Active Teacher for Semi-Supervised Object Detection

Peng Mi[1*], Jianghang Lin[1*], Yiyi Zhou[1*], Yunhang Shen[1], Gen Luo[1], Xiaoshuai Sun[1],
Liujuan Cao[1†], Rongrong Fu[2], Qiang Xu[2], Rongrong Ji[1]
[1]Media Analytics and Computing Lab, School of Informatics, Xiamen University, 361005, China.
[2]Ascend Enabling Laboratory, Huawei Technologies, China.
{mipeng,hunterjlin007,luogen}@stu.xmu.edu.cn, {zhouyiyi,xssun,caoliujuan,rrji}@xmu.edu.cn,
shenyunhang01@gmail.com, {furongrong, xuqiang40}@huawei.com

Abstract

In this paper, we study teacher-student learning from the perspective of data initialization and propose a novel algorithm called Active Teacher[1] *for semi-supervised object detection (SSOD). Active Teacher extends the teacher-student framework to an iterative version, where the label set is partially initialized and gradually augmented by evaluating three key factors of unlabeled examples, including difficulty, information and diversity. With this design, Active Teacher can maximize the effect of limited label information while improving the quality of pseudo-labels. To validate our approach, we conduct extensive experiments on the MS-COCO benchmark and compare Active Teacher with a set of recently proposed SSOD methods. The experimental results not only validate the superior performance gain of Active Teacher over the compared methods, but also show that it enables the baseline network,* i.e., *Faster-RCNN, to achieve* 100% *supervised performance with much less label expenditure,* i.e. 40% *labeled examples on MS-COCO. More importantly, we believe that the experimental analyses in this paper can provide useful empirical knowledge for data annotation in practical applications.*

1. Introduction

Recent years have witnessed the rapid development of object detection supported by a flurry of benchmark datasets [9, 11, 22, 32] and methods [13–15, 23, 26, 28, 29]. Despite great success, the expensive instance-level annotation has long plagued the advancement and application of existing detection models. To this end, how to save labeling expenditure has become a research focus in object detection [4, 5, 17, 17, 24, 25, 33, 35, 36, 38, 40, 41].

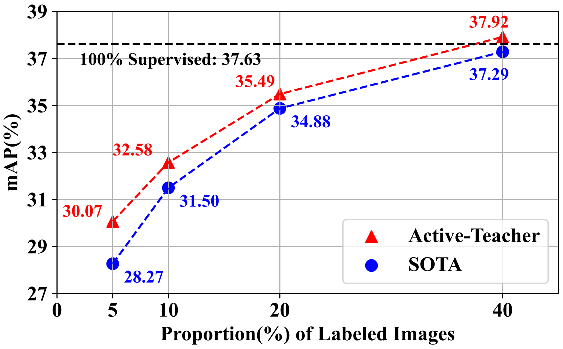


Figure 1. The performance comparison between Active Teacher and the state-of-the-art (SOTA) method [24] with different proportions of labeled data in MS-COCO. Active Teacher exceeds 100% fully supervised performance with only 40% label information.

Inspired by recent success in image classification [2, 3, 19, 34, 39], some practitioners resort to teacher-student learning for semi-supervised object detection (SSOD) [24, 35, 37]. Specifically, this methodology uses a teacher network with weakly augmented labeled data to generate high-quality pseudo-labels for the student network with strong data augmentation [8, 10, 44]. This self-training process helps the model explore large amounts of unlabeled data based on a very limited number of annotations. Following this methodology, Sohn *et al.* [35] proposed the first teacher-student framework called STAC for SSOD. This simple framework outperforms the existing semi-supervised methods [4, 33, 36, 38] by a large margin, showing the great potential of teacher-student learning in object detection.

Some very recent SSOD works [24, 37, 50] are proposed to improve this methodology. For instances, Liu *et al.* [24] apply *exponential moving average* (EMA) [39] to train a gradually progressing teacher to alleviate the class-

[*] Equal Contribution. † Corresponding Author.
[1] Source code are available at: https://github.com/HunterJ-Lin/ActiveTeacher

imbalance and over-fitting issues. Zhou *et al*. [50] propose an instant pseudo labeling strategy to reduce the impact of the confirmation bias and improve the quality of pseudo labeling. In [37], Tang *et al*. adopt a detection-specific data ensemble to produce more reliable pseudo-labels. Conclusively, these methods mainly focus on the framework optimization or the negative impact of noisy pseudo-labels, of which contributions are orthogonal to ours.

In this paper, we study this semi-supervised methodology from the perspective of data initialization. More specifically, we investigate how to select the optimal labeled examples for teacher-student learning in SSOD. To explain, although a plenty of pseudo-labels are generated for self-training, ground-truth label information still plays a key role in the infant training phase, which determines the quality of pseudo-labels and the performance lower-bound of the teacher networks [24, 35, 50]. Meanwhile, in some teacher-student methods [24,31], the pseudo-labels are only used to optimize the predictions of object categories and foreground-background proposals, while the optimization of bounding boxes regression still relies on the ground-truth annotations. In this case, we observe that ground-truth label information plays an important role in SSOD, which, however, is still left unexplored.

To this end, we propose a new teacher-student method, coined as *Active Teacher*, for semi-supervised object detection. As shown in Fig. 2, Active Teacher extends the conventional teacher-student framework to an iterative one, where the label set is partially initialized and gradually augmented via a novel active sampling strategy. With this modification, Active Teacher can maximize the effect of limited label information by active sampling, which can also improve the quality of pseudo-labels. We further investigate the selection of labeled examples from the aspects of *difficulty*, *information* and *diversity*, and the values of these metrics are automatically combined without hyper-parameter tunning. Through these metrics, we can explore what kind of data are optimal for SSOD.

