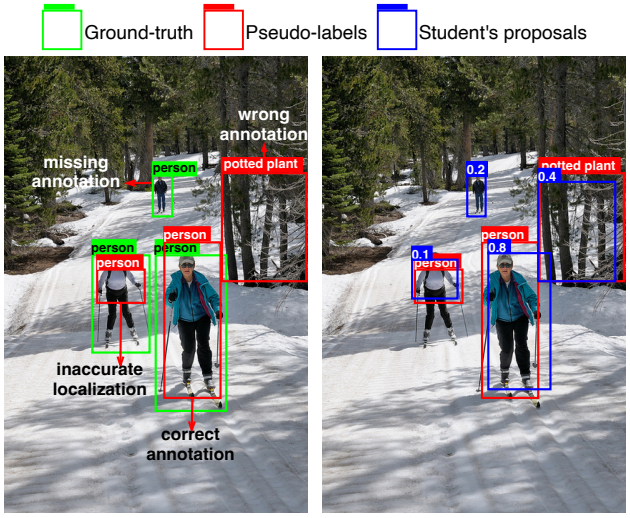
SCMT: Self-Correction Mean Teacher for Semi-supervised Object Detection

Feng Xiong, **Jiayi Tian**, **Zhihui Hao**, **Yulin He** and **Xiaofeng Ren**
Alibaba Group
{xf250971, tianjiayi.tjy, hzh106945, harrylin95, x.ren}@alibaba-inc.com,

Abstract

Semi-Supervised Object Detection (SSOD) aims to improve performance by leveraging a large amount of unlabeled data. Existing works usually adopt the teacher-student framework to enforce student to learn consistent predictions over the pseudo-labels generated by teacher. However, the performance of the student model is limited since the noise inherently exists in pseudo-labels. In this paper, we investigate the causes and effects of noisy pseudo-labels and propose a simple yet effective approach denoted as **S**elf-**C**orrection **M**ean **T**eacher (SCMT) to reduce the adverse effects. Specifically, we propose to dynamically re-weight the unsupervised loss of each student's proposal with additional supervision information from the teacher model, and assign smaller loss weights to possible noisy proposals. Extensive experiments on MS-COCO benchmark have shown the superiority of our proposed SCMT, which can significantly improve the supervised baseline by more than 11% mAP under all 1%, 5% and 10% COCO-*standard* settings, and surpasses state-of-the-art methods by about 1.5% mAP. Even under the challenging COCO-*additional* setting, SCMT still improves the supervised baseline by 4.9% mAP, and significantly outperforms previous methods by 1.2% mAP, achieving a new state-of-the-art performance.



(a) (b)

Figure 1: The green and the red rectangles denote the ground-truth and the corresponding pseudo annotations from the teacher model, while the blue rectangles and their captions denote the bounding box positions of the student's proposals and their self-correction weights. (a) Illustration of the existing SSOD methods based on pseudo-labels. Missing or wrong annotations from the teacher model exist in the pseudo-labelling process. (b) Illustration of our SCMT. The adverse effects of wrong pseudo-labels will be partially eliminated by the self-correction weights based on the teacher model.

1 Introduction

Supervised object detection requires annotations with accurate class labels and bounding box coordinates, which is expensive and time-consuming. Semi-supervised learning is one of the most dominant approaches for solving this issue, which leverages the large-scale unlabeled data to improve the performance of the detector trained on a small amount of labeled data. In this work, we focus on semi-supervised object detection to reduce the efforts of box-level annotations while improving the performance of supervised object detection.

Recently, most semi-supervised object detection methods are based on pseudo-labeling strategy [Sohn *et al.*, 2020c; Liu *et al.*, 2021], where a teacher-student framework are utilized. Specifically, a teacher model is firstly trained on the labeled data and then used to generate pseudo-labels for the unlabeled data. After that, the unlabeled data with pseudo-labels is combined with labeled data to train the student model. In order to obtain stable and reliable pseudo-labels, the exponential moving average (EMA) [Liu *et al.*, 2021; Xu *et al.*, 2021; Tang *et al.*, 2021] is adopted to gradually update the teacher model. However, the final performance is still limited since noisy pseudo-labels inherently exist in the learning process.

