delve into this problem and propose a simple yet effective LabelMatch framework from two different yet complementary perspectives, i.e., distribution-level and instance-level. For the former one, it is reasonable to approximate the class distribution of the unlabeled data from that of the labeled data according to Monte Carlo Sampling. Guided by this weakly supervision cue, we introduce a re-distribution mean teacher, which leverages adaptive label-distribution-aware confidence thresholds to generate unbiased pseudo labels to drive student learning. For the latter one, there exists an overlooked label assignment ambiguity problem across teacher-student models. To remedy this issue, we present a novel label assignment mechanism for self-training framework, namely proposal self-assignment, which injects the proposals from student into teacher and generates accurate pseudo labels to match each proposal in the student model accordingly. Experiments on both MS-COCO and PASCAL-VOC datasets demonstrate the considerable superiority of our proposed framework to other state-of-the-arts. Code will be available at https://github.com/HIK-LAB/SSOD.

1. Introduction

Supervised learning has advanced object detection in the past few years, benefited from tremendous labeled training data [5, 17, 26, 28, 34]. However, it is extremely expensive and time-consuming to collect accurate annotations. As an alternative, semi-supervised object detection (SSOD) is proposed to use a small amount of labeled data in conjunc-
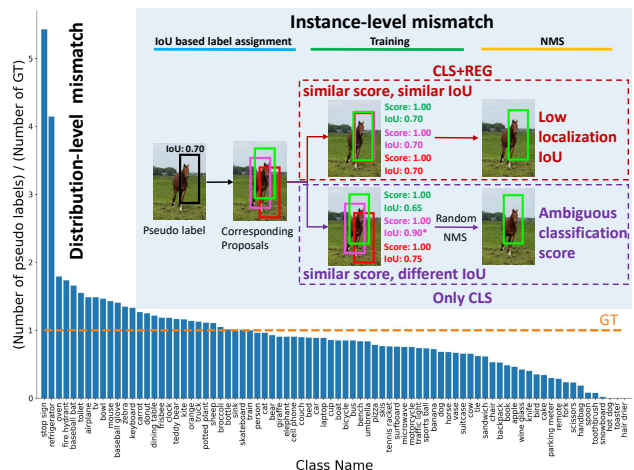

Figure 1. Label mismatch problems on the MS-COCO dataset. 1) **Distribution-level mismatch**: there exists a bias between the pseudo labels produced by the single confidence threshold and the ground truth labels (GT) during self-training, as shown in the relation of the blue bar and the orange dotted line. 2) **Instance-level mismatch**: there are two kinds of training patterns for the unlabeled data in the previous SSOD frameworks. One is the same as supervised learning, using both classification and box regression for optimization, which will overfit the poor-quality pseudo labels and result in low localization accuracy. To avoid incorrect box regression, another one merely exploits a classification objective [22], which will bring ambiguity due to the similar classification scores to confuse the post-processing of Non-Maximum-Suppression (NMS).

tion with a large amount of unlabeled data to optimize the detectors [14, 22, 31, 37, 40]. Recently, SSOD has achieved growing interest in the object detection community.

Self-training has been proven useful in SSOD, especially mean teacher framework [22, 37], which annotates the unlabeled data by a gradually evolving teacher and guides the learning of a student in a mutually beneficial manner. As the key process of mean teacher, the existing pseudo labeling methods [22, 37, 40] simply utilize a hand-crafted confidence threshold to filter out low-quality pseudo labels and

---

†Corresponding author

directly treat the remaining ones as reliable pseudo labels. However, it is inevitable to encounter the label mismatch problem, leading to severe confirmation bias [1] during self-training. In this paper, we delve into this problem from two perspectives, *i.e.*, *distribution-level* and *instance-level*.

From the perspective of the distribution-level label mismatch problem, it is extremely difficult to generate unbiased pseudo labels to match the ground-truth labels with consistent class distribution by using a single and fixed confidence threshold due to the class-imbalanced data distribution. As shown in Fig. 1, the number of pseudo labels is much higher than the ground-truth labels in some classes, while far less in some other classes, resulting in abundant false positives and false negatives. From the perspective of the instance-level label mismatch problem, the existing methods directly follow supervised object detection [28] for label assignment. However, the situation is totally different in semi-supervised learning since the quality of pseudo label cannot be guaranteed, leading to label assignment ambiguity problem as illustrated in Fig. 1. Especially in mean teacher driven self-training framework, it is crucial to study how to assign the pseudo labels generated by the mean teacher to the proposals generated by the student network rather than a rough IoU based label assignment manner [28]. Based on the aforementioned challenges of label mismatch in two different yet complementary granularities, we begin our study and develop a *LabelMatch* framework.

To address the first challenge, we present a very simple *re-distribution mean teacher*. Assumed that the labeled data is selected from the entire data gallery via Monte Carlo Sampling. In this way, the label distribution of the unlabeled data can be approximated from that of the labeled data. In fact, we have evaluated the label distribution of the labeled and unlabeled data in several popular SSOD datasets, and they all meet this hypothesis which can be exploited as a weakly supervision cue for pseudo labeling. Under this inspiration, in contrast to a single and fixed confidence threshold, we utilize *adaptive label-distribution-aware confidence thresholds* (ACT) to generate unbiased pseudo labels for the unlabeled data, supervised by the label distribution in the labeled data. The ACT are category-specific and adaptively up-to-date during self-training.