To validate the proposed method, we conduct extensive experiments on the benchmark dataset, namely MS COCO [22][2]. The experimental results not only confirm the significant performance gains of Active Teacher against a set of state-of-the-art SSOD methods, *e.g*., +6.3% and +23.3% compared with Unbiased Teacher [24] and STAC [35] on 5% MS-COCO, respectively. It also shows that Active Teacher enables the baseline detection network, *i.e*., Faster-RCNN [29], to achieve 100% supervised performance with much less labeling expenditure, *e.g*., with 40% labeled examples on MS-COCO, as shown in Fig. 1. More importantly, we also provide the in-depth analyses for active sampling, which can give useful hints for data annotation in practical applications of object detection.

In summary, our contribution is two-fold:

- We present the first attempt of studying data initialization in teacher-student based semi-supervised object detection (SSOD), and conduct extensive experiments for different sampling strategies. These quantitative and qualitative analyses can provide useful references for data annotation in practical applications.
- We propose a new teacher-student framework for SSOD called *Active Teacher*, which not only outperforms a set of SSOD methods on the benchmark dataset, but also enables the baseline detection network achieve 100% fully supervised performance with much less label expenditure.

2. Related Work

Object Detection. With the rapid development of deep neural networks, object detection has achieved great progress both academically and industrially [13–16, 20, 23, 26–29]. Object detection is roughly divided into two genres: one-stage and two-stage detectors. The representative work of one-stage methods includes YOLO [13, 26–28], SSD [23], *etc*., and the ones of two-stage models include RCNN series [14, 15, 29] and its variants [16, 20]. The main difference between these two methodologies is that the one-stage method directly predicts the coordinates and probability distribution of the object based on the feature map, while the two-stage methods use region proposal networks [29] to sample potential objects, and further predict the probability distribution and coordinate information of the object, respectively. Following the prior works [24, 35, 50], we focus the semi-supervised learning of two-stage models and use Faster-RCNN [29] as our baseline network.

Semi-Supervised Object Detection. In the field of computer vision, most existing researches on semi-supervised learning mainly focus on image classification [1, 7, 19, 45], which can be roughly divided into consistency-based and pseudo labeling based methods, respectively . Consistency-based approaches [2, 3, 12, 17, 34] constrain the model to make it robust to noise via producing consistent prediction results. Pseudo labeling [1, 24, 34, 35, 37, 50] methods firstly train the classifiers with ground-truth annotations and generate pseudo-labels for unlabeled data, and finally retrain models with all data. Recently, some works [17, 18, 24, 35, 50] apply semi-supervised learning to object detection. CSD [17] randomly flips images multiple times, driving the model to produce consistent predictions for these flipped images. ISD [18] uses *mixup* [49] to constrain model training. Following the popular teacher-student framework [39], STAC [35] proposes the first teacher-student based framework for SSOD. Due to the static annotating strategy, the pseudo-labels in STAC are fixed, which limits the final detection performance. In Instant-Teaching [50], both teacher

[2] More experimental results can be found in our Github project.

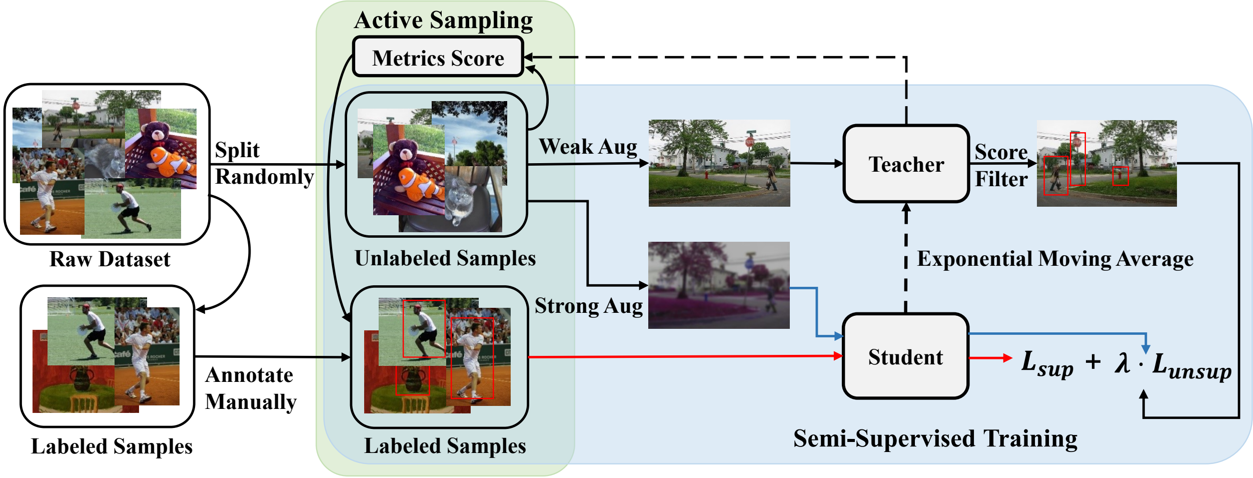


Figure 2. The overall framework of the proposed Active Teacher. In Active Teacher, the label set is partially initialized and gradually augmented after each semi-supervised training. Active Teacher includes two detection networks, *i.e*., Faster-RCNN [29], with the same configurations, namely *Teacher* and *Student*. Teacher is used to generate pseudo-labels for training Student, and its parameters are gradually updated from Student via EMA [39]. Student is trained with both ground-truth and pseudo-labels, denoted as $\mathcal{L}_{sup}$ and $\mathcal{L}_{unsup}$, respectively. Teacher also serves to estimate the unlabeled examples for active sampling.

and student share the same parameters to deal with above problem. However, they still suffer from extreme instability in the initial training phase and require a high confidence score threshold for generating pseudo-labels. Unbiased teacher [24] exploits EMA [39] to optimize teacher from student gradually. In addition, Unbiased teacher apply EMA [39] and focal loss [21] to address the pseudo-label over-fitting problem in teacher-student learning.

