
CONSISTENT TARGETS PROVIDE BETTER SUPERVISION IN SEMI-SUPERVISED OBJECT DETECTION

Xinjiang Wang^{1*}, Xingyi Yang^{3*‡}, Shilong Zhang², Yijiang Li^{1‡},
 Litong Feng¹, Shijie Fang^{4‡}, Chengqi Lyu², Kai Chen^{1,2}, Wayne Zhang¹

¹Sensetime Research

²Shanghai AI Lab

³National University of Singapore

⁴Peking University

wangxinjiang@sensetime.com, xyang@u.nus.edu

October 3, 2022

ABSTRACT

In this study, we dive deep into the inconsistency of pseudo targets in semi-supervised object detection (SSOD). Our core observation is that the oscillating pseudo targets undermine the training of an accurate semi-supervised detector. It not only injects noise into student training but also leads to severe overfitting on the classification task. Therefore, we propose a systematic solution, termed **Consistent-Teacher** , to reduce the inconsistency. First, adaptive anchor assignment (ASA) substitutes the static IoU-based strategy, which enables the student network to be resistant to noisy pseudo bounding boxes; Then we calibrate the subtask predictions by designing a 3D feature alignment module (FAM-3D). It allows each classification feature to adaptively query the optimal feature vector for the regression task at arbitrary scales and locations. Lastly, a Gaussian Mixture Model (GMM) dynamically revises the score threshold of the pseudo-bboxes, which stabilizes the number of ground-truths at an early stage and remedies the unreliable supervision signal during training. **Consistent-Teacher** provides strong results on a large range of SSOD evaluations. It achieves 40.0 mAP with ResNet-50 backbone given only 10% of annotated MS-COCO data, which surpasses previous baselines using pseudo labels by around 3 mAP. When trained on fully annotated MS-COCO with additional unlabeled data, the performance further increases to 47.2 mAP. Our code will be open-sourced soon.

1 Introduction

The goal of semi-supervised object detection (SSOD) [1, 2, 3, 4, 5, 4, 6, 7, 8, 9, 10, 11] is to facilitate the training of object detectors with the help of a large amount of unlabeled data. The common practice is first to train a teacher model on the labeled data and then generate pseudo labels and boxes on unlabeled sets, which act as the ground truth (GT) for the student model. Student detectors, on the other hand, are anticipated to make consistent predictions regardless of network stochasticity [12] or data augmentation [4, 3]. In addition, to improve pseudo-label quality, the teacher model is updated as a moving average [5, 6, 7] of the student parameters.

However, the performance of SSOD still lags far behind the current development of its fully-supervised counterpart. In this study, we point out that the inconsistency of pseudo-targets is one major factor that hinders the performance of semi-supervised detectors. **Inconsistency** refers to the fact that the pseudo boxes may be highly inaccurate and vary greatly at different stages of training. The oscillating bounding boxes would bias the student’s predictions with accumulated error [13]. In contrast to labels in semi-supervised classification, SSOD assigns a set of dense tuples of classification and regression targets (c_i, bbox_i) to each RoI/anchor (i). Thus, a small perturbation in the teacher’s

*Equally contributed.

‡Work done during internship at SenseTime.

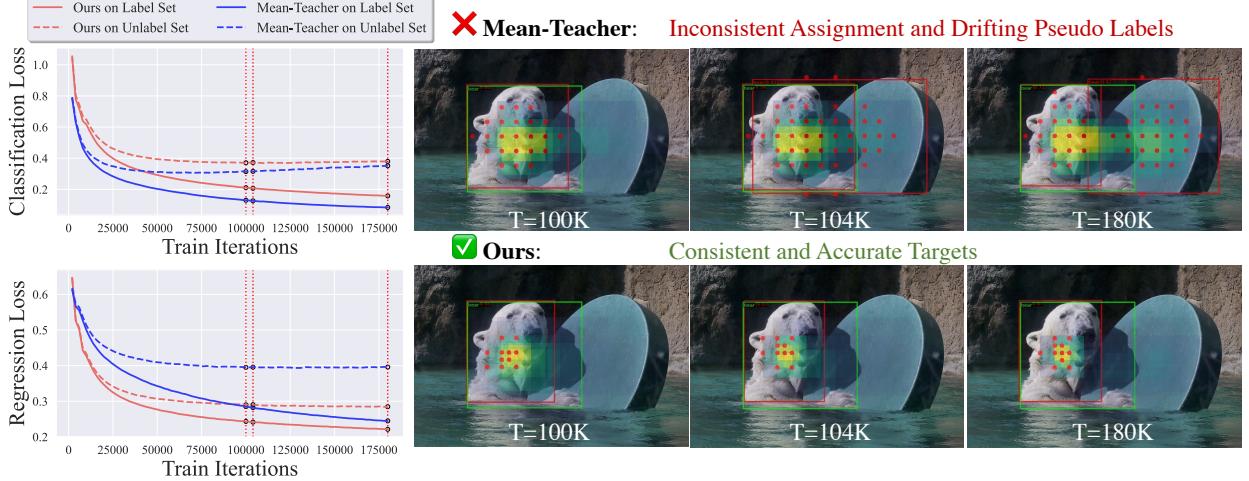


Figure 1: Illustration of inconsistency problem in SSOD on COCO 10% evaluation. (Left) We compare the training losses between the Mean-Teacher and our **Consistent-Teacher**. In Mean-Teacher, inconsistent pseudo targets lead to overfitting on the classification branch, while regression losses become difficult to converge. In contrast, our approach sets consistent optimization objectives for the students, effectively balancing the two tasks and preventing overfitting. (Right) Snapshots for the dynamics of pseudo labels and assignment. The **Green** and **Red** bboxes refer to the ground-truth and pseudo bbox for the polar bear. **Red dots** are the assigned anchor boxes for the pseudo label. The heatmap indicates the dense confidence score predicted by the teacher (brighter the larger). The nearby board is finally misclassified as a polar bear in the baseline while our adaptive assignment prevents overfitting.

prediction could also affect the assignment results and the regression target dramatically. In general, inconsistent targets inject substantial noise into the student network and may even lead to severe overfitting on unlabeled images.