To improve the performance of semi-supervised object detection, it is natural to ask: *what causes the inherently noisy pseudo-labels in the teacher-student framework?* Note that in this framework, the pseudo-labels are selected from the prediction results of teacher based on their corresponding classification score with a single threshold. In semi-

supervised classification, a higher filtering threshold is commonly adopted to ensure high-quality pseudo-labels. However, it is not applicable in SSOD for two reasons: (**i**) The single threshold cannot guarantee both precision and recall for foreground pseudo-labels. In object detection, a training image usually contains multiple foreground instances, and the detector needs to distinguish the foreground from the background and classify the foreground into specific categories. A higher threshold can guarantee the high precision of foreground pseudo-labels, but introduce missing annotations for some foreground on the same image. Conversely, a lower threshold will alleviate the issue of missing annotations, but more wrong annotations will be introduced simultaneously. Therefore, no matter how the single threshold is set, the noise problem cannot be avoided in pseudo-labeling. (**ii**) The classification score cannot reflect the quality of bounding box coordinates. Object detection is formulated as a multi-task learning problem involving both classification and localization, but the objectives of these two tasks are different. Specifically, the proposals are treated equally in classification learning if their IoUs with any ground-truth box are larger than a lower bound threshold, while the localization learning seeks for the proposal that can best fit the ground-truth bounding box. This phenomenon is referred to as misalignment between classification and localization in [Jiang *et al.*, 2018], where the bounding boxes with higher classification scores may have poor localization quality, leading to the inaccurate bounding box coordinates in pseudo-labels. As a result, only relying on a high threshold to select pseudo-labels will inevitably cause the existence of noise in SSOD, as shown in Figure 1a.

In this paper, we propose a novel **S**elf-**C**orrection **M**ean **T**eacher (SCMT) framework to alleviate the noisy pseudo-labels problem in SSOD. The motivation is intuitive, which is to dynamically adjust the loss weights of box candidates in student's classification and regression learning, and assign smaller loss weights to possible noisy box candidates. The dynamic weights are illustrated in Figure 1b. We observe that the network can easily learn simple box candidates with high confidence, and tend to make uncertain predictions on difficult and noisy box candidates. In light of this, the network outputs are used as proxy indicators to estimate confidence scores for box candidates. Specifically, for each student-generated box candidate, we define the confidence score as a combination of its corresponding localization accuracy and classification score from teacher's network outputs. The confidence score is served as the dynamic loss weight for student-generated box candidates, where the box candidate with a higher confidence score will have a greater loss weight. As the confidence score is generated by teacher and can be regarded as the self-correction value of teacher-generated pseudo-labels, we name our method as Self-Correction Mean Teacher.

We highlight the contributions of this paper as follows:

- For semi-supervised object detection, we first investigate the limitations of existing methods based on teacher-student framework, where the noise inherently exist in pseudo-labels, and then we propose a simple yet effective **S**elf-**C**orrection **M**ean **T**eacher (SCMT) framework.
- In SCMT, the losses of student-generated box candidates are dynamically re-weighted with confidence scores estimated by the teacher, which aims to reduce the adverse effects of noisy box candidates.
- Our SCMT achieves state-of-the-art performance on the MS-COCO dataset under both the COCO-*standard* and COCO-*additional* settings, which demonstrates the effectiveness of our proposed framework.

2 Related Work

Semi-supervised Learning in Classification. Significant progress has been made in semi-supervised image classification. The majority of the recent work can be roughly categorized into pseudo-labeling and consistency regularization. The main objective of pseudo-labeling methods is to generate pseudo-labels for unlabeled data with a model trained on labeled data. [Lee and others, 2013] firstly introduces pseudo-labeling in semi-supervised learning, the pseudo-labels are generated from predictions of a trained neural network. Recently, Noisy Student [Xie *et al.*, 2020] demonstrates that pseudo-labeling can further improve the performance on ImageNet with an additional large amount of unlabeled data. Consistency regularization methods [Sajjadi *et al.*, 2016; Tarvainen and Valpola, 2017] utilize the assumption that no matter what perturbations are applied, the model should keep the consistent output for the same unlabeled data. [Sajjadi *et al.*, 2016] introduces random dropout and random data augmentation as input perturbations. Several approaches combine ideas from both pseudo-labeling and consistency regularization [Berthelot *et al.*, 2019b; Berthelot *et al.*, 2019a; Sohn *et al.*, 2020a]. In [Sohn *et al.*, 2020a], pseudo-labels are generated by the teacher training on weakly augmented images, and the student trains on strongly augmented images to learn consistent predictions over pseudo-labels.