Active Learning. There are also some active-learning based methods proposed to reduce the labeling expenditure of object detection [42, 47, 48]. For instance, Wang *et al*. [42] use different active sampling metrics for different stages in object detection. CALD [47] measures information by calculating the data consistency of bounding boxes before and after augmentation. MI-AOD [48] applies multi-instance learning to suppress the pseudo-label noises.

In this paper, we focus on the teacher-student based semi-supervised learning for object detection.

3. Active Teacher

The overall framework of the proposed Active Teacher is illustrated in Fig. 2. As shown in this figure, Active Teacher consists of an iterative teacher-student structure, where the limited label set is partially initialized and gradually augmented. After each iteration, the well-trained teacher network is used to evaluate the importance of unlabeled examples in terms of the proposed metrics, *i.e*., *information*, *diversity* and *difficulty*, based on which active data augmentation is performed. The detailed procedure is depicted in Algorithm 1. In the following section, we introduce Active Teacher from the aspects of semi-supervised learning and active sampling, respectively.

Algorithm 1 Pseudo Code of Active Teacher

```
Input: Labeled Dataset $\{\mathcal{X}_L^0, \mathcal{Y}_L^0\}$, Unlabeled Dataset $\{\mathcal{X}_U^0\}$, Maximum Iteration $K$
Output: Teacher Model $M^t$
for all $x_l \in \mathcal{X}_L^0$ and $x_u \in \mathcal{X}_U^0$ do
    Update the parameters of Student $M_0^s$ by Eq. (1)
    Update the parameters of Teacher $M_0^t$ by Eq. (6)
end for
for all i=1,...,K do
    for all $x_u \in \{\mathcal{X}_U^{i-1}\}$ do
        Calculate sampling score of unlabeled data using Teacher network $M_{i-1}^t$ by Eq. (11);
    end for
    Rank the data based on score.
    Select the top-N data $\{\mathcal{X}_P^i\}$ and annotate them with label $\{\mathcal{Y}_P^i\}$;
    Update labeled set $\{\mathcal{X}_L^i, \mathcal{Y}_L^i\} = \{\mathcal{X}_L^{i-1}, \mathcal{Y}_L^{i-1}\} \cup \{\mathcal{X}_P, \mathcal{Y}_P\}$;
    Update unlabeled set $\{\mathcal{X}_U^i\}$ = $\{\mathcal{X}_U^{i-1}\}$ - $\{\mathcal{X}_P\}$
    for all $x_l \in \mathcal{X}_L^i$ and $x_u \in \mathcal{X}_U^i$ do
        Update the parameters of Student $M_i^s$ by Eq. (1)
        Update the parameters of Teacher $M_i^t$ by Eq. (6)
    end for
end for
return $M_K^t$
```

3.1. Semi-Supervised Learning

Given a set of labeled data $\mathcal{D}_L = \{\mathcal{X}_L, \mathcal{Y}_L\}$ and a set of unlabeled data $\mathcal{D}_U = \{\mathcal{X}_U\}$, where $\mathcal{X}$ denotes the examples and $\mathcal{Y}$ is the label set, the target of semi-supervised learning is to maximize model performance based on both labeled and unlabeled data.

Similar to prior works [24, 37], our semi-supervised learning paradigm also includes two detection networks with the same configurations, namely *Teacher* and *Student*, as shown in Fig. 2. In this paper, we use Faster-RCNN [29] as our baseline detection network. The teacher network is in charged of pseudo-label generation, while the student one is optimized with both ground-truth and pseudo-labels. Specifically, the optimization loss for the student network can be defined as:

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda \cdot \mathcal{L}_{unsup}, \tag{1}$$

where $\mathcal{L}_{sup}$ and $\mathcal{L}_{unsup}$ denote the losses for supervised and unsupervised learning, respectively, and λ is the hyper-parameter to trade-off between $\mathcal{L}_{sup}$ and $\mathcal{L}_{unsup}$.

For object detection, $\mathcal{L}_{sup}$ consists of the classification loss $\mathcal{L}_{cls}$ of RPN and ROI head, and the one for bounding box regression $\mathcal{L}_{loc}$. Then, $\mathcal{L}_{sup}$ is defined as

$$\mathcal{L}_{sup} = \frac{1}{N_l} \sum_{i=1}^{N_l} (\mathcal{L}_{cls}(x_l^i, y_{cls}^i) + \mathcal{L}_{loc}(x_l^i, y_{loc}^i)), \tag{2}$$

where $\mathcal{L}_{cls}$ and $\mathcal{L}_{loc}$ are calculated by

$$\begin{aligned} \mathcal{L}_{cls}(x_l^i, y_{cls}^i) &= \mathcal{L}_{cls}^{rpn}(x_l^i, y_{cls}^i) + \mathcal{L}_{cls}^{roi}(x_l^i, y_{cls}^i), \\ \mathcal{L}_{loc}(x_l^i, y_{loc}^i) &= \sum_{c \in \{\text{x,y,h,w}\}} \text{Smooth}_{L1}(t_c^i - y_c^i). \end{aligned} \tag{3}$$

Here, x_l refers to the labeled example, y_{cls} and y_{loc} are its labels, and N_l denotes the number of x_l. t_c is the c-th coordinate of the output image x_i. In terms of L_{loc}, we use the smooth *L*-1 loss for the bounding box regression:

$$\text{Smooth}_{L1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1, \\ |x| - 0.5 & \text{otherwise.} \end{cases} \tag{4}$$

For $\mathcal{L}_{unsup}$, we only use the pseudo-labels of RPN and ROI head predictions, similar to that in [24]. It is formulated as

$$\mathcal{L}_{unsup} = \frac{1}{N_u} \sum_{i=1}^{N_u} \mathcal{L}_{cls}(x_u^i, \hat{y}_{cls}^i), \tag{5}$$

where $\mathcal{L}_{cls}$ is the same as Eq. (2), and $\hat{y}_{cls}^i$ is the pseudo-labels generated by the teacher network.