To address the second challenge, we propose a novel *proposal self-assignment* method. Before introducing our method, we should highlight that it is infeasible to set all pseudo labels as hard labels due to the poor quality of the pseudo labels, especially at the beginning of self-training. Under this consideration, we divide the pseudo labels into reliable ones and uncertain ones according to the confidence score. We treat the reliable labels as hard labels for model optimization identically to the supervised manner, while exploiting the uncertain ones via the proposal self-assignment method for soft learning. Detailedly, the proposals from the student are injected into the teacher for proposals correction, which can provide corresponding soft labels to rectify each proposal accordingly. Besides, to encourage positive feedback during self-training, we introduce a *reliable pseudo label mining* (RPLM) strategy to further improve the performance, which aims to convert the high-quality uncertain pseudo labels into reliable ones in a curriculum way.

We benchmark LabelMatch with the same experimental settings to Unbiased-Teacher [22] using the MS-COCO [20] and PASCAL-VOC [7] datasets, namely *COCO-standard*, *COCO-additional*, and *VOC*. LabelMatch achieves new state-of-the-art results across all benchmarks. Especially in the settings with scarce labeled data, *i.e.*, *COCO-standard* with only 1% labeled data and *VOC*, our method can surpass the previous state-of-the-arts by a large margin.

The contributions of this paper are listed as follows:

- We contribute to analyzing the label mismatch problem from the perspectives of distribution-level and instance-level, which provides a brand-new direction for SSOD.

- We propose a simple yet effective LabelMatch framework to address the label mismatch problems in SSOD. In this framework, we 1) present a re-distribution mean teacher to address the distribution-level label mismatch problem; 2) design a proposal self-assignment scheme to address the instance-level label mismatch problem; 3) introduce a reliable pseudo label mining strategy for pseudo label re-calibration during self-training.

- The LabelMatch framework achieves new state-of-the-arts on many popular SSOD benchmarks. Also, we build a MMDetection-based semi-supervised object detection codebase for the fair study of SSOD algorithms.

2. Related Work

Semi-Supervised Classification. The general methods can be roughly categorized into two types. One is consistency regularization, assuming the model's predictions to be invariant even if various perturbations are applied. There are different kinds of perturbations, including model-level perturbations [13, 29, 33], image augmentations [35], and adversarial training [23]. Another one is self-training, aka pseudo labeling, which regards the predictions as pseudo labels. For instance, NoisyStudent [36] evolves the pseudo labels for model optimization iteratively. MixMatch [2] uses mixup augmentation and averages different augmented predictions to generate pseudo labels. FixMatch [30] uses the weakly augmented data for pseudo labeling while exploiting the strongly augmented data for model training.

Semi-Supervised Object Detection. The technologies in SSOD are inherited from semi-supervised classification, dividing into consistency regularization [14, 32] and self-training [22, 31, 37, 38, 40]. In this paper, we mainly focus
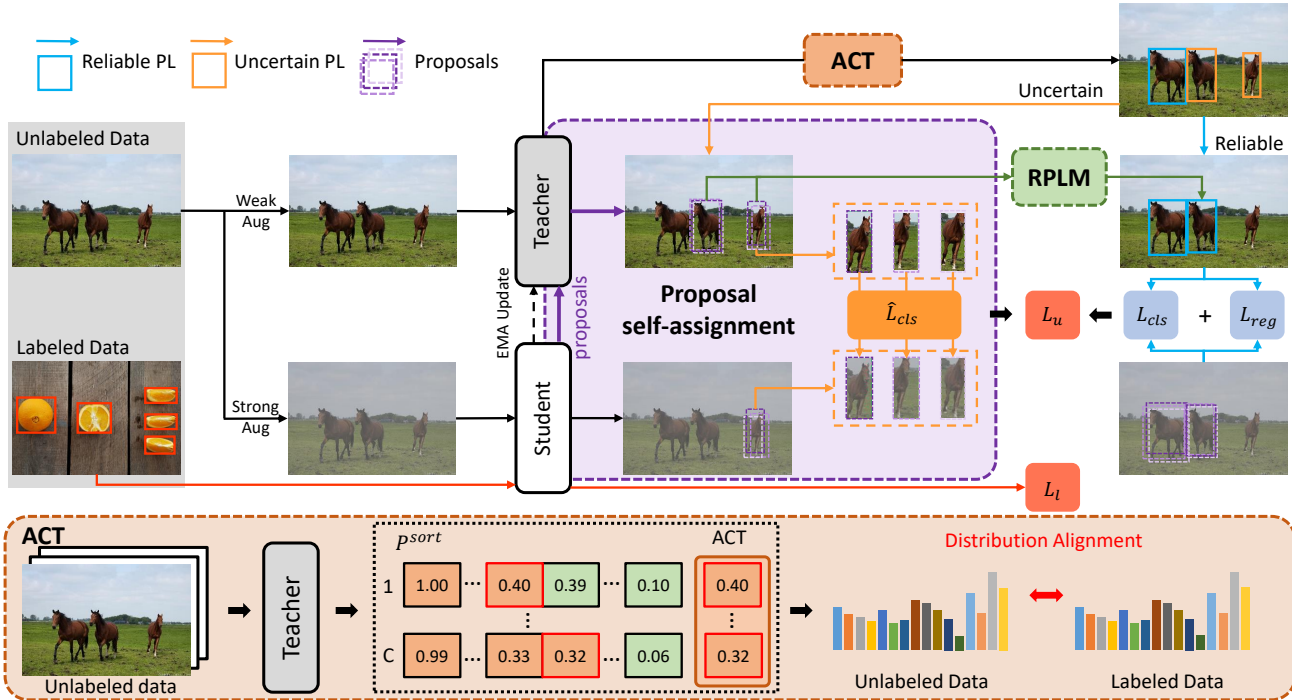
Figure 2. An overview of LabelMatch framework. Labeled data: only applied to the student with a supervised loss. Unlabeled data: annotated by the teacher to get pseudo labels (PL) according to the *adaptive label-distribution-aware confidence thresholds* (ACT), which are then split into reliable ones and uncertain ones for separated optimization. Reliable pseudo labels directly follow the IoU based assignment strategy, acting as hard labels to train the student model. As for uncertain labels, the *proposal self-assignment* method guides the student training with the supervision provided by the corresponding proposal prediction in the teacher. Besides, a *reliable pseudo label mining* (RPLM) strategy is utilized to convert the high-quality uncertain pseudo labels into reliable ones as the training goes on.