Semi-supervised Learning in Object Detection. Currently, SSOD has developed rapidly, inspired by semi-supervised image classification. State-of-the-art results of SSOD have been established by combining pseudo-labeling and consistency regularization. STAC [Sohn *et al.*, 2020c] proposes to use weakly augmented images for pseudo-labeling and strongly augmented images for model consistency training. However, the pseudo-labels are generated only once and are fixed throughout the rest of training, which seriously limits the performance of SSOD. To alleviate this issue, [Liu *et al.*, 2021; Xu *et al.*, 2021; Tang *et al.*, 2021] introduce mean teacher to generate pseudo-labels from teacher and train the student simultaneously, while the teacher is gradually updated by the student. The quality of pseudo-labels is crucial to pseudo-labeling methods. Unbiased-Teacher [Liu *et al.*, 2021] introduces a class-balance loss to address the pseudo-labeling bias problem caused by class-imbalance. The current state-of-the-art method Softer-Teacher [Xu *et al.*, 2021] adopts a high filtering threshold to guarantee a high precision of the positive pseudo-labels, and proposes using a reliability measure to weight the loss of each background box candidate to address the problem of missing annotations. When compared with Softer Teacher, we propose to re-weight losses for both positive and negative samples from both RPN and

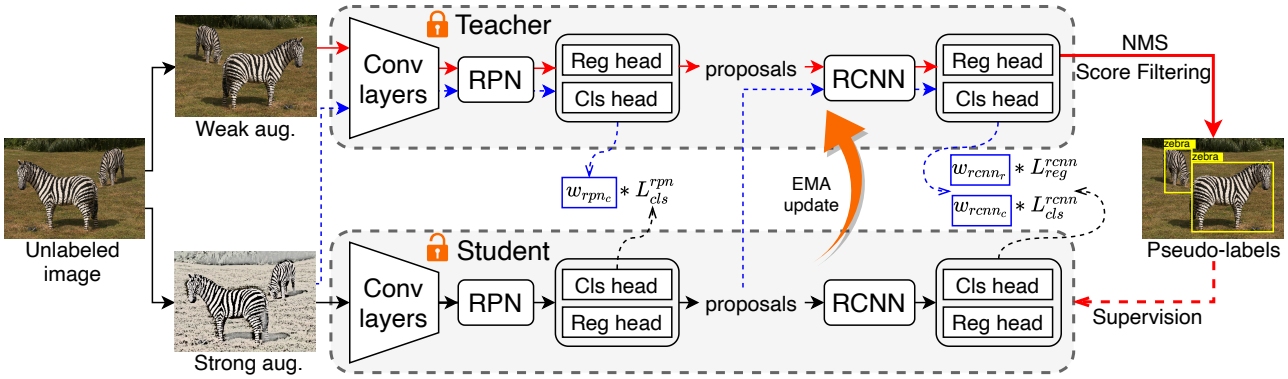


Figure 2: Overview of our proposed Self-Correction Mean Teacher framework. We generate the pseudo-labels by using teacher model for each unlabeled data based on its weak augmented version. Then, the unsupervised loss is computed according to the pseudo-labels with the novel proposed self-correction weights. We denote the blue dash line as the process of generating self-correction weights.

RCNN stages, where the loss weight is dynamically obtained based on the classification score and localization accuracy.

3 Method

For semi-supervised object detection, we are given a set of labeled data $\mathcal{D}_s = \{\mathbf{x}_i^s, \mathbf{y}_i^s\}_{i=1}^{N_s}$ and a set of unlabeled data $\mathcal{D}_u = \{\mathbf{x}_i^u\}_{i=1}^{N_u}$, where $\mathbf{x}_i$ and $\mathbf{y}_i$ denote the i-th sample and its corresponding label, N_s and N_u represent the numbers of supervised and unsupervised samples. Our goal is to leverage the unlabeled data to exceed the performance of models trained only using labeled data. For the rest of this section, we first review a widely used teacher-student framework, then introduce our proposed improvement based on this framework.

3.1 Teacher-Student Framework

Teacher-student is a typical semi-supervised object detection framework, which includes two networks: teacher network and student network. Both networks use the same architecture, but are in charge for different mission.