To avoid the class-imbalance and over-fitting issues, we follow [24, 37] to freeze the optimization of the teacher network during semi-supervised training, and update its parameters from the student network via *Exponential Moving Average* (EMA) [39]:

$$\theta_t^i \leftarrow \alpha \theta_t^{i-1} + (1 - \alpha)\theta_s^i, \tag{6}$$

where θ_t and θ_s are the parameters of the teacher and student networks, respectively, and i denotes the i-th training step. α is the hyper-parameter to determine the speed of parameter transmission, which is normally close to 1. To improve the quality of pseudo-labels, we also apply non-maximum suppression (NMS) [15] and confidence threshold to filter repetitive and uncertain pseudo-labels.

3.2. Active Sampling

In Active Teacher, the label set is partially initialized and augmented through the teacher network after each semi-supervised training. We explore what kind of examples (or images) are critical for semi-supervised object detection, and introduce three active sampling metrics, namely *difficulty*, *information* and *diversity*.

Difficulty is the widely-used metric for active learning [6, 51], and is normally measured based on the entropy of the probability distribution predicted by the model. A higher entropy shows that the model is more uncertain about its prediction, suggesting that the example is more difficult.

In SSOD, we measure the difficulty score s_i^{diff} of an unlabeled example based on the category prediction of the teacher network, which is defined as

$$s_i^{\text{diff}} = -\frac{1}{n_b^i} \sum_{j=1}^{n_b^i} \sum_{k=1}^{N_c} p(c_k; b_j, \theta_t) \log p(c_k; b_j, \theta_t), \tag{7}$$

where n_b^i is the number of the predicted bounding box after NMS and confidence filtering, N_c is the number of object categories and $p(c_k; b_j, \theta_t)$ is the prediction probability of the k-th category by the teacher network. With Eq. (7), we can judge whether the image is difficult for SSOD based on the prediction uncertainty of the teacher network.

Information is a metric to measure the amount of information of the unlabeled image for SSOD. In some classification tasks [6, 51], it is often calculated by prediction entropies, similar to *difficulty*. However, in object detection, richer information means that more visual concepts appear in the image, so the model can learn more detection patterns. To this end, we use the prediction confidence to measure this metric:

$$s_i^{\text{info}} = \sum_{j=1}^{n_b^i} \text{confidence}(b_j, \theta), \tag{8}$$

where the confidence(b_j, θ_t) is the highest confidence score in j-th bounding box predicted by the teacher network. From Eq. (8), we can see that the larger s^{info}, the more visual concepts recognized by the teacher network, suggesting that the image has richer information.

Diversity is a metric to measure the distribution of object categories in an image. The diversity score s^{dive} is calculated by

$$s_i^{\text{dive}} = |\{c_j\}_{j=1}^{n_b^i}| \tag{9}$$

where c_j is the predicted category of the $j-$th bounding box, and $|\cdot|$ is the cardinality. The difference between information and diversity is that the former will sample images of more visual instances that might belong to only one or a few categories, while the later will favor those involving more different concepts.

Metrics Combination. The introduced metrics may be able to answer which type of examples are suitable for SSOD. However, a practical problem is that the models in different states may have different requirements for label information. Besides, how to maximize the benefits of these metrics without extensive trials remains a challenge. To this end, we propose a simple yet efficient solution to automatically combine these metrics, termed *AutoNorm*.

Before combining these metrics, we notice that the value ranges of these metrics differ greatly. For instance, the *difficulty* scores is usually between 0.3 and 0.8 with a theoretical maximum of $\log N_c$, while the *information* score often ranges from 4.0 to 6.0. In this case, the first step of combination is to normalize their values:

$$\hat{s_i^m} = \frac{s_i^m}{s_{\text{max}}^m} \tag{10}$$

where $m \in \{\textit{difficulty}, \textit{information}, \textit{diversity}\}$ represent the metrics, the s_{max}^m is the maximum value of this metric.

Since these metrics represent image information from different aspects, we further build a three-dimensional sampling space to represent each example as $\vec{s_i} = (s_i^{\text{diff}}, s_i^{\text{info}}, s_i^{\text{dive}})$. The evaluation result of each unlabeled example can be regarded as a point in this space. Afterwards, we use *L-p* normalize the data points into a single scalar s_{L_p}, which is obtained by

$$s_{L_p} = L_p(\vec{s}) = ||\mathbf{s}||_p = \sqrt[p]{\sum_{i=1}^{3} s_i^p} \tag{11}$$

where $\vec{s} = (s_1, s_2, s_3) = (\hat{s_i^{\text{diff}}}, \hat{s_i^{\text{info}}}, \hat{s_i^{\text{dive}}})$. Empirically, we use L_1 norm to combine these three metrics. When using *L-p* (p>1) norm, the metrics with higher values will receive more sampling weights, *e.g.*, *difficulty*, which is found to be suboptimal in our experiments.