Common two-stage [3, 5, 6] and single-stage [14, 15] SSOD networks adopt static criteria for anchor assignment, e.g. IOU score or centerness. The static assignment methods are sensitive to noise in the bounding boxes (bboxes) predicted by the teacher. To illustrate this, we train a single-stage detector with standard IoU-based assignment on MS-COCO %10 data. As shown in Fig. (1), a small change in the teacher’s output results in strong noise in the boundaries of pseudo-bboxes, which associate the erroneous targets to an unrelated but nearby object under static IoU-based assignment. It is due to the fact that the high-responding anchors in the teacher network may not necessarily be assigned positive in the student network. Consequently, the network overfits as it produces inconsistent label to neighboring objects. The overfitting is also observed in the classification loss curve on unlabeled images¹.

A second cause for inconsistency lies in the tuples of classification and regression labels (c_i , bbox_i) in SSOD. Typically, only the classification score is used to filter the pseudo labels. However, the classification score does not necessarily reflect the quality of its bbox [6]. The misalignment between c_i and bbox_i also accounts for the oscillation in the pseudo-bbox boundaries, which further exacerbates the inconsistency caused by static assignment in SSOD.

The hard threshold scheme in common SSOD methods also causes temporal inconsistencies in pseudo labels. Traditional SSOD methods [3, 5, 6] usually adopt a hard threshold on top of the confidence score to distinguish pseudo-bboxes for student training. However, the hard threshold, as a hyper-parameter, not only needs to be carefully tuned for each model-task combination, but should also be dynamically adjusted in accordance with model’s capability at different time-steps. In the Mean-Teacher[16] training paradigm, the number of pseudo-bboxes may increase from too few to too many under hard threshold scheme, which incurs inefficient and biased training for the student.

Therefore, we propose **Consistent-Teacher** in this study to address the inconsistency issues. First, we find that a simple replacement of the static IOU-based anchor assignment by cost-aware adaptive sample assignment (ASA) [17, 18] greatly alleviates the inconsistency in dense pseudo targets. During each training step, we calculate the matching cost between each pseudo-bbox with the student network’s predictions. Only feature points with lowest costs are assigned as positive. It reduces the mismatch between the teacher’s high-response features and the student’s assigned positive pseudo targets, which inhibits overfitting.

¹All GT bboxes on unlabeled data are only used to calculate the loss value but not for updating the parameters.

Then, we calibrate the classification and regression tasks such that high teacher’s classification confidence c_j provides a good proxy of the tuple (c_j, bbox_j) quality, which reduces the oscillation in pseudo-bbox boundaries and make consistent targets for the student network. Inspired by TOOD [19], we propose a 3-D feature alignment module (FAM-3D) that allows classification feature to sense and adopt the best feature in its neighborhood for regression. Different from the single scale searching, FAM-3D reorders the features pyramad for regression across scales as well as locations. In this way, a unified confidence score accurately measures the quality of classification and regression with the improved alignment module, and ultimately brings consistent pseudo-targets for the student in SSOD.

As for the temporal inconsistency in pseudo-bboxes, we apply Gaussian Mixture Model (GMM) to generate an adaptive threshold for each category at training time. We consider the confidence scores of each class as the weighted sum of positive and negative distributions and predict the parameters of each Gaussian with maximum likelihood estimation. It is expected that the model is able to adaptively infer the optimal threshold at different training steps so as to stabilize the number of positive samples.

The proposed **Consistent-Teacher** greatly surpasses current SSOD methods. **Consistent-Teacher** reaches 40.0 mAP with 10% of labeled data on MS-COCO, which is $\tilde{3}$ mAP ahead of the state-of-the-art [11]. When using the 100% labels together with extra unlabeled MS-COCO data, the performance is further boosted to 47.2 mAP. The effectiveness of **Consistent-Teacher** is also testified on other ratios of labeled data and on other datasets as well. Concretely, the paper contributes in the following aspects.

- We provide the first in-depth investigation for the inconsistency target problem in object detection under semi-supervised situation, which incurs severe overfitting issues.
- We introduce the adaptive sample assignment to stabilize the matching between noisy pseudo-bboxes and anchors, leading to robust training objective for the student.
- We develop a 3-D feature alignment module (FAM-3D) to calibrate the classification confidence and regression quality, which helps stabilize pseudo-bbox boundaries of high confidence scores.
- We adopt GMM to flexibly determine the threshold for each class during training. The adaptive threshold evolves through time and reduces the temporal consistencies for SSOD.
- **Consistent-Teacher** achieves compelling improvement on a wide range of evaluations and serves as a new solid baseline for SSOD.

2 Related Work

Semi-supervised object detection (SSOD). It is a common practice for SSOD to generate pseudo bounding boxes using a teacher model and expect the student detectors to make consistent prediction on augmented input samples[4, 3, 20, 5, 6, 7, 21, 22, 23]. Two-stage detectors [4, 5, 6] have been dominant in traditional SSOD methods while single-stage detectors also shown the advantages for its simplicity and higher performance [14, 15, 11]. In this study, we adopt a single-stage framework yet a different path and focus on the inconsistency problem in SSOD. To resolve the inconsistency issues, we design the adaptive anchor assignment, feature alignment and GMM-based threshold to improve the label quality.