Teacher network is firstly obtained by training on labeled data $\mathcal{D}_s$ using supervised loss $\mathcal{L}_s$ as in Faster RCNN [Ren *et al.*, 2015], which is constructed by four components: the RPN classification loss L_{cls}^{rpn}, the RPN regression loss L_{reg}^{rpn}, the RCNN classification loss L_{cls}^{roi} and the RCNN regression loss L_{reg}^{roi},

$$\mathcal{L}_s = \frac{1}{N_s}\sum_{i=1}^{N_s}(L_{cls}^{rpn}(\mathbf{x}_i^s, \mathbf{y}_i^s) + L_{reg}^{rpn}(\mathbf{x}_i^s, \mathbf{y}_i^s) + L_{cls}^{rcnn}(\mathbf{x}_i^s, \mathbf{y}_i^s) + L_{reg}^{rcnn}(\mathbf{x}_i^s, \mathbf{y}_i^s)). \tag{1}$$

Then, the parameters of student network are initialed by the parameters of the teacher network. To address the lack of annotations for unsupervised data, motivated by FixMatch [Sohn *et al.*, 2020b], a weak-strong data augmentation strategy and pseudo-labeling method are adopted. Specifically, given an unlabeled sample $\mathbf{x}_i^u$, the prediction result of teacher model are first obtained based on a weakly augmented version of $\mathbf{x}_i^u$, then after applying non-maximum suppression (NMS) and confidence-based box filtering, we acquire the pseudo annotation $\hat{\mathbf{y}}_i^u$ for $\mathbf{x}_i^u$. Using $\hat{\mathbf{y}}_i^u$ as pseudo-labels, we compute the classification loss and regression loss based on the student's output for a strongly augmented version of $\mathbf{x}_i^u$. The unsupervised loss in this step is computed as follows:

$$\mathcal{L}_u = \frac{1}{N_u}\sum_{i=1}^{N_u}(L_{cls}^{rpn}(\mathbf{x}_i^u, \hat{\mathbf{y}}_i^u) + L_{reg}^{rpn}(\mathbf{x}_i^u, \hat{\mathbf{y}}_i^u) + L_{cls}^{rcnn}(\mathbf{x}_i^u, \hat{\mathbf{y}}_i^u) + L_{reg}^{rcnn}(\mathbf{x}_i^u, \hat{\mathbf{y}}_i^u)). \tag{2}$$

Finally, the objective function of the student network is composed of a supervised loss $\mathcal{L}_s$ applied on labeled data and an unsupervised loss $\mathcal{L}_u$ applied on unlabeled data, which can be written as follows:

$$\mathcal{L} = \mathcal{L}_s + \lambda_u \mathcal{L}_u, \tag{3}$$

where λ_u is a fixed scalar to control the influence of the unsupervised loss, and we set $\lambda_u = 4$ for all experiments setting.

In teacher-student framework, the parameters of teacher is obtained using the supervised loss as Eq. (1) and fixed for the training of student.To obtain more stable pseudo-labels, [Liu *et al.*, 2021] introduces the **mean teacher** framework to further gradually refine the parameters of teacher network θ_t as an exponential moving average(EMA) of the parameters of student network θ_s,

$$\theta_t \leftarrow \alpha\theta_t + (1-\alpha)\theta_s, \tag{4}$$

where α is a smoothing coefficient hyper-parameter, which is usually set to 0.999.

3.2 Self-Correction Mean Teacher Framework

The quality of pseudo-labels is important for mean teacher based semi-supervised object detection. However, as discussed in Section 1, noise inherently exists in pseudo-labels, which may not only lead to poor classification generalization capability, but also harm the bounding box regression ability. In this paper, we propose SCMT to deal with the above issues from both classification and regression learning, the overall framework is illustrated in Figure 2.

Self-Correction Weight for Noise-aware Classification. For two-stage object detector, the classification learning is involved in both RPN and RCNN stages, where the box candidates are assigned labels according to their overlaps with

ground-truth labels. As discussed in Section 1, the noisy pseudo-labels are inevitable in SSOD, which will cause the box candidates being assigned with wrong labels and harm the classification learning.