4. Experiment

4.1. Dataset and Metric

We evaluate our approach on the main benchmark for object detection, namely MS-COCO [22]. Specifically, MS-COCO divides the examples into two splits, namely *train2017* and *val2017*. The *train2017* has 118k labeled images. During our experiments, this split is further divided into the labeled set and the unlabeled one, similar to the prior works in SSOD [24, 35]. In practice, we adopt the settings of 1%, 2%, 5%, 10% and 20% labeled data of *train2017* for experiments and the comparisons with the other SSOD methods [24, 35, 37, 46, 50]. The rest examples are regarded as unlabeled data. In terms of model evaluation, we follow the previous works [17, 24, 35, 37, 46, 50] adopt mAP (50:95) [22] as the metric of our experiments. And *val 2017*, which has 5k images, is used for evaluation.

4.2. Experimental Settings

Following the most work in SSOD [17, 24, 35, 37, 46, 50], we use Faster-RCNN with ResNet-50 as our baseline detection network. The implementation and hyper-parameter setting are the same as those in Detectron2 [43]. In terms of semi-supervised learning, we also follow the works in [24] to pre-train the teacher network with the supervised objectives defined in Eq. (2). The numbers of pre-training steps is set to 2k for all experimental settings. Afterwards, the student network is initialized with the parameters of the teacher one. The total training steps for each semi-supervised learning are 180k. The optimizer used is SGD [30], and the learning rate linearly increases from 0.001 to 0.01 at the first 1k iterations, and is divided by 10 at 179,990 iteration and 179,995 iteration, respectively. Similar to [24], we apply *random horizontal flip* as weak augmentation for the teacher, and the strong augmentations for the student include *horizontal flip*, *color jittering*, *grey scale*, *gaussian blur* and *CutOut* [10]. We use a threshold $\tau = 0.7$ to filter the pseudo-labels of low quality. We use $\alpha = 0.9996$ for EMA and $\lambda = 4$ for the unsupervised loss on all experiments. In terms of active sampling, we set the iteration number in Algorithm 1 as 2 in this paper. For all experiments, half of the label set are randomly selected, and the other half are actively selected after semi-supervised learning. The batch size is set to 64, which consists 32 labeled and 32 unlabeled images via random sampling.

4.3. Experimental Result

4.3.1 Quantitative Comparisons

Comparisons with the state-of-the-arts. We first compare Active Teacher with a set of teacher-student based SSOD methods, of which results are given in Table 1. From this table, we can first observe that all teacher-student based methods greatly outperform the traditional supervised learning. Besides, we can also notice that with the careful designs in framework, those recently proposed teacher-student methods, *e.g.* Unbiased Teacher [24], improve the pioneer obviously, *i.e.* STAC, suggesting the notable progresses made in teacher-student based SSOD. However, their competition also becomes more fierce. Even so, we still observe that

Table 1. Comparison between the proposed Active Teacher and other SSOD methods on MS-COCO *val2017*. The metric we used is mAP (50:95). “Supervised” refers to the performance of the model trained with labeled data only. * is our re-implementation. Δ: AP gain to the supervised performance. Our method consistently outperforms the compared methods.

	COCO-Standard									
	1%	Δ	2%	Δ	5%	Δ	10%	Δ	20%	Δ
Supervised [29]	9.05	+0.00	12.70	+0.00	18.47	+0.00	23.86	+0.00	26.88*	+0.00*
STAC [35]arXiv2020	13.97	+4.92	18.25	+5.55	24.38	+5.91	28.64	+4.78	/	/
ISMT [46]CVPR2021	18.88	+9.83	22.43	+9.73	26.37	+7.90	30.53	+6.67	/	/
Instant-Teaching [50]CVPR2021	18.05	+9.00	22.45	+9.75	26.75	+8.28	30.40	+6.54	/	/
Humble-Teacher [37]CVPR2021	16.96	+7.91	21.72	+9.02	27.70	+9.23	31.61	+7.75	/	/
Unbiased-Teacher [24]ICLR2021	20.75	+11.70	24.30	+11.60	28.27	+9.80	31.50	+7.64	34.88*	+8.00*
Active-Teacher(Ours)	**22.20**	**+13.15**	**24.99**	**+12.29**	**30.07**	**+11.60**	**32.58**	**+8.72**	**35.49**	**+8.61**

Table 2. Experiment of how much labeled data is for achieve 100% supervised performance(37.63 [24]) by Unbiased-Teacher [24] and our Active-Teacher on MS-COCO.

	COCO-Standard			
	5%	10%	20%	40%
Unbiased-Teacher	28.27	31.50	34.88	37.29
Active-Teacher	30.07	32.58	35.49	37.92

Table 3. The result of Active Teacher on STAC [35]. We just replace the initial data while keep the rest settings the same.

	COCO-Standard		
	1%	5%	10%
STAC	13.97	24.38	28.64
STAC+Ours	14.79	26.19	29.77

the proposed Active Teacher can achieve obvious performance gains on all experimental settings, *e.g.*, +6.3% than Unbiased-Teacher with 5% label information. These results greatly confirm the effectiveness of our method.

Requirement of labeled data to achieve supervision. In practical applications, the minimum amount of labeled data required to achieve supervised performance is more concerned. For this purpose, we conduct a comparison between Unbiased-Teacher [24] and our Active Teacher. As shown in Table 2, with 40% labeled data our method could achieve supervised performance easily.

Effect of Active Teacher on different AP metrics. Fig. 4 shows the detailed performance gains of Active Teacher against Unbiased Teacher on more metrics. On 5% labeled data, Active Teacher can greatly improve the performance on the detection of medium and small objects, *i.e.*, APs and APm, suggesting that Active Teacher can sample images with more small objects. On 20% labeled data, all AP metrics can obtain obvious improvements by Active Teacher, which also suggests its change in data sampling.