on the latter one. STAC [31] first generates pseudo labels by the pre-trained model and then feeds them back into the network with strong augmentation for model fine-tuning. To simplify this offline pseudo labeling, mean teacher based methods [22, 37, 40] perform a weak data transformation for online pseudo labeling and a strong data transformation for model training. However, there lies both foreground-background imbalance and foreground classes imbalance in SSOD, which makes it a more challenging task than semi-supervised classification. Unbiased-Teacher [22] and Soft-Teacher [37] use the focal-loss and the soft-weight to alleviate these problems, respectively. Despite the great progress, the label mismatch problem during pseudo labeling still exists in the previous works. In contrast, we propose a Label-Match framework to solve this problem from perspectives of distribution-level and instance-level.

Label Assignment. It is necessary to assign the target of classification and localization for each proposal or anchor in object detection, known as label assignment [8], which can be categorized into fixed and dynamic variants [8]. IoU based and center based label assignment are two common fixed assigning strategies, the first of which shows effectiveness in both RCNN-series [6, 9, 10, 24, 28] and one-stage detectors [19, 21, 26, 27], while the second one is popular

in many anchor-free object detection [16, 25, 34]. Recently, many adaptive mechanisms have been proposed to promote the label assignment, such as ATSS [39], PAA [15], AutoAssign [41], OTA [8], etc. However, all of these methods are only applied in supervised object detection, leaving a blank in SSOD due to the complex situation. To cope with the instance-level label mismatch problem, a novel label assignment is proposed to facilitate self-training in this paper.

3. Methodology

In SSOD, a set of labeled images $D_l=\{x_i^l, y_i^l\}_{i=1}^{N_l}$ and a set of unlabeled images $D_u=\{x_i^u\}_{i=1}^{N_u}$ are provided, where N_l and N_u represent the number of labeled and unlabeled data, respectively. The annotation y_i^l contains both categories and bounding boxes information.

3.1. Overview

The pipeline of LabelMatch framework is illustrated in Fig. 2, which is derived from a basic mean teacher framework. The main idea of the mean teacher framework is to drive the teacher and student to evolve in a mutual learning mechanism. However, previous mean teacher based works inevitably suffer from the label mismatch problems, which we divide into two granularities, including distribution-level

and instance-level. To solve these problems, we modify the mean teacher framework and develop a LabelMatch framework, consisting of a re-distribution mean teacher to solve the distribution-level label mismatch problem and a proposal self-assignment method to deal with the instance-level label mismatch problem. Also, it is beneficial to explore more high-quality pseudo labels. Thus, we further equip the proposed LabelMatch with a reliable pseudo label mining strategy to improve performance.

3.2. Preliminary: Mean Teacher Framework

Our approach follows the regimen of mean teacher, which contains a teacher model for pseudo label generation and a student model to improve the teacher model by updating knowledge. Both labeled and unlabeled data jointly constitute the batch of data. In each iteration, the teacher model first generates pseudo labels on the weakly-augmented unlabeled data, which are served as supervision signals for the corresponding strongly-augmented version. Subsequently, the student model is trained on the labeled data and the strongly-augmented unlabeled data with pseudo labels. In this way, the final training objective consists of a supervised loss and an unsupervised loss:

$$\mathcal{L}_l = \sum_i \mathcal{L}_{cls}(x_i^l, y_i^l) + \mathcal{L}_{reg}(x_i^l, y_i^l), \tag{1}$$

$$\mathcal{L}_u = \sum_i \mathcal{L}_{cls}(x_i^u, y_i^u) + \mathcal{L}_{reg}(x_i^u, y_i^u), \tag{2}$$

where $\mathcal{L}_{cls}$ is the classification loss, $\mathcal{L}_{reg}$ is the box regression loss, and y_i^u is the pseudo label annotated by the teacher model. The overall loss is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_l + \lambda \mathcal{L}_u, \tag{3}$$

where λ is a weight to balance the unsupervised loss, which is set 2.0 by default in this paper. During self-training, the teacher gradually updates its weights from the student via an exponential moving average (EMA) strategy.

3.3. LabelMatch

We claim that the main obstacle to hinder the performance of mean teacher framework lies in the label mismatch problem. The proposed LabelMatch framework adopts the same mean teacher scheme, but develops a re-distribution mean teacher, which utilizes adaptive label distribution-aware confidence thresholds (ACT) to achieve unbiased pseudo labels. Moreover, a proposal self-assignment method and a reliable pseudo label mining strategy are introduced to rectify self-training.