Therefore, we propose to reduce the impact of box candidates with wrong labels by dynamically re-weighting their classification losses, where we introduce a confidence score to quantify the importance of candidate boxes by taking into account the classification score as well as the localization accuracy. In specific, for student-generated box candidates classification, the prediction results with high confidence of these candidates are likely to be correct, and those with low confidence are usually confused and uncertain ones. While for student-generated box candidate localization, the localization accuracy can be represented as IoU between the region proposal and its corresponding pseudo bounding box, where the box candidate with higher IoU is more accurate. Thus, we propose to define the confidence score of each box candidate as the product of its classification score and localization accuracy. Motivated by this observation, for each student-generated box candidate, we use its corresponding confidence score from teacher as the dynamic loss weight, and greater loss weights will be assigned to box candidates with high confidence. Here we use teacher to estimate confidence scores of box candidates, because the prediction results of teacher are more stable than those of student. Contrary to the widely used hard sample mining approaches, e.g. Focal Loss [Lin *et al.*, 2017b], where the loss is designed to down-weight the contributions of the well-classified samples. The proposed self-correction weight pays more attention to the learning of convincing samples, as we argue that the convincing samples are usually the important samples in classification. The self-correction weight can be easily obtained based on the mean teacher framework. However, in the RPN and RCNN stage, the processes of generating weights are different, and the details are described below.

In the RPN stage, the strongly augmented unlabeled image used in student is simultaneously put through the teacher's RPN head to obtain the classification score and box coordinates of each anchor's corresponding region proposal, as shown in Figure 2. Note that the anchor set is the same for the teacher and the student. Therefore, the RPN classification self-correction weight of each student's anchor can be formulated as:

$$w_{rpn_c} = \begin{cases} \frac{1}{1+e^{-p_i}} \cdot IoU(\hat{a}_i, g_i), & \text{for } a_i \in \mathcal{A}_{pos} \\ 1 - \frac{1}{1+e^{-p_i}}, & \text{for } a_i \in \mathcal{A}_{neg} \end{cases} \tag{5}$$

Here, a_i and p_i denote the i-th candidate anchor and its logit output generated by teacher, respectively. $\mathcal{A}_{pos}$ and $\mathcal{A}_{neg}$ represent positive and negative candidate anchor sets, separately. Besides, $IoU(\hat{a}_i, g_i)$ denotes the IoU between the region proposal $\hat{a}_i$ from the teacher with respect to anchor a_i and its corresponding pseudo bounding box g_i.

In the RCNN stage, each student-generated region proposal is also input into the teacher's RCNN head to obtain the classification score and box candidates of the refined region proposal. Since the RCNN needs to classify region proposals into specific categories, which is a multi-class classification task. For each region proposal, the classification score under the corresponding category based on the assigned pseudo-labels is used to calculate the self-correcting weight. We formulate the RCNN classification self-correction weight as:

$$w_{rcnn_c} = \begin{cases} \frac{e^{p_j^c}}{\sum_{k=1}^{C+1} e^{p_j^k}} \cdot IoU(\hat{b}_j^c, g_j), & \text{for } b_j^c \in \mathcal{B}_{pos} \\ \frac{e^{p_j^c}}{\sum_{k=1}^{C+1} e^{p_j^k}}, & \text{for } b_j^c \in \mathcal{B}_{neg} \end{cases} \tag{6}$$

Here b_j^c is the j-th candidate proposal assigned to the c-th category based on pseudo-labels, p_j^c denotes the logit output of the proposal with respect to the c-th category, which is generated by teacher. $\mathcal{B}_{pos}$ and $\mathcal{B}_{neg}$ represent positive and negative candidate proposal sets, separately. Besides, C represents the total number of foreground categories, and $C+1$-th class denotes the background category. In addition, $IoU(\hat{b}_j^c, g_j)$ denotes the IoU between the refined region proposal $\hat{b}_j^c$ from the teacher with respect to b_j^c and its corresponding pseudo bounding box g_j.

Self-Correction Weight for Classification-aware Regression. As same in classification, regression learning is also needed in both RPN and RCNN stages, but the regression learning in RCNN stage takes a more important role than in RPN stage. The difference from classification is that the regression learning is performed only for positive box candidates with foreground categories. We argue that wrong labels of box candidates will also damage the localization ability, because they may lead to unstable regression loss. For example, given an actual background box candidate, which is assigned as foreground class by incorrect pseudo-labels, the regression branch will continue to learn offsets of this candidate mistakenly.

To tackle the problem of unstable regression, we introduce self-correction weight to adjust loss weights of the noisy box candidates in the regression learning. The self-correction weight is only adopted in the RCNN stage, since the fine regression learning in the RCNN stage is more important for accurate localization. The RCNN regression self-correction weight is the same as the RCNN classification self-correction weight, and can be formulated as:

$$w_{rcnn_r} = w_{rcnn_c}, \text{for } b_j^c \in \mathcal{B}_{pos}. \tag{7}$$

The proposed RCNN regression self-correction weight can not only alleviate the impact of noisy pseudo-labels in the regression learning, but also promote the consistency of regression and classification learning, as the regression branch allocates more attention for the candidate proposals with high classification scores.