Label assignment in object detection. Defining positive and negative sample [24] plays a substantial role in object detection. Typical Anchor-based or anchor-free detectors either adopt hard IoU thresholding [25, 26, 27, 28, 29, 30, 31, 32] or the centerness prior [33, 34, 35] as the assigning criterion. In contrast, modern detectors have been shifting to adaptive assignment strategies. [36, 37, 38, 39, 17] For example, PAA [39] adaptively differentiates the positive anchors and negative ones by fitting the anchor scores distribution. OTA [17] treats the label assignment as an optimal transport problem so that the assignment cost is minimized.

However, these assignment methods are all carried out in fully-supervised settings, whereas we find that the static assignment induces new inconsistency issues and accumulates error in SSOD. We show that a simple cost-aware adaptive assignment stabilize the label noisy and greatly benefits the SSOD task.

3 Consistent-Teacher

In this section, we elaborate on how our **Consistent-Teacher** works to address the SSOD inconsistencies. It is composed of three key modules, namely Adaptive Sample Assignment, 3D Feature Alignment Module and Gaussian Mixture based thresholding. The full pipeline is in Figure 2.

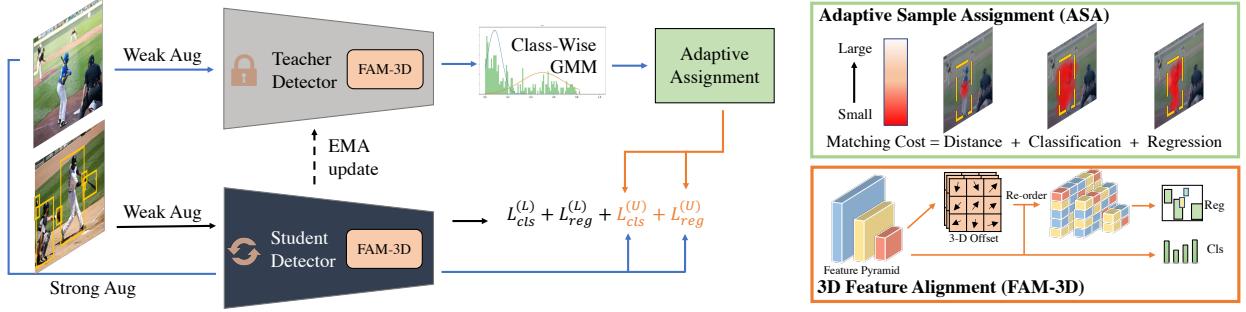


Figure 2: The pipeline of Consistent-Teacher . We design three modules to address the inconsistency in SSOD, where GMM dynamically determines the threshold; 3D feature alignment calibrates regression quality; Adaptive assignment assigns anchor based on matching cost.

3.1 Baseline SSOD Detector

We adopt a general SSOD paradigm as our baseline, namely a Mean-Teacher [5, 6, 16] pipeline with a RetinaNet [28] object detector. The teacher model is an exponential moving average [16] of a student network. Unlabeled images first go through weak augmentations and are fed into the teacher detector to generate pseudo bboxes. Pseudo-bboxes are then used as supervision for the student network, whose unlabeled images are strongly jittered. In the meantime, student detector takes the labeled images as input to learn discriminative representation for both classification and regression. Given a labeled set $\mathcal{D}_L = \{\mathbf{x}_i^l, \mathbf{y}_i^l\}^N$ with N samples and an unlabeled set $\mathcal{D}_U = \{\mathbf{x}_j^u\}^M$ with M samples, we maintain a teacher detector $f_t(\cdot; \Theta_t)$ and a student detector $f_s(\cdot; \Theta_s)$ that minimize the loss

$$\begin{aligned} \mathcal{L} = & \frac{1}{N} \sum_i \left[\mathcal{L}_{cls}(f_s(T(\mathbf{x}_i^l); \Theta_s), \mathbf{y}_i^l) + \mathcal{L}_{reg}(f_s(T(\mathbf{x}_i^l); \Theta_s), \mathbf{y}_i^l) \right] \\ & + \lambda_u \frac{1}{M} \sum_j \left[\mathcal{L}_{cls}(f_s(T'(\mathbf{x}_j^u); \Theta_s), \hat{\mathbf{y}}_j^u) + \mathcal{L}_{reg}(f_s(T'(\mathbf{x}_j^u); \Theta_s), \hat{\mathbf{y}}_j^u) \right], \end{aligned} \quad (1)$$

where T and T' stands for weak and strong image transformations, $\mathbf{y} = \{y_l = (c_l, \text{bbox}_l)\}_{l=1}^L$ is the ground truth including L bboxes with classification label c_l . $\hat{\mathbf{y}} = f_t(T(\mathbf{x}); \Theta_t)$ is the pseudo-bboxes generated by the teacher model. Teacher parameter is updated as $\Theta_t \leftarrow (1 - \gamma)\Theta_t + \gamma\Theta_s$. λ_u is a weighting parameter. To ensure a fair comparison, Focal Loss [28] and GIoU loss [40] are set for \mathcal{L}_{cls} and \mathcal{L}_{reg} for all models in this study.