Re-distribution Mean Teacher. In semi-supervised learning, the labeled data and the unlabeled data are from the same data distribution. Intuitively, we can obtain adaptive thresholds by minimizing the discrepancy of class distributions between the labeled and the unlabeled data, which can

be formulated as follows:

$$\underset{t_1,\ldots,t_C}{\mathrm{argmin}} \quad D_{KL}([\underbrace{r_1^l,\ldots,r_C^l}_{f-f}, \underbrace{r_f^l}_{f-b}], [\underbrace{r_1^u,\ldots,r_C^u}_{f-f}, \underbrace{r_f^u}_{f-b}])$$

$$\text{s.t.} \quad r_c^l = \frac{n_c^l}{\sum_{i=1}^C n_i^l},$$

$$r_f^l = \frac{\sum_{i=1}^C n_i^l}{N_l},$$

$$r_c^u = \frac{\sum_{j=1}^{N_u} \sum(P_c^j > t_c)}{\sum_{i=1}^C (\sum_{j=1}^{N_u} \sum(P_i^j > t_i))}, \tag{4}$$

$$r_f^u = \frac{\sum_{i=1}^C (\sum_{j=1}^{N_u} \sum(P_i^j > t_i))}{N_u},$$

where D_{KL} represents the Kullback-Leibler divergence between two distributions, n_i^l denotes the box number of the i-th class in the labeled data, C is the entire foreground class number, P_i^j is a list of prediction scores of the i-th class in the j-th unlabeled image, f-f means the foreground-foreground class distribution, and f-b means the foreground-background ratio. Note that all the predictions in the unlabeled data are post-processed by NMS. t_c is the optimized variant, aka the confidence threshold to filter pseudo boxes for the c-th category, determined as:

$$t_c = P_c^{sort}[n_c^l \cdot \frac{N_u}{N_l}], \tag{5}$$

where P_c^{sort} is a list of prediction scores of the c-th class, which are sorted by descending. For efficient implementation, only a subset of unlabeled data are selected to estimate the distribution for thresholds determination. While the model is consecutively optimized during training, the previous thresholds are however imprecise for pseudo labeling, failing to be consistent with the truth class distribution. We thus simply update these thresholds every K iterations to dynamically adjust to the current teacher model. Such that, the thresholds are category-specific and adaptively up-to-date, termed as adaptive label-distribution-aware confidence thresholds (ACT), which we identify as the critical step to solve the distribution-level mismatch problem.

Proposal Self-Assignment. It is worth noting that the quality of pseudo labels cannot be guaranteed, especially at the early beginning of self-training. Inspired by the noise label learning [4, 11], we divide pseudo labels into reliable ones and uncertain ones according to the confidence score. Denoting $\alpha\%$ as the pre-defined proportion of reliable pseudo labels, the confidence thresholds t_c^r to filter reliable pseudo labels for the c-th category can be written as:

$$t_c^r = P_c^{sort}[\alpha\% \cdot n_c^l \cdot \frac{N_u}{N_l}], \tag{6}$$

Pseudo labels with confidence higher than t_c^r are regarded as hard labels for student model optimization in a supervised manner. In contrast, the remaining uncertain ones are treated as soft labels for soft learning.

Obviously, the uncertain pseudo labels potentially lead to low localization accuracy. To avoid poor box regression, Unbiased-Teacher removes the box regression loss for the unlabeled data, yet resulting in ambiguity in label assignment as shown in Fig. 1. For example, supposed that the proposals with an IoU overlap higher than 0.5 are optimized to the same uncertain pseudo labels, they will tend to be the same in classification score but different in localization after being refined by ROIHead. These refined proposals behave indistinguishably for the NMS post-processing, which confuses NMS to suppress redundant boxes randomly. We refer to this situation as an instance-level label mismatch problem, lacking attention in the previous SSOD works. To this end, we present a novel proposal self-assignment method for proposal re-calibration. Specifically, we utilize the proposals matched to the uncertain pseudo labels generated by the student to extract the corresponding features in the teacher, and then feed these features into the ROIHead of the teacher to achieve the refined boxes. Different from the IoU based label assignment, each proposal in the student uses the corresponding soft labels refined by the ROIHead of the teacher model for self-training, and finally, varying from each other in classification score to avoid NMS confusion. In this way, we optimize the student model with the uncertain pseudo labels via a soft classification loss:

$$\hat{\mathcal{L}}_{cls} = \sum_{i=1}^{n_p} \sum_{c=1}^{C} -p_{i,c}^t \log p_{i,c}^s, \qquad (7)$$

where n_p is the number of the corresponding proposals matched to the uncertain pseudo labels, C is the class number, $p_{i,c}^s$ is the probability of the c-th class in the i-th proposal from the student model, and $p_{i,c}^t$ denotes the corresponding soft label from the teacher model matched to $p_{i,c}^s$.

Incorporating the re-distribution mean teacher and the proposal self-assignment for pseudo labeling, the unsupervised loss in Eq. (2) can be reformulated as:

$$\mathcal{L}_u = \sum_i \mathcal{L}_{cls}(x_i^u, y_i^{ur}) + \mathcal{L}_{reg}(x_i^u, y_i^{ur}) + \hat{\mathcal{L}}_{cls}(x_i^u, y_i^{uu}), \quad (8)$$

where y_i^{ur} and y_i^{uu} denote the reliable pseudo labels and the uncertain pseudo labels, respectively.

Reliable Pseudo Label Mining. To benefit from the continuously evolved teacher model and encourage the cycle positive feedback during self-training, we present a reliable pseudo label mining strategy to convert the high-quality uncertain pseudo labels into reliable ones. First, it is known that a set of adjacent boxes will be suppressed into one box after NMS. In this paper, we claim that these adjacent boxes before NMS can be exploited to evaluate the quality of the corresponding pseudo label after NMS. In this way, we present two evaluation metrics in this paper, *i.e.*, *mean score* and *mean IoU*, which are the average classification scores and IoU of this set of adjacent boxes matched to the corresponding bounding box after NMS. Note that We use the predictions from the teacher to compute *mean score* and
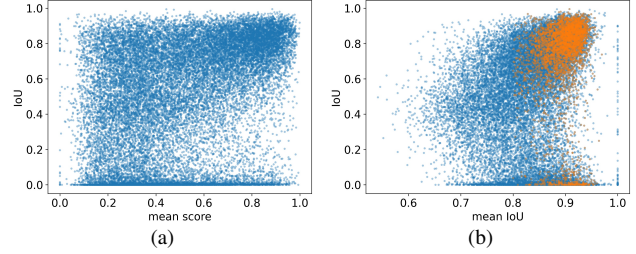