Overall Loss Optimization. In the self-correction mean teacher framework, the overall loss is composed of supervised loss from the labeled images and unsupervised loss from the unlabeled images. The supervised loss is the same as which in the teacher-student framework. The unsupervised loss is modified with the self-correction weight, and can be formulated as:

$$\begin{aligned} \mathcal{L}_u = \frac{1}{N_u} \sum_{i=1}^{N_u} (& w_{rpn_c} L_{cls}^{rpn}(\mathbf{x}_i^u, \hat{\mathbf{y}}_i^u) + L_{reg}^{rpn}(\mathbf{x}_i^u, \hat{\mathbf{y}}_i^u) \\ & + w_{rcnn_c} L_{cls}^{rcnn}(\mathbf{x}_i^u, \hat{\mathbf{y}}_i^u) + w_{rcnn_r} L_{reg}^{rcnn}(\mathbf{x}_i^u, \hat{\mathbf{y}}_i^u)). \end{aligned} \tag{8}$$

Method \ Setting	COCO-*standard*			COCO-*additional*
	1%	5%	10%	100%
Supervised Baseline(Ours)	9.05±0.16	18.47±0.22	23.86±0.81	37.63
CSD [Jeong *et al.*, 2019]	10.20± 0.15 (+1.15)	18.90± 0.10(+0.43)	24.50± 0.15(+0.64)	38.82(+1.19)
STAC [Sohn *et al.*, 2020c]	13.97 ± 0.35(+4.92)	24.38 ± 0.12(+5.91)	28.64 ± 0.21(+4.78)	39.21(+1.58)
Instant-Teaching [Zhou *et al.*, 2021]	18.05± 0.15(+9.00)	26.75± 0.05(+8.28)	30.40± 0.05(+6.54)	40.20(+2.57)
Unbiased-Teacher [Liu *et al.*, 2021]	20.75± 0.12(+11.70)	28.27± 0.11(+9.80)	31.50± 0.10(+7.64)	41.30(+3.67)
Softer-Teacher [Xu *et al.*, 2021]	20.46± 0.39(+11.41)	30.74± 0.08(+12.27)	34.04± 0.14(+10.18)	41.40(+3.77)
SCMT(Ours)	**23.09± 0.16(+14.04)**	**32.14± 0.06(+13.67)**	**35.42± 0.12(+11.56)**	**42.56(+4.93)**

Table 1: Comparison of mAP for different semi-supervised methods on MS-COCO benchmark. The Softer-Teacher's **COCO-*additional*** result is reported with a total training step of 360k as re-implemented based on ours training schedule.

4 Experiments

4.1 Dataset and Evaluation

We evaluate our approach on the large-scale dataset MS-COCO [Lin *et al.*, 2014]. For MS-COCO, we use version 2017 in all experiments, including *train2017*, *unlabeled2017* and *val2017*. *train2017* set has a total of 118k labeled images, *unlabeled2017* set contains 123k unlabeled images and *val2017* set has 5k images in total. For a fair comparison, we follow STAC to validate the performance on COCO-*standard* and COCO-*additional* settings. For COCO-*standard* setting, we set up three different proportions: 1%, 5%, and 10% to sample images from *train2017* as the labeled training data, and the remaining unsampled images are used as the unlabeled data. For COCO-*additional* setting, we use the fully *train2017* as the labeled data and the additional *unlabeled2017* as the unlabeled data. Besides, *val2017* set is used to evaluate our approach on both settings.

4.2 Implementation Details

We implement our SCMT framework based on the MMDetection toolbox [Chen *et al.*, 2019]. To get a fair comparison, we follow STAC to use Faster R-CNN with ResNet-50 backbone and FPN [Lin *et al.*, 2017a] as the default base model. The filtering threshold is one of the main hyper-parameters in our SCMT framework, and is set to 0.7 throughout our experiments. Since the amount of training data of COCO-*standard* setting and COCO-*additional* setting has large differences, the training schedules are slightly different. For both settings, the initial learning rate is set to 0.01. For COCO-*standard* setting, the learning rate decays by 10× at 120k and 160k, with a total training step of 180k. For COCO-*additional* setting, the learning rate decays by 10× at 240k and 320k, with a total training step of 360k. All models are trained with batch size of 32, and the other training hyper-parameters are the same as standard Faster R-CNN. As for the data augmentation, we follow Softer-Teacher to use different data augmentation for labeled image training and unlabeled image training. Specifically, we apply scale jitter and color jitter for weak augmentation and additional geometric augmentations and cutout patches for strong augmentations.