3.2 Consistent Adaptive Sample Assignment (ASA).

Each anchor in RetinaNet is assigned as positive only if its IOU with ground truth (GT) bbox is larger than a threshold. As described in Sec. 1, the static IOU-based assignment could assign irrelevant anchors as positive under noise in pseudo-bbox boundaries. Therefore, we propose to adopt Adaptive Sample Assignment (ASA). Specifically, a matching cost between each anchor² and ground truths (also including pseudo bboxes) is calculated [17, 18, 41], and anchors with lowest matching costs are assigned as positive. Given an anchor i , the cost between each GT y_l and the prediction p_i from the anchor considers the quality of classification, regression and prior information, as seen in Fig. 2, and is calculated as

$$C_{ij} = \mathcal{L}_{cls}(p_i, y_l) + \lambda_{reg}\mathcal{L}_{reg}(p_i, y_l) + \lambda_{dist}C_{dist}, \quad (2)$$

where λ_{reg} and λ_{dist} are weighting parameters. C_{dist} measures the center distance between anchor i and GT bbox y_l , which acts as a small ($\lambda_{dist} \sim 0.001$) center prior to stabilize training. After sorting out the matching cost for each GT y_l , anchors with top K lowest costs are assigned as positive. Since the assignment is made in accordance with the model’s detection quality, noise in pseudo-bboxes would have a negligible impact on feature points assignment. Therefore, the optimization target would be more consistent, as seen in Fig. 1.

3.3 BBox consistency via 3-D Feature Alignment Module (FAM-3D).

In common SSOD frameworks, pseudo bboxes are generated purely according to classification scores. A high-confidence prediction, however, does not always guarantee accurate bbox localization [6]. It again contributes to the

²The anchor definition in this study generalizes to feature points in anchor-free detectors.

noise in the pseudo-bbox. Therefore, inspired by TOOD [19], we introduce a 3-D Feature Alignment Module (FAM-3D) to calibrate the bbox localization with the classification confidence. It allows each classification feature to adaptively cast about the optimal feature vector for regression task.

Assume the feature pyramid is \mathbf{P} , we would like to construct a re-sampling function $\mathbf{P}' \leftarrow s(\mathbf{P})$ to rearrange the feature map to conduct the regression task, such that \mathbf{P}' better aligns with the classification features. Different from the single-scale feature re-sampling in [19], we extend the process to multi-scale feature space, considering the fact that the optimal features for classification and regression could be at different scales [42].

Our feature alignment is realized via a sub-branch in the detection head that predicts the 3-D offset with the feature pyramid for regression. As illustrated in Fig. 2, we add one extra $\text{CONV}_{3 \times 3}(\text{RELU}(\text{CONV}_{1 \times 1}))$ layer at different FPN levels and estimates an offset vector $\mathbf{d} = (d_0, d_1, d_2) \in \mathbb{R}^3$ for each regression prediction. \mathbf{P} is then re-ordered using the predicted offsets in two steps

$$\mathbf{P}'(i, j, l) \leftarrow \mathbf{P}(i + d_0, j + d_1, l) \quad (3)$$

$$\mathbf{P}'(i, j, l) \leftarrow \mathbf{P}'(i', j', l + d_2), \quad (4)$$

where Eq. 3 is to conduct feature offset in a 2-D space and Eq. 4 is the offset across different scales. i, j represent the planer feature coordinates while l indexes an FPN layer. In Eq. 4, i' and j' are the rescaled coordinates of i and j at different FPN levels. Notably, the extra CONV layers increases the computational cost slightly ($\sim 1\%$), but significantly improves the performance.

3.4 Temporal consistency using Gaussian Mixture Model (GMM)

Previous works [3, 5] require a static hyperparameter τ for pseudo bboxes filtering. It fails to take into account that the model's prediction confidence varies across categories and iterations, which makes inconsistent target and has a profound effect on performance [15]. Furthermore, tuning the threshold on different datasets is tedious.

Our goal is to find a way to automatically distinguish the positive from negative pseudo-bboxes. Specifically, we hypothesize that the score prediction s^c in each category c is sampled from a Gaussian mixture (GMM) distribution $\mathcal{P}(s^c)$ on all unlabeled data with two modalities, positive and negative. (see the score distribution in the subfigure of Fig. 2)

$$\mathcal{P}(s^c) = w_n^c \mathcal{N}(s^c | \mu_n^c, (\sigma_n^c)^2) + w_p^c \mathcal{N}(s^c | \mu_p^c, (\sigma_p^c)^2), \quad (5)$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian distribution, $w_n^c, \mu_n^c, (\sigma_n^c)^2$ and $w_p^c, \mu_p^c, (\sigma_p^c)^2$ represent the weight, mean and variance of negative and positive modalities, respectively. Expectation-Maximization (EM) algorithm is then used to infer the posterior $\mathcal{P}(\text{pos}|s^c, \mu_p^c, (\sigma_p^c)^2)$ which is the probability that a detection should be set as the pseudo-target for the student, and the adaptive score threshold is determined as

$$\tau^c = \underset{s^c}{\operatorname{argmax}} \mathcal{P}(\text{pos}|s^c, \mu_p^c, (\sigma_p^c)^2) \quad (6)$$

In practice, we maintain a prediction queue of size N ($N \sim 200$) for each class to fit GMM. Considering that the score distribution from a single-stage detector is strongly imbalanced as the majority of prediction is negative, only the top $K = \sum_k (s_k)$ number of predictions are stored in a queue. The EM algorithm only accounts for $\sim 7\%$ training time increase. The threshold can then be adaptively determined *w.r.t.* the model's performance at different training stages.

4 Experiments

In this section, we first evaluate our solution on a series of SSOD benchmarks, and then validate the effectiveness of each components through extensive ablation studies.

Datasets and Evaluation Setup. we conduct comprehensive experiment on the MS-COCO 2017 [43] benchmark and PASCAL VOC datasets [44].