Figure 3. 5k images are selected to estimate the quality of pseudo labels. (a) the correlation between the IoU with ground truth and *mean score*. (b) the correlation between the IoU with ground truth and *mean IoU*. The orange points represent the predictions with the *mean score* larger than 0.8 and *mean IoU* larger than 0.8

mean IoU, and the IoU scores are determined by the IoU between the suppressed boxes and the selected box in NMS. We claim that the pseudo labels with higher quality usually correspond to higher mean scores and higher mean IoU. The empirical study in Fig. 3 gives a demonstration of our hypothesis. In this way, the uncertain pseudo labels with mean scores larger than T_{score} and mean IoU larger than T_{iou} will be transferred to reliable pseudo labels.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate our method on the MS-COCO [20] and PASCAL-VOC [7] datasets. There are three settings following the existing works [22,31]: (1) COCO-standard: 1%, 5%, 10% images of `train2017` set are sampled as the labeled training data and the remaining ones as the unlabeled data. (2) COCO-additional: we use the entire `train2017` set as the labeled data and the additional `COCO2017-unlabeled` set as the unlabeled data. (3) VOC: we use VOC07 `trainval` set as the labeled data and the VOC12 `trainval` set as the unlabeled data. The validation sets in the COCO setting and VOC setting are COCO `val2017` and VOC07 test set, respectively.

Network. For a fair comparison, we use Faster-RCNN [28] with FPN [18] and ResNet-50 backbone [12] as the detector. Our framework can be easily extended to other detectors.

Implementation Details. We implement our method with MMDetection [3]. For data augmentation, we apply random horizontal flipping and multi-scale for weak augmentation. Based on this augmentation, we then add random color jittering, grayscale, gaussian blurring and cutout patches for strong augmentation, which is similar to [22]. The T_{score} and T_{iou} in RPLM are set to 0.8 by default. More training and implementation details are introduced in the Appendix.

4.2. Results

COCO-standard. We evaluate the proposed method on COCO-standard (Tab. 1). Our method consistently outper-

	Threshold	1%	5%	10%
Supervised [22]	-	9.05 ± 0.16	18.47 ± 0.22	23.86 ± 0.81
STAC [31]	0.9	13.97 ± 0.35 (+4.92)	24.38 ± 0.12 (+5.91)	28.64 ± 0.21 (+4.78)
ISMT [38]	0.9	18.88 ± 0.74 (+9.83)	26.37 ± 0.24 (+7.90)	30.53 ± 0.52 (+6.67)
Instant Teaching [40]	0.9	18.05 ± 0.15 (+9.00)	26.75 ± 0.05 (+8.28)	30.40 ± 0.05 (+6.54)
Unbiased Teacher [22]	0.7	20.75 ± 0.12 (+11.70)	28.27 ± 0.11 (+9.80)	31.50 ± 0.10 (+7.64)
Soft Teacher [37]	0.9	20.46 ± 0.39 (+11.41)	30.74 ± 0.08 (+12.27)	34.04 ± 0.14 (+10.18)
LabelMatch (Ours)	ACT	$\mathbf{25.81 \pm 0.28}$ (+16.76)	$\mathbf{32.70 \pm 0.18}$ (+14.23)	$\mathbf{35.49 \pm 0.17}$ (+11.63)

Table 1. Experimental results on COCO-standard ($AP_{50:95}$). All the results are the average of all 5 folds.

	Iterations	$AP_{50:95}$
STAC [31]	540k	$39.5 \xrightarrow{-0.3} 39.2$
Unbiased Teacher [22]	270k	$40.2 \xrightarrow{+1.1} 41.3$
Soft Teacher [37]	370k	$40.9 \xrightarrow{+3.6} 44.5$
LabelMatch (Ours)	540k	$40.3 \xrightarrow{+5.0} \mathbf{45.3}$

Table 2. Experimental results on COCO-additional.

	AP_{50}	$AP_{50:95}$
Supervised [22]	72.63	42.13
STAC [31]	77.45 (+4.82)	44.64 (+2.51)
ISMT [38]	77.23 (+4.60)	46.23 (+4.10)
Instant Teaching [40]	79.20 (+6.57)	50.00 (+7.87)
Unbiased Teacher [22]	77.37 (+4.74)	48.69 (+6.56)
LabelMatch (Ours)	$\mathbf{85.48}$ (+12.85)	$\mathbf{55.11}$ (+12.98)

Table 3. Experimental results on VOC.

forms the previous state-of-the-arts with different percentages of labeled data. It is worth mentioning that the proposed method achieves 25.81 mAP on 1% labeled data, which is even higher than the supervised baseline trained on 10% labeled data.

COCO-additional. We verify whether the model trained on 100% COCO data can be further improved by using additional unlabeled COCO data. As shown in Tab. 2, our method boosts the supervised baseline with +5.0 mAP, while the existing SOTA improvement is +3.6 mAP.

VOC. We evaluate the proposed method on PASCAL-VOC. As is shown in Tab. 3, our method achieves 85.48 mAP on AP_{50}, which outperforms previous state-of-the-arts by +6.28 absolute mAP improvement.

4.3. Ablation Studies

In ablation studies, we conduct experiments in the setting of 1% COCO-standard (one of 5 folds), without RPLM strategy if not specified. The training iterations are reduced to 40k for all of the experiments. More implementation details and ablation studies can be found in the Appendix.