4.3 Overall Performance

In this section, we compare our approach with previous state-of-the-art SSOD methods. As shown in Table 1, our SCMT performs favorably against other methods under all settings, including various COCO-*standard* setting and COCO-*additional* setting. Specifically, for the 1% protocol, SCMT improves the mAP of Softer-Teacher from 20.46% to 23.09%, which achieves 2.63% mAP improvement; for the 10% protocol, SCMT improves the mAP of Softer-Teacher from 34.04% to 35.42%, which achieves 1.38% mAP improvement. It is worth noting that our SCMT trained on 5% labeled data achieves 32.14% mAP, which is even higher than most of other SSOD methods trained on 10% labeled data. SCMT consistently outperforms Supervised Baseline by more than 11% absolute mAP under all 1%, 5% and 10% protocols, and as there is less labeled data, the improvement becomes greater. COCO-*additional* setting is more challenging. Our SCMT still achieves 4.93% improvement in mAP over Supervised Baseline and about 1.16% mAP improvement under the high benchmark (Softer-Teacher) of 41.40% mAP. The result demonstrates that our proposed SCMT can further improve the performance of a well-trained object detector on a large-scale labeled dataset by using more unlabeled data.

w_{rpn_c}	w_{rcnn_c}	w_{rcnn_r}	mAP	AP@0.5
			31.9	52.5
✓			32.3	53.4
	✓		32.9	54.3
✓	✓		33.4	54.7
✓	✓	✓	33.8	54.9

Table 2: Ablation studies on the effectiveness of different components, w_{rpn_c}, w_{rcnn_c} and w_{rcnn_r} represent the self-correction weights of RPN classification, RCNN classification and RCNN regression, respectively. The first row denotes the results of semi-supervised baseline based on the Mean Teacher framework.

4.4 Ablation Studies

We validate the effectiveness of our key designs in this section. All the ablation experiments are performed on COCO-*standard* setting with 10% labeled data. Due to the limitation of machine resources, we train the model with a total training step of 90k, and the learning rate decays by 10× at 60k and 80k, which is different from the setting used in Table 1.

Effectiveness of different components: We ablate each component of the proposed SCMT step by step. Table 2 reports the results. The basic framework Mean Teacher de-

Method	AR_s@1000	AR_m@1000	AR_l@1000
w/o w_{rpn_c}	31.9	54.3	68.9
w_{rpn_c}	32.7	54.6	69.1

Table 3: Results of RPN region proposals measured by Average Recall (AR), and AR_s, AR_m and AR_l represents the AR of small size objects, meduim size objects and large size objects respectively. w/o w_{rpn_c} denotes without the RPN classification self-correction weight.

Threshold	mAP	AP@0.5	AP@0.75	AP@0.9
0.9	32.5	52.6	34.7	7.3
0.8	33.0	53.8	35.3	7.0
0.7	33.8	54.9	35.6	7.5
0.6	33.2	54.8	34.9	6.8

Table 4: Ablation study on the effects of different thresholds.

scribed in Section 3.1 is adopted as our semi-supervised baseline, which achieves 31.9% mAP. As shown in Table 2, the RPN classification self-correction weight and RCNN classification self-correction weight improve the mAP of baseline by 0.4% and 1.0%, respectively. Furthermore, the combination of these two weights greatly outperforms the baseline by 1.5% mAP. This gain mainly comes from reducing the classification problems caused by noisy pseudo-labels, including missing detection and background false detection. Besides, when applying the RCNN regression self-correction weight, the performance reaches 33.8% mAP, which improves the baseline by 1.9%. We further investigate the effects of RPN classification self-correction weight. From Table 3, we observe that the RPN classification self-correction weight makes an improvement of 0.8%, 0.3% and 0.2% for AR_s, AR_m and AR_l , respectively. Note that larger improvements are achieved for AR_s, because the poor features of small objects cause more noisy pseudo-labels, and the self-correction weight can avoid the impact of this issue.

Effects of other hyper-parameters: We study the effects of different filtering thresholds for generating pseudo-labels. Table 4 reports the results. The best performance is achieved when the filtering threshold is set to 0.7. And we find that a too high threshold is not the optimal choice, which may introduce more missing annotations in pseudo-labels, resulting in hindering the training and limiting the performance of the object detector.