Generalization capability of active sampling. Active Teacher is also highly generalized. Table 3 illustrates the performance changes of STAC after using the selected label information by Active Teacher. Without bells and whistles, this simple modification can lead to obvious performance gains of STAC on all experimental settings, strongly suggesting the generalization ability of our method.

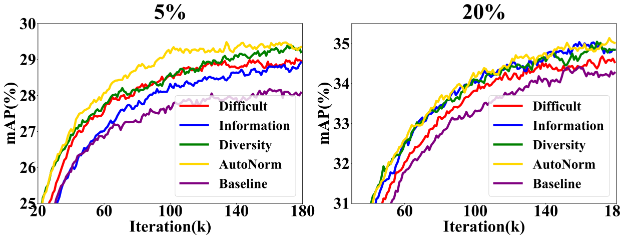


Figure 3. Training curves of active sampling with different sampling metrics on 5% and 20% labeled data. The proposed AutoNorm can well combine the advantages of three metrics.

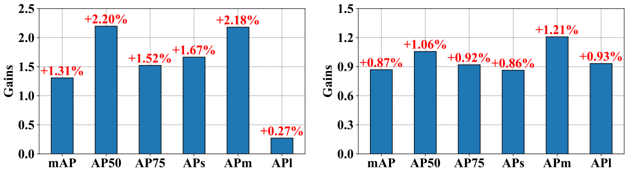


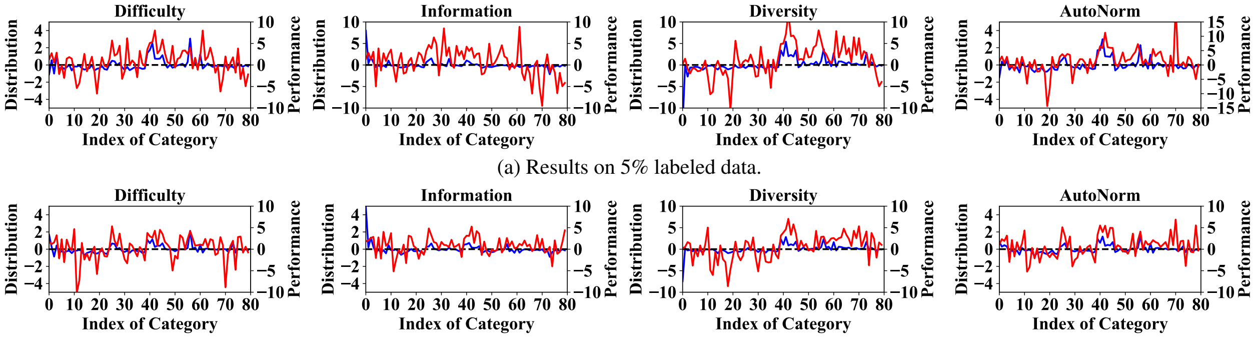
Figure 4. Changes of specific AP indicators of Active Teacher compared with Unbiased Teacher on 5% and 20% labeled data. Active Teacher is more sensitive to small and medium sized object.

Ablation. We also ablate the proposed metrics with different proportions of labeled data, as shown in Table 4. From this table, we can see that three metrics, *i.e.*, *difficulty*, *information* and *diversity*, are all beneficial for SSOD. However, under different settings of label proportions, their performance is also different, which verifies the assumption we made in Sec 3.2. For instance, with more label examples, the metric of information will performs better, and *vice verse*. In addition, as shown in Fig. 3, AutoNorm is superior than the other metrics during the training and obtains the overall better performance finally, which well confirms its effectiveness.

Sampling distributions and performance changes. To obtain deep insight into these metrics, we further com-

Table 4. Ablation study of different sampling strategies in Active Teacher. Note that these results are experimented with a smaller batch size, *i.e.* 32, which are slightly inferior than those in Table 1.

Strategy	Metric			COCO-Standard		
	Difficulty	Information	Diversity	5%(2.5%+2.5%)	10%(5%+5%)	20%(10%+10%)
Baseline	-	-	-	27.84	31.39	34.26*
Difficulty	✓	-	-	29.03	32.13	34.68
Information	-	✓	-	28.92	31.98	35.04
Diversity	-	-	✓	29.40	32.26	35.05
AutoNorm	✓	✓	✓	29.48	32.08	35.13



(a) Results on 5% labeled data.

(b) Results on 20% labeled data.

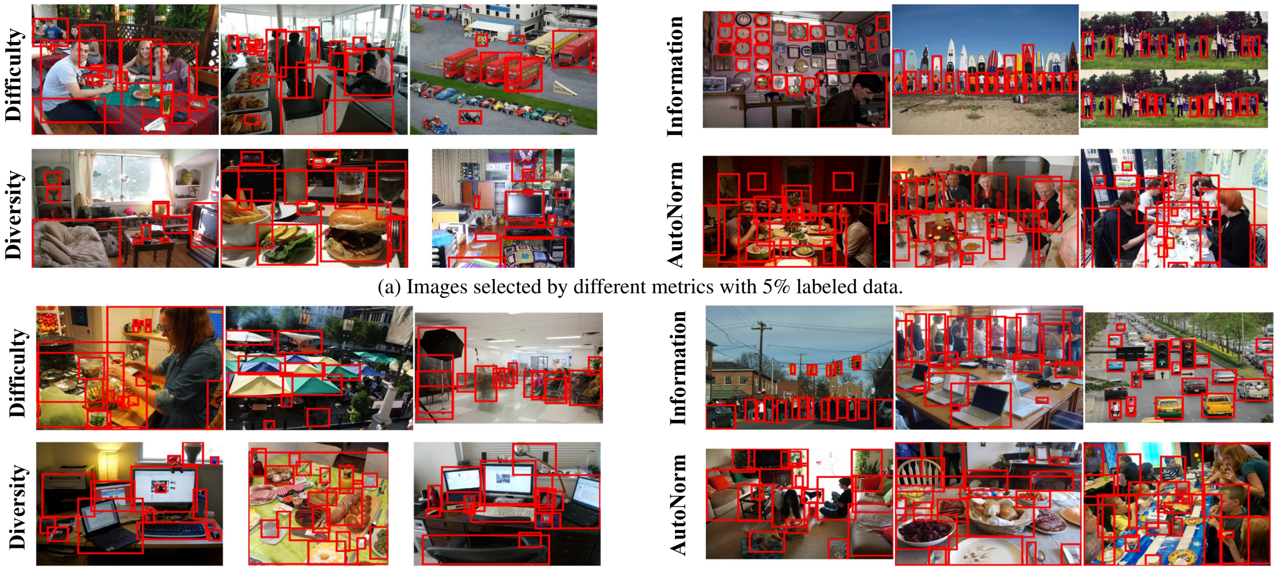
Figure 5. The relative changes of sample distribution (**blue**) and performance (**red**) of Activate Teacher with and without active sampling on different metrics. The results are obtained on 5% and 20% labeled data.