The quality of pseudo labels. The quality of pseudo la-

bels can be reflected in three aspects: 1) the accuracy of pseudo labels; 2) foreground-background distribution; 3) class distribution (foreground-foreground distribution). We compare the proposed method against the single confidence threshold based mean teacher framework. Label-Match shows superior advancement, which we attribute to the following aspects:

- More accurate pseudo labels. As is shown in Fig. 4a, the accuracy of pseudo labels decreases when using threshold=0.7 and threshold=0.8. In contrast, the accuracy increases in LabelMatch. Although the accuracy achieves the best in threshold=0.9, the number (recall) of foregrounds is much less than the ground truth.

- Unbiased foregrounds-background distribution. As is shown in Fig. 4b, the number of pseudo labels in our method is nearly the same as the ground truth, while the number in the single confidence threshold based method is much lower, especially when threshold=0.9.

- Consistent class distribution. Fig. 4c demonstrates that LabelMatch guarantees both the foreground-background distribution and the class distribution to be nearly consistent with the ground truth. The situation is totally different when using the single confidence threshold, which brings large gaps with ground truth in many categories.

To show the quality of pseudo labels more intuitively, we give the quantitative and qualitative demonstrations in Tab. 4 and Fig. 5, respectively. From Tab. 4, we observe obvious gains in AP on top20 tail and head categories by leveraging LabelMatch compared with the single and fixed threshold. As shown in Fig. 5, there are many false positives with threshold=0.7 and many false negatives with threshold=0.9, which are removed by leveraging LabelMatch. These experimental results indicate the effectiveness of LabelMatch to re-distribute the pseudo label distribution, preventing self-training from collapsing to dominant classes. More qualitative results are shown in the Appendix.

The necessity of ACT adaptation. We randomly select a subset of unlabeled data to determine ACT for every K iterations in our implementation. As mentioned in Sec. 3.3, the proposed ACT are updated to the evolved teacher during the training phase, avoiding a false bias caused by the
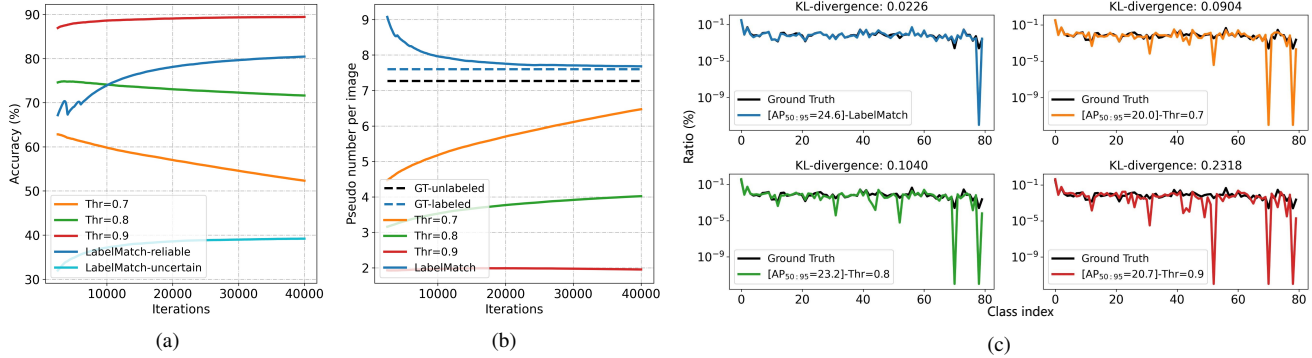
Figure 4. Ablation study about the quality of pseudo labels. (a) the accuracy of pseudo labels in the training phase. (Note: the pseudo labels with IoU overlapping the ground truth greater than 0.5 are regarded as true positives) (b) The average number of pseudo labels per image in the training phase. (c) KL divergence of the class distribution between the pseudo labels and the ground truth in the training phase.

	toaster	hair drier	scissors	microwave	toothbrush	parking meter	snowboard	bear	stop sign	fire hydrant	mous	frisbee	refrigerator	hot dog	oven	airplane	baseball bat	baseball glove	keyboard	bed	mean
thr=0.7	0.0	**0.0**	7.4	37.9	**4.8**	31.9	1.8	47.6	17.6	30.2	33.8	36.7	14.8	2.0	11.9	15.9	6.9	16.4	24.4	21.8	18.2
thr=0.8	0.0	0.0	7.7	37.2	4.7	32.0	3.0	**51.5**	32.4	46.2	44.8	44.5	33.7	4.6	**18.5**	34.8	11.1	**24.0**	21.8	27.1	24.0
thr=0.9	0.0	0.0	5.3	**37.9**	2.2	17.6	2.8	48.4	**53.4**	46.8	44.8	37.8	31.7	1.2	16.8	33.1	12.4	23.7	26.8	26.6	23.5
LabelMatch	**4.6**	0.0	**16.5**	37.5	4.6	**34.9**	9.5	49.9	50.4	49.2	47.4	45.0	34.4	8.5	17.7	**39.3**	13.2	21.0	**29.3**	**32.3**	**27.3**

	person	car	chair	book	bottle	cup	dining table	traffic light	bowl	handbag	bird	boat	truck	umbrella	bench	cow	banana	carrot	motorcycle	backpack	mean
thr=0.7	**39.1**	30.5	9.2	2.1	20.3	24.0	12.5	19.4	22.2	4.3	16.1	**14.4**	3.5	20.0	7.7	31.8	8.9	3.1	27.4	3.8	16.0
thr=0.8	39.0	31.8	10.3	**2.1**	25.8	26.3	14.1	20.2	25.6	4.1	**18.1**	13.6	8.6	20.0	11.4	32.1	8.3	3.9	**28.0**	4.8	17.4
thr=0.9	32.0	26.6	7.1	1.0	13.6	19.9	**14.9**	17.8	23.0	1.1	14.9	8.1	12.7	15.4	10.3	24.2	6.6	1.8	24.5	2.5	13.9
LabelMatch	38.8	**31.9**	**11.7**	1.9	25.6	26.1	14.6	**20.8**	**29.2**	**4.8**	17.9	13.2	**14.8**	**21.2**	**13.6**	**39.0**	**9.1**	**4.9**	27.4	**6.1**	**18.6**

Table 4. Quantitative results on top20 tail categories (upper) and top20 head categories (lower) ($AP_{50:95}$).