4.5 Further Investigation

We further investigate the superiority of SCMT over other methods by comparing the error type analysis generated by the TIDE toolbox [Bolya *et al.*, 2020] in Figure 3, and the qualitative results of GradCAM [Selvaraju *et al.*, 2017] in Figure 4. The analysis reveals more granular information about where SCMT improves mAP relative to other methods. We observe that the Supervised-Baseline has a very serious miss error problem, and the learned features cannot focus on the foreground area, which indicates the difficulty of object detection with a small amount of labeled data. Surprisingly, SCMT significantly closes the gap between semi-supervised object detection and fully supervised object detection. Both the classification error and localization error can be well solved, thus leading to better mAP. Besides, we find that SCMT concentrates more on the foreground area, which can encourage the network to extract informative and meaningful features as shown in Figure 4.

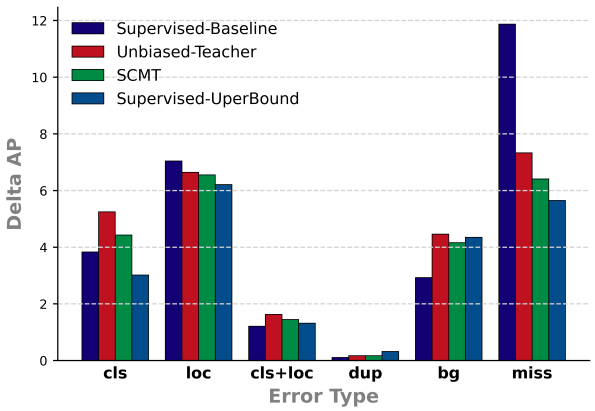


Figure 3: TIDE error analysis. We plot the Delta AP metric at a IoU threshold of 0.5. Each bar shows how much AP can be added to the detector if a certain error type is fixed. The error types are same as in [Bolya *et al.*, 2020]. Supervised-UperBound represents the fully supervised object detection with 100% labeled data.

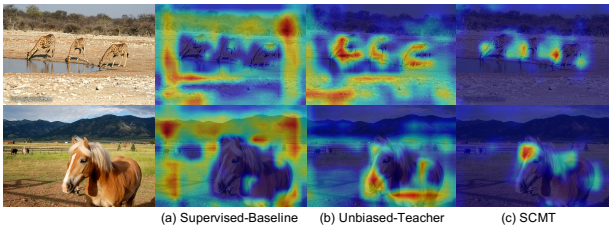


Figure 4: Class-Activation-Map (CAM) illustration of different methods, whose experimental settings are the same as those in Figure 3.

5 Conclusion

In this paper, we propose a simple yet effective training framework for semi-supervised object detection, which can dynamically re-weight the unsupervised loss to mitigate the effects of noisy pseudo-labels in both classification and regression. Based on the teacher-student framework in SSOD, we first generate the pseudo-labels by teacher model, then obtain the confidence score combining classification score and localization accuracy from teacher model for each student's proposal as the dynamic loss weight, which is referred to as self-correction weight. Namely, we use the loss weights provided by the teacher model to alleviate the side effect of the noisy pseudo-labels from the teacher model. The proposed method outperforms state-of-the-art methods by a large margin on MS-COCO benchmark under both partially labeled data and fully labeled data settings.

References

[Berthelot *et al.*, 2019a] 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.

[Berthelot *et al.*, 2019b] 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.

[Bolya *et al.*, 2020] Daniel Bolya, Sean Foley, James Hays, and Judy Hoffman. Tide: A general toolbox for identifying object detection errors. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 558–573. Springer, 2020.

[Chen *et al.*, 2019] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

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

[Jiang *et al.*, 2018] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 784–799, 2018.

[Lee and others, 2013] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013.

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

[Lin *et al.*, 2017a] Tsung-Yi Lin, Piotr Dollar, 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 (CVPR)*, July 2017.

[Lin *et al.*, 2017b] 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.

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

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

[Sajjadi *et al.*, 2016] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29:1163–1171, 2016.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[Sohn *et al.*, 2020a] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

[Sohn *et al.*, 2020b] 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.

[Sohn *et al.*, 2020c] 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.

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

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

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

[Xu *et al.*, 2021] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. *arXiv preprint arXiv:2106.09018*, 2021.

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