pare their detailed sampling distributions and performance changes on all categories, of which results are depicted in Fig. 5. From these results, we can first observe that *information* is easy to suffer from *inverse compensation effect*. Specifically, the categories that already take a large proportion of data will receive more samplings from this metric. As a result, the biased distribution and unbalanced performance will become more prominent under this metric. Notably, *diversity* is the opposite of *information*, which can also address *diminishing marginal effect*. From Fig. 5, we can find that the performance gains of the major categories will not keep increasing with more examples. In contrast, some small categories will obtain more improvements via data augmentation, which can be achieved by *diversity*. However, due to the obvious difference between its sampling distribution and the real one, the advantage of *diversity* will be weaken as the number of labeled examples increases. The distribution of *difficulty* matches the real one. Due to the preference of outliers, its overall performance is not significant. Instead, the proposed *AutoNorm* can make good use of three metrics, while maintaining the amount of information and the diversity of examples. Besides, it is also closer to the real data distribution.

4.3.2 Qualitative analysis

What examples are selected by these metrics? In Fig. 6, we visualize the examples selected by these metrics based on 5% and 20% labeled data. From Fig. 6, we can first observe that the selected examples well correspond to the definitions of these metrics. For instance, *difficulty* will sample examples with objects that are difficult to detect, *e.g.*, small objects, and *information* prefers the ones with more instances, *e.g.* street views. *Diversity* will select the images containing more categories, *e.g.* dining room. In addition, we can also notice some slightly difference between the samplings with 5% and 20% labeled data. Specifically, under 5%, the teacher is not sufficiently trained, so it can only estimate the examples of the common categories. For instance, *information* will sample a picture of only people, which also explains why its sampling is less effective on 5%. In contrast, under 20%, the example estimation becomes more comprehensive. Besides, we can find that the proposed AutoNorm is the optimal strategy on both settings. The images sampled by AutoNorm are *full of information, rich in categories and different in object sizes*. We believe that this is also the proper criteria of data sampling for SSOD from an overall perspective.

Effects on pseudo-labels. We further visualize the pseudo-labels of Active Teacher with and without active sampling on different training steps. Firstly, we can find that there is still an obvious gap between the qualities of the pseudo-labels and the ground-truth ones. Even so, with the help of active sampling, Active Teacher can still generate more pseudo-labels with better qualities in different training steps. As shown in Fig. 7, Active Teacher can also detect more small objects in image. This result greatly confirms our argument that data initialization also affects the qualities of pseudo-labels.



(a) Images selected by different metrics with 5% labeled data.

(b) Images selected by different metrics with 20% labeled data.

Figure 6. Visualization of the images with top ranks with 5% and 20% labeled proportions and different sampling metrics. The bounding boxes in red are predicted by the teacher network.

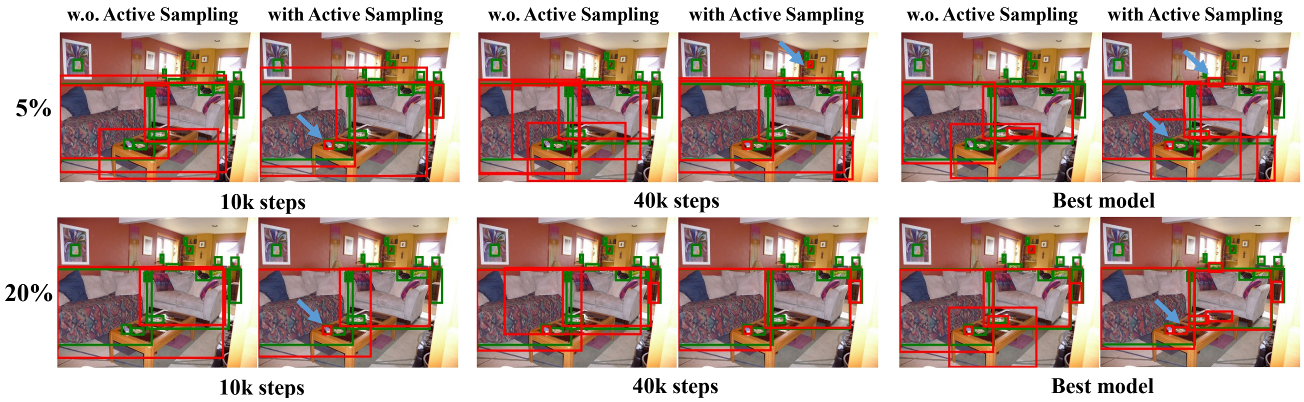


Figure 7. Visualization of the pseudo-labels predicted by Active Teacher with and without active sampling at different training steps. The **green** bounding boxes are the ground-truths, while the **red** ones are pseudo-labels predicted by the teacher network.