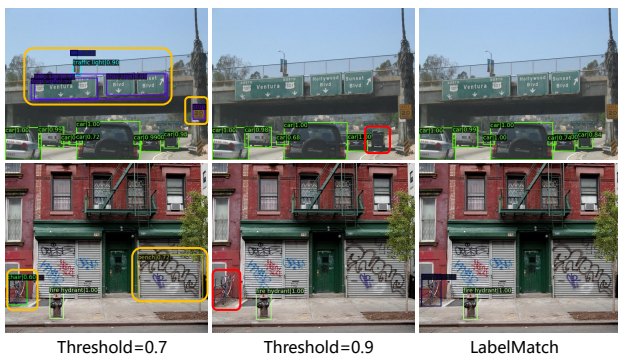Threshold=0.7　　　Threshold=0.9　　　LabelMatch

Figure 5. Qualitative comparisons between the single confidence threshold and our proposed method. Red rectangles highlight the false negatives, and yellow rectangles highlight the false positives. The score threshold for visualization is 0.6.

outdated predictions. Fig. 6a proves the necessity of ACT adaptation, where the performance is pretty bad without adaptation ($K = +\infty$). In our experiments, we simply use $K = 1000$ and a subset of unlabeled data (10,000) to refresh the thresholds.

Effect of reliable ratio α. In the training phase, we split the candidate pseudo labels into reliable pseudo labels and uncertain pseudo labels by $\alpha\%$ according to the confidence scores. Here, we analyze the influence of different ratios. As shown in Fig. 6b, if we directly set all the candidate pseudo labels as uncertain pseudo labels ($\alpha = 0$), the performance is worse than splitting some pseudo labels to reliable, which is mainly caused by the lack of box regression optimization for the unlabeled data. However, setting too many pseudo labels as reliable pseudo labels is also harmful due to the noisy boxes. We use $\alpha = 20$ for all experiments.

Proposal self-assignment. We compare different label assignment strategies for uncertain pseudo labels: 1) Ignore assignment; 2) IoU based label assignment; 3) Proposal self-assignment. Here the ignore assignment means that the uncertain pseudo labels are directly set as ignore labels. Fig. 6c shows the superiority of the proposal self-assignment strategy over other label assignments. For the ignore assignment, there exists an imbalance between foreground and background, and background dominates the object detection training, which makes the ignored objects tend to be regarded as background after training, imped-
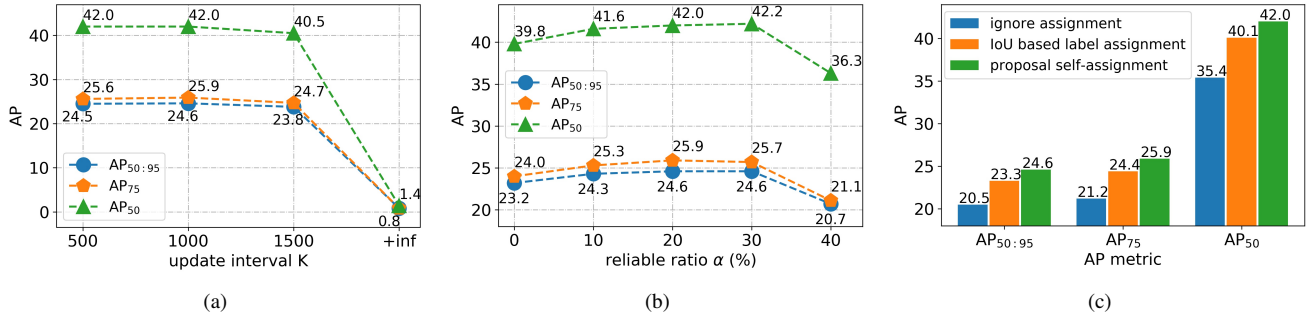
Figure 6. Ablation study about: (a) the updating interval of ACT. (b) the effect of reliable ratio α. (c) different label assignment strategies.
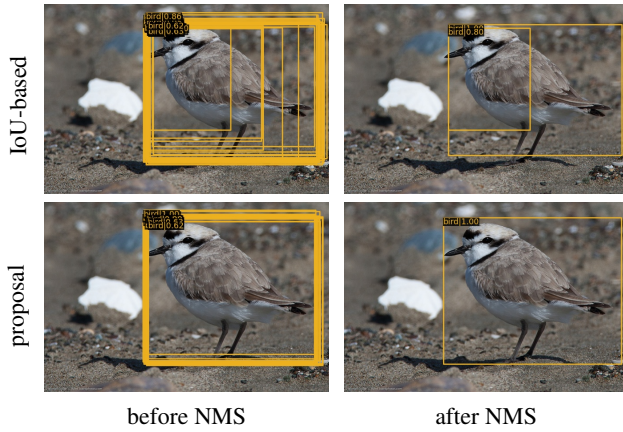

before NMS   after NMS

Figure 7. Visualization of the predictions before and after NMS post-processing. IoU based: the model is trained based on IoU based label assignment; proposal: the model is trained based on proposal self-assignment

Figure 8. Ablation study about RPLM. (a) RPLM in different percentages of labeled data. (Δ means the number of mining pseudo labels per image.) (b) RPLM in different reliable ratio α.

ing performance improvement. For IoU based label assignment, it will produce many ambiguous boxes after training due to the instance mismatch as illustrated in Fig. 7.