We include three evaluation protocols: (1) COCO-PARTIAL: We randomly sample 1%/2%/5%/10% of the images in train2017 as labeled data and treat the rest as unlabeled data. We report the $AP_{50:95}$ ³ results on the val2017 as the evaluation metrics. (2) COCO-ADDITION: We use the full train2017 as labeled set and include the official unlabeled set unlabeled2017 as unlabeled set. The trained models are evaluated on val2017. (3) VOC-PARTIAL: We utilize the VOC2007 trainval set as the labeled data and make use of the VOC2012 trainval as our unlabeled data. The final model is

³ $AP_{50:95}$ is interchangeable with mAP in this study.

verified on VOC2007 test set using both AP_{50} and $AP_{50:95}$ following [3]. In addition, some of the model improvements are also evaluated on the standard fully-supervised COCO-1x training [28] to compare the relative benefits on semi- and fully-supervised regimes.

Implementation Details. To ensure a fair comparison, all detectors are trained on 8 GPUs with 5 images per GPU (1 labeled and 4 unlabeled images) similar to [6]. The detectors are optimized using SGD with a constant learning rate of 0.01, momentum of 0.9 and weight decay of 0.0001. The unlabeled data weight is $\lambda_U = 2$. No learning rate decay is applied. In COCO-PARTIAL and VOC-PARTIAL evaluation, we train the detectors for 180K iterations, whereas we increase the training time on COCO-ADDITION to 720K for better convergence. The teacher model is updated through EMA with a momentum of 0.9995. We follow the same data prepossessing and augmentation pipeline in [6]. We adopt RetinaNet [28] with ResNet-50 [27] backbone as our baseline. ImageNet [45]-pretrained model is used as initialization.

We compare our **Consistent-Teacher** with numerous prevailing SSOD approaches including CSD [4], STAC [3], Instant Teaching [7], Humble Teacher [23], Unbiased Teacher v1 and v2 [5, 9], Soft Teacher [6], ACRST [46], DSL [15], S4OD [14], Dense Teacher [11] and PseCo [10]. In addition, we implement a baseline method where students are trained using labeled and pseudo-labeled data, and the teacher is updated through EMA of student. We name it the Mean-Teacher baseline [16]. The default confidence threshold is set as 0.4.

4.1 Troubleshooting the Inconsistency Problems in SSOD

In the first step, we provide a thorough analysis to justify inconsistencies in SSOD, and how our solution addresses them.

Inconsistency Leads to Noisy Labels. We plot the mAP of the pseudo bboxes against the GT targets on unlabeled data in Figure 3(Left axis). It stands for the quality for the labels. In addition, the *inconsistency* is measured, which is an accumulation of the mismatch between the pseudo-bboxes of two consecutive teacher checkpoints (Right axis). Please refer to the Appendix 1.1 for the full formulation.

According to Figure 3(Right axis), while the Mean-Teacher suffers from unfavorable large inconsistencies during training, Consistent-Teacher significantly reduces the target discrepancy at different time steps. Consequently, our model enjoys continuous improvement overtime, therefore provides high quality labels for its student, as shown in Figure 3(Left axis).

Inconsistency Leads to Dynamic Definition of Targets.

Figure 4 plots the number of pseudo GTs per image on the unlabeled data using different thresholding schedules. Notably, it reveals a critical problems that, with static confidence thresholds $\tau = 0.4, 0.5, 0.6$, the number of pseudo label keeps going up as detector becomes more confident. GMM-based approach, on the other hand, adaptively adjust the best threshold according to the model capacity, with a nearly constant number of GTs, which reduces temporal inconsistency.

In Figure 5, we plot the estimated threshold curve obtained by GMM on COCO 1%/5%/10%. The value steadily increases as training proceeds. Further more, with less labeled samples, GMM sets higher confident threshold in accordance with more overfitting issues. Typical static threshold setting is incapable to address the inconsistency in learning targets, while GMM provides a gratifying solution.

Inconsistency Introduces Classification-Regression Misalignment.

It is a well-known problem in object detection that, the classification score may not fully reflect the regression quality [6, 24]. It deters the essence of SSOD since we rely heavily on the prediction score to filter labels. Figure 6 visualizes the confidence-IOU heatmap of all predicted bounding boxes on the COCO val2017. For each predicted bbox, we plot the confidence of the maximum category and its maximum IOU with the GT boxes in the corresponding class. As highlighted in the red squares, Mean-Teacher predicts low-confidence but high-IOU bboxes. On the other hand, our model generates predictions that are concentrated in high-confidence, high-IOU region. Consistent-Teacher gives rise to more calibrated predictions.

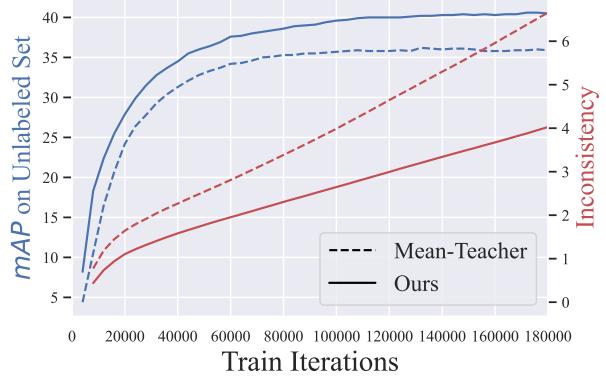


Figure 3: **Consistent-Teacher** improves the training consistency in SSOD. (Left axis) mAP on the unlabeled set at different time. (Right axis) The inconsistency of pseudo labels.

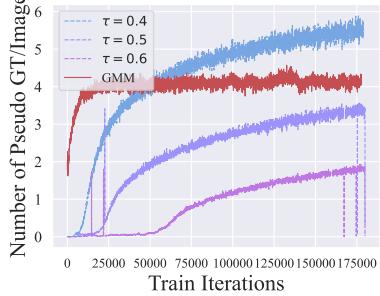


Figure 4: Number of pseudo labels/image with threshold schedules on COCO 10%.