5. Conclusion

In this paper, we propose a novel teacher-student based method for semi-supervised object detection (SSOD), termed *Active Teacher*. Different from prior works, Active Teacher studies SSOD from the perspective of data initialization, which is supported with a novel active sampling strategy. Meanwhile, we also investigate the selection of examples from the aspects of *information*, *diversity* and *difficulty*. The experimental results not only show the superior performance gains of Active Teacher over the existing methods, but also show that it can help the baseline network achieve 100% supervised performance with much less label expenditure. Meanwhile, the quantitative and qualitative analyses provide useful hints for the data annotation in practical applications.

Limitation. A potential issue of Active Teacher is that it theoretically takes $k-1$ times more training steps than the other teacher-student methods, where k is the number of training iterations in Algorithm 1. In our experiments, we find that $k=2$ can already help the model obtain obvious performance gains. Considering the fact that data annotation is much more expensive than model training in some practical applications of SSOD, *e.g.*, security surveillance and industrial inspection, we believe that the doubled training time is still acceptable.

Acknowledgements. This work was supported by the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No.U21B2037, No.62176222, No.62176223, No.62176226, No.62072386, No.62072387, No.62072389, and No.62002305), Guangdong Basic and Applied Basic Research Foundation (No.2019B1515120049), and the Natural Science Foundation of Fujian Province of China (No.2021J01002). We also thank Huawei Ascend Enabling Laboratory for the continuous support in this work.

References

[1] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27:3365–3373, 2014. 2

[2] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019. 1, 2

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019. 1, 2

[4] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1

[5] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 1

[6] Jae Won Cho, Dong-Jin Kim, Yunjae Jung, and In So Kweon. Mcdal: Maximum classifier discrepancy for active learning. *arXiv preprint arXiv:2107.11049*, 2021. 4

[7] Mirsad Cosovic, Achilleas Tsitsimelis, Dejan Vukobratovic, Javier Matamoros, and Carles Anton-Haro. 5g mobile cellular networks: Enabling distributed state estimation for smart grids. *IEEE Communications Magazine*, 55(10):62–69, 2017. 2

[8] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019. 1

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[10] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 1, 5

[11] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 1

[12] Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan Ö Arık, Larry S Davis, and Tomas Pfister. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *European Conference on Computer Vision*, pages 510–526. Springer, 2020. 2

[13] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 1, 2

[14] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1, 2

[15] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 1, 2, 4

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[17] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. *Advances in neural information processing systems*, 32:10759–10768, 2019. 1, 2, 5

[18] Jisoo Jeong, Vikas Verma, Minsung Hyun, Juho Kannala, and Nojun Kwak. Interpolation-based semi-supervised learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11602–11611, 2021. 2

[19] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 1, 2

[20] 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

[21] 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. 3

[22] 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, 2, 5

[23] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 1, 2

[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, 4, 5, 6

[25] Ishan Misra, Abhinav Shrivastava, and Martial Hebert. Watch and learn: Semi-supervised learning for object detectors from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3593–3602, 2015. 1

[26] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1, 2

[27] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 2

[28] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1, 2

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region

proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 1, 2, 3, 4, 6

[30] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951. 5

[31] Aruni RoyChowdhury, Prithvijit Chakrabarty, Ashish Singh, SouYoung Jin, Huaizu Jiang, Liangliang Cao, and Erik Learned-Miller. Automatic adaptation of object detectors to new domains using self-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 780–790, 2019. 2

[32] 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

[33] Yunhang Shen, Rongrong Ji, Zhiwei Chen, Yongjian Wu, and Feiyue Huang. Uwsod: Toward fully-supervised-level capacity weakly supervised object detection. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7005–7019. Curran Associates, Inc., 2020. 1

[34] 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. *arXiv preprint arXiv:2001.07685*, 2020. 1, 2

[35] 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, 5, 6

[36] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1

[37] 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. 1, 2, 4, 5, 6

[38] Yuxing Tang, Josiah Wang, Boyang Gao, Emmanuel Dellandréa, Robert Gaizauskas, and Liming Chen. Large scale semi-supervised object detection using visual and semantic knowledge transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2119–2128, 2016. 1

[39] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017. 1, 2, 3, 4

[40] Jiajie Wang, Jiangchao Yao, Ya Zhang, and Rui Zhang. Collaborative learning for weakly supervised object detection. *arXiv preprint arXiv:1802.03531*, 2018. 1

[41] Keze Wang, Xiaopeng Yan, Dongyu Zhang, Lei Zhang, and Liang Lin. Towards human-machine cooperation: Self-supervised sample mining for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[42] Zhenyu Wang, Yali Li, Ye Guo, Lu Fang, and Shengjin Wang. Data-uncertainty guided multi-phase learning for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4568–4577, 2021. 3

[43] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019. 5

[44] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019. 1

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

[46] 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. 5, 6

[47] Weiping Yu, Sijie Zhu, Taojiannan Yang, Chen Chen, and Mengyuan Liu. Consistency-based active learning for object detection. *arXiv preprint arXiv:2103.10374*, 2021. 3

[48] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple instance active learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5330–5339, 2021. 3

[49] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 2

[50] 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, 5, 6

[51] Zongwei Zhou, Jae Shin, Lei Zhang, Suryakanth Gurudu, Michael Gotway, and Jianming Liang. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7340–7351, 2017. 4