Effect of RPLM. To verify the effectiveness of RPLM, we estimate it in the setting of COCO-standard. As depicted in Fig. 8a, RPLM slightly boosts the performance, but remitting the sensitivity of the reliable ratio α and showing an obvious improvement even α dropping to zero (Fig. 8b). This implies that, with the favor of RPLM, we can easily adjust α to make stable performance improvement without expert techniques.

5. Conclusion and Future Work

In this paper, we first diagnose the existing SSOD frameworks experimentally and figure out the common limitation, namely label mismatch problem, including two different but complementary granularities, *i.e.*, distribution-level and instance-level. To solve the above problems, we propose the LabelMatch framework. For distribution-level mismatch, LabelMatch develops a re-distribution mean teacher to derive adaptive label-distribution-aware confidence thresholds
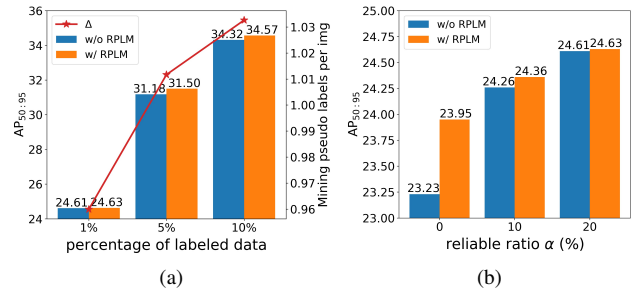
by narrowing the class distribution discrepancy between labeled and unlabeled data, and then generating unbiased pseudo labels. For instance-level mismatch, LabelMatch adopts a proposal self-assignment method, injecting the proposals generated by the student into the teacher model to supervise proposals correction. Furthermore, a reliable pseudo label mining strategy is introduced to convert high-quality uncertain pseudo labels to reliable ones, facilitating the cycle of positive feedback during self-training. Extensive experimental results verify the efficacy of the proposed LabelMatch, which establishes a new state-of-the-art on both PASCAL-VOC and MS-COCO datasets.

Limitations. While we have shown the superiority of LabelMatch, there is still a non-negligible problem that the labeled and unlabeled data are assumed to follow the same distribution. Hence, LabelMatch relies on the class distribution prior, which is inaccessible in some scenarios, *e.g.*, unsupervised domain adaptive object detection. It is beneficial to study the label mismatch problems between two different distributions and, intuitively, advance the LabelMatch to more complex situations without class distribution prior, which is an interesting future work.

Acknowledgements

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020. 2

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A. Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, volume 32, pages 5049–5059, 2019. 2

[3] 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. 5

[4] Weijie Chen, Luojun Lin, Shicai Yang, Di Xie, Shiliang Pu, Yueting Zhuang, and Wenqi Ren. Self-supervised noisy label learning for source-free unsupervised domain adaptation. *CoRR*, abs/2102.11614, 2021. 4

[5] Weijie Chen, Di Xie, Yuan Zhang, and Shiliang Pu. All you need is a few shifts: Designing efficient convolutional neural networks for image classification. In *CVPR*, 2019. 1

[6] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, volume 29, pages 379–387, 2016. 3

[7] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 2, 5

[8] 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. 3

[9] Ross Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015. 3

[10] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR '14 Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. 3

[11] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, volume 31, pages 8527–8537, 2018. 4

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[13] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pages 876–885, 2018. 2

[14] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In *Advances in Neural Information Processing Systems*, volume 32, pages 10758–10767, 2019. 1, 2

[15] Kang Kim and Hee Seok Lee. Probabilistic anchor assignment with iou prediction for object detection. In *European Conference on Computer Vision*, pages 355–371, 2020. 3

[16] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. Foveabox: Beyound anchor-based object detection. *IEEE Transactions on Image Processing*, 29:7389–7398, 2020. 3

[17] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. *International Journal of Computer Vision*, 128(3):642–656, 2020. 1

[18] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017. 5

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020. 3

[20] Tsung-Yi Lin, Michael Maire, Serge J. 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, 2014. 2, 5

[21] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In *14th European Conference on Computer Vision, ECCV 2016*, pages 21–37, 2016. 3

[22] 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: The Ninth International Conference on Learning Representations*, 2021. 1, 2, 3, 5, 6

[23] 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, 2019. 2

[24] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra r-cnn: Towards balanced learning for object detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 821–830, 2019. 3

[25] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. 3

[26] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, 2017. 1, 3

[27] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 3

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. 1, 2, 3, 5

[29] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *NIPS'16 Proceedings of the 30th International Conference on Neural Information Processing Systems*, volume 29, pages 1171–1179, 2016. 2

[30] 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 596–608, 2020. 2

[31] 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, 5, 6

[32] Peng Tang, Chetan Ramaiah, Yan Wang, Ran Xu, and Caiming Xiong. Proposal learning for semi-supervised object detection. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2290–2300, 2021. 2

[33] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *ICLR (Workshop)*, 2017. 2

[34] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9627–9636, 2019. 1, 3

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

[36] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10687–10698, 2020. 2

[37] 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. *arXiv preprint arXiv:2106.09018*, 2021. 1, 2, 3, 6

[38] 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 Computer Society Conference on Computer Vision and Pattern Recognition*, pages 5941–5950, 2021. 2, 6

[39] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z. Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9759–9768, 2020. 3

[40] 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. 1, 2, 3, 6

[41] Benjin Zhu, Jianfeng Wang, Zhengkai Jiang, Fuhang Zong, Songtao Liu, Zeming Li, and Jian Sun. Autoassign: Differentiable label assignment for dense object detection. *arXiv preprint arXiv:2007.03496*, 2020. 3