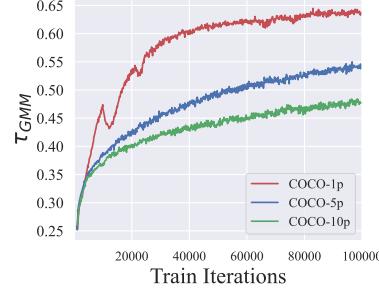


Figure 5: GMM threshold dynamics along training.

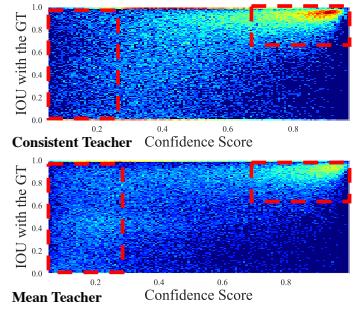


Figure 6: Heatmap of predicted bboxes confidence and its IOU score with GTs.

4.2 Semi-supervised Object Detection

In this section, we compare our method with previous state-of-the-arts under COCO-PARTIAL, VOC-PARTIAL and COCO-ADDITION evaluation protocol.

COCO-PARTIAL Results. Table 1 systematically compares mAP of all aforementioned semi-supervised detectors trained with COCO 1%/2%/5%/10% labels. We first note that the simple Mean Teacher baseline with RetinaNet detector constitutes a strong method for SSOD. It achieves an mAP of 35.5 on COCO 10% experiments without sophisticated data re-weighting strategy or pseudo labeling selection methods. More surprisingly, Consistent-Teacher achieves remarkable progress over current methods on 2%/5%/10% experiments. It scores 36.1 and 40.0 mAP on COCO 5%/10% data, largely surpassing the best-performed model Dense Teacher by ~ 3.1 and ~ 3 mAP.

Table 1: COCO-PARTIAL comparison with other semi-supervised detector on val2017. The results for two-stage (upper half) and single-stage (lower half) detectors are listed separately. We also report the Faster-RCNN and RetinaNet performance trained on labeled data only. All models adopt ResNet50 with FPN as backbone. We highlight the previous best record with underline.

Method	1% COCO	2% COCO	5% COCO	10% COCO
Labeled Only	9.05	12.70	18.47	23.86
CSD	<u>10.51</u>	13.93	18.63	22.46
STAC	13.97	18.25	24.38	28.64
Instant Teaching	18.05	22.45	26.75	30.40
Humble teacher	16.96	21.72	27.70	31.61
Unbiased Teacher	20.75	24.30	28.27	31.50
Soft Teacher	20.46	-	30.74	34.04
ACRST	<u>26.07</u>	<u>28.69</u>	31.35	34.92
PseCo	22.43	27.77	32.50	36.06
Labeled Only	10.22	13.80	19.40	24.10
Unbiased Teacher v2	22.71	26.03	30.08	32.61
DSL	22.03	25.19	30.87	36.22
Dense Teacher	22.38	27.20	<u>33.01</u>	<u>37.13</u>
S4OD	20.10	-	30.00	32.90
Mean-Teacher	20.40	26.00	30.40	35.50
Consistent-Teacher	25.30	30.40	36.10	40.00

VOC-PARTIAL Results. In addition to the COCO evaluations, we compare our proposed model against other SSOD approaches on VOC0712 datasets in Table 3. Again, we notice that our Consistent-Teacher makes outstanding improvements over its counterparts. Our method shows an improvement of 2.2 absolute mAP compared with the latest state-of-the-art [9, 15].

COCO-addition Results. Now we would like to push our model to its limits by taking the full COCO train train2017 as labeled data and additional unlabeled2017 as unlabeled data. As shown in Table 2, in the case of COCO-ADDITION, our model achieves 47.20 mAP, surpassing all previous state-of-the-arts.

Table 2: COCO-ADDITION experimental results on val2017 with unlabeled2017 as unlabeled set. Note that $1\times$ represents 90K training iterations, and $N\times$ represents $N\times 90K$ training iterations.

Method	$AP_{50:95}$
CSD(3 \times)	40.20 $\xrightarrow{-1.38}$ 38.82
STAC(6 \times)	39.48 $\xrightarrow{-0.27}$ 39.21
Unbiased Teacher(3 \times)	40.20 $\xrightarrow{+1.10}$ 41.30
ACRST(3 \times)	40.20 $\xrightarrow{+2.59}$ 42.79
Soft Teacher(16 \times)	40.90 $\xrightarrow{+3.70}$ 44.50
DSL(2 \times)	40.20 $\xrightarrow{+3.60}$ 43.80
PseCo(8 \times)	41.00 $\xrightarrow{+5.10}$ 46.10
Dense Teacher(8 \times)	41.24 $\xrightarrow{+4.88}$ 46.12
Consistent-Teacher (8\times)	40.50 $\xrightarrow{+6.70}$ 47.20

Table 3: VOC-PARTIAL experimental results comparison with other semi-supervised detector on VOC07 labeled and VOC12 unlabeled set.

Method	AP_{50}	$AP_{50:95}$
Labeled Only	72.63	42.13
CSD	74.70	-
STAC	77.45	44.64
ACRST	78.16	50.12
Instant Teaching	79.20	50.00
Humble Teacher	80.94	53.04
Unbiased Teacher	77.37	48.69
Unbiased Teacher v2	<u>81.29</u>	<u>56.87</u>
Mean-Teacher	77.02	53.61
Consistent-Teacher	81.00	59.00

Table 4: Comparisons between IoU-based and our adaptive anchor assignment on COCO.

Assignment	$AP_{50:95}^{1\times}$	$AP_{50:95}^{10\%}$
IOU-based	38.4	35.50
our ASA	40.1($+1.7$)	38.50($+3.0$)

Table 5: Ablation Study on detection head structure. We compare the performance, model size and FLOPs on different head structures on COCO 10% and standard 1 \times evaluation. FLOPs are measured on the input image size of 1280 \times 800.

Method	FLOPs (G)	$AP_{50:95}^{1\times}$	$AP_{50:95}^{10\%}$
Ours w/o FAM	205.21	40.1	38.5
Ours w FAM-2D	205.70	40.4($+0.3$)	39.1($+0.6$)
Ours w FAM-3D	208.49	40.7($+0.6$)	39.5($+1.0$)

4.3 Ablation Study

In this section, we validate the effectiveness of our 3 major designs on the MS-COCO dataset.

Adaptive Sample Assignment. We first examine the effect of ASA strategy. To enable a fair comparison between all assigners, we utilize the Mean Teacher with fixed confidence threshold 0.4 and unlabeled weight of 2 as our baseline and replace its IOU-based assignment with our proposed ASA. Since the adaptive assignment is also applicable to the supervised scenario, we further experiment on the supervised MS-COCO with the standard 1 \times (12 epochs) training setting. It is notable that, as shown in Table 4, robust sample assignment plays a pivotal role in SSOD. By specializing the assignment policy on semi-supervised tasks, our ASA achieves 38.50 mAP on COCO 10%, with an improvement of 3 mAP compared with the heuristic matching cost using IOU. **Another finding** is that, the performance benefits from ASA is almost doubled on SSOD (3.0 mAP) than on the fully supervised setting (1.7 mAP). It suggests our proposed ASA is particularly beneficial in the evaluation of the SSOD tasks, as also seen in Fig. 1 of its ability to suppress the confirmation bias in SSOD.

3D Feature Alignment Module. To testify the effectiveness of FAM, we first replace the FAM-3D as a 2-D counterpart, which is adopted in [19]. Table 5 provides the ablative study of our method with different FAM structure. We observe that the FAM-3D surpasses the setting without feature alignment by 1.0 mAP and FAM-2D by 0.4 mAP on COCO 10% evaluation, with negligible parameters and FLOPs. It is shown that, by automatically estimating the best 3D feature location for classification and regression, the semi-supervised detector are better calibrated to identify high quality pseudo-labels.

GMM. We testify the detector performance with or without the GMM-based pseudo-labelling. We replace it with a hard confidence threshold $\tau \in (0.2, 0.3, 0.4, 0.5, 0.6, 0.9)$. Figure 7 illustrates the test mAP on val2017. Notice that the detetor is highly sensitive to confidence threshold, with the optimal constant threshold at 0.4. By fitting the distribution of confidence, GMM dynamically adjusts the threshold for selecting pseudo-labels. This allows our model to gain more accuracy and stable supervision signal than a fixed threshold, achieving the final performance of 40.00 mAP with 0.5 mAP improvement on 10% labeled data. GMM is also higher than the model using hard threshold (0.4) at different ratios of labeled data as well, as illustrated in Figure 8.

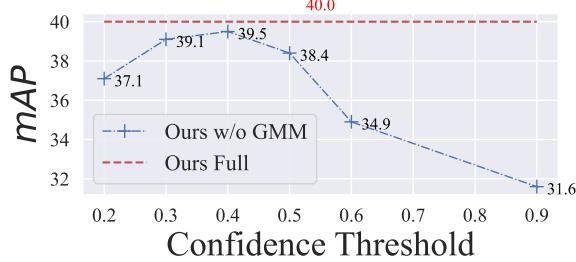


Figure 7: Ablative study of GMM-based pseudo-label filtering. Each value represents the mAP score on COCO 10% data.



Figure 8: Ablation of GMM at different labeled data ratio on COCO. Models are compared to baselines with a hard threshold 0.4.

5 Conclusion and Future work

In this paper, we systematically investigate the inconsistency problems in semi-supervised object detection, where the pseudo boxes may be highly inaccurate and vary greatly at different stages of training. To alleviate the aforementioned problem, we present a simple yet effective semi-supervised object detector termed **Consistent-Teacher**. We introduce adaptive anchor assignment which selects the positive anchor with lowest matching costs and FAM which regress the 3-D feature pyramid offsets that aligns classification and regression tasks. To solve the temporal inconsistency in pseudo-bboxes, we leverage GMM to dynamically adjust the threshold for self-training. Through integration of three designs, our **Consistent-Teacher** is able to simultaneously obtain a robust anchor assignment with consistent pseudo-bboxes, outperforming the state-of-the-art methods by a large margin on a series of SSOD benchmarks.

References

- [1] 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] Cong Chen, Shouyang Dong, Ye Tian, Kunlin Cao, Li Liu, and Yuanhao Guo. Temporal self-ensembling teacher for semi-supervised object detection. *IEEE Transactions on Multimedia*, 2021.
- [3] 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.
- [4] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. *Advances in neural information processing systems*, 32, 2019.
- [5] 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.
- [6] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3060–3069, 2021.
- [7] 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.
- [8] Binbin Chen, Weijie Chen, Shicai Yang, Yunyi Xuan, Jie Song, Di Xie, Shiliang Pu, Mingli Song, and Yueting Zhuang. Label matching semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14381–14390, 2022.
- [9] Yen-Cheng Liu, Chih-Yao Ma, and Zsolt Kira. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9819–9828, 2022.
- [10] Gang Li, Xiang Li, Yujie Wang, Shanshan Zhang, Yichao Wu, and Ding Liang. Pseco: Pseudo labeling and consistency training for semi-supervised object detection. *arXiv preprint arXiv:2203.16317*, 2022.

- [11] Hongyu Zhou, Zheng Ge, Songtao Liu, Weixin Mao, Zeming Li, Haiyan Yu, and Jian Sun. Dense teacher: Dense pseudo-labels for semi-supervised object detection. *arXiv preprint arXiv:2207.02541*, 2022.
- [12] 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.
- [13] 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. IEEE, 2020.
- [14] Yueming Zhang, Xingxu Yao, Chao Liu, Feng Chen, Xiaolin Song, Tengfei Xing, Runbo Hu, Hua Chai, Pengfei Xu, and Guoshan Zhang. S4od: Semi-supervised learning for single-stage object detection. *arXiv preprint arXiv:2204.04492*, 2022.
- [15] Binghui Chen, Pengyu Li, Xiang Chen, Biao Wang, Lei Zhang, and Xian-Sheng Hua. Dense learning based semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4815–4824, 2022.
- [16] 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.
- [17] 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.
- [18] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.
- [19] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3510–3519, 2021.
- [20] Xingyi Yang, Jingwen Ye, and Xinchao Wang. Factorizing knowledge in neural networks. *arXiv preprint arXiv:2207.03337*, 2022.
- [21] Yandong Li, Di Huang, Danfeng Qin, Liqiang Wang, and Boqing Gong. Improving object detection with selective self-supervised self-training. In *European Conference on Computer Vision*, pages 589–607. Springer, 2020.
- [22] 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.
- [23] 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.
- [24] 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 *CVPR*, 2020.
- [25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.
- [26] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.
- [27] 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.
- [28] 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.
- [29] 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.
- [30] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems*, 29, 2016.
- [31] 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.

- [32] Xingyi Yang, Yong Wang, and Robert Laganière. A scale-aware yolo model for pedestrian detection. In *International Symposium on Visual Computing*, pages 15–26. Springer, 2020.
- [33] 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.
- [34] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019.
- [35] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. Foveabox: Beyond anchor-based object detection. *IEEE Transactions on Image Processing*, 29:7389–7398, 2020.
- [36] Xiaosong Zhang, Fang Wan, Chang Liu, Rongrong Ji, and Qixiang Ye. Freeanchor: Learning to match anchors for visual object detection. *Advances in neural information processing systems*, 32, 2019.
- [37] Wei Ke, Tianliang Zhang, Zeyi Huang, Qixiang Ye, Jianzhuang Liu, and Dong Huang. Multiple anchor learning for visual object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10206–10215, 2020.
- [38] 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.
- [39] Kang Kim and Hee Seok Lee. Probabilistic anchor assignment with iou prediction for object detection. In *European Conference on Computer Vision*, pages 355–371. Springer, 2020.
- [40] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.
- [41] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.
- [42] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018.
- [43] 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.
- [44] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>.
- [45] 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.
- [46] Fangyuan Zhang, Tianxiang Pan, and Bin Wang. Semi-supervised object detection with adaptive class-rebalancing self-training. *arXiv preprint arXiv:2107.05031*, 2021.

In this supplementary material, we provide additional experimental quantitative results, model size comparison, as well as bounding boxes visualization to further support the effectiveness of our proposed **Consistent-Teacher**. In addition, we delineate more experimental details, implementation information, and hyper-parameter settings of our method. Our code is also attached for your reference.

1 More details in Consistent-Teacher

1.1 Inconsistency measurement.

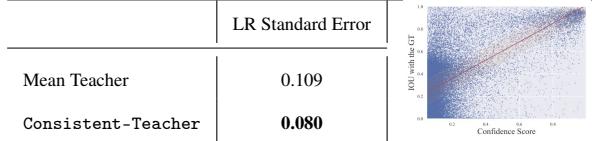
Inconsistency refers to the fact that the pseudo boxes may be highly inaccurate and vary greatly at different stages of training. Therefore, we measure the pseudo-bboxes variation across different training steps. Specifically, we store the checkpoints every 4000 training steps. We then run inference using these checkpoints on a subset with 5000 images from the unlabeled set. The prediction output from the previous checkpoint is then set as GT and we evaluate the mAP of the current checkpoint with the previous predictions. Therefore, a higher mAP implies a more consistent pseudo targets. Then the inconsistency is measured by accumulating $1 - mAP$ for these checkpoints to reflect the accumulated effect of noisy targets.

2 Verify the Inconsistency in SSOD

Assignment Inconsistency under Noisy Pseudo Labels. To illustrate that the conventional IOU-based or heuristic label assignment is problematic in SSOD, we intentionally inject random noise to the ground-truth bounding boxes and testify the assignment consistency by quantifying the assignment IOU (A-IOU) of clean and noisy assignments. Suppose a bounding box $b = (x_1, y_1, x_2, y_2)$ is assigned to a set of k anchors $A = \{a_1, \dots, a_k\}$. We add Gaussian noise to its coordinate with a noise ratio ρ , so that $b' = (x_1 + \epsilon_{x_1} \times w, y_1 + \epsilon_{y_1} \times h, x_2 + \epsilon_{x_2} \times w, y_2 + \epsilon_{y_2} \times h)$, in which w and h are width and height of the box. $\epsilon_{x_1}, \epsilon_{y_1}, \epsilon_{x_2}, \epsilon_{y_2}$ are sampled from a normal distribution $\mathcal{N}(0, \rho)$. The perturbed box b' is matched to a new set of l anchors $A' = \{a'_1, \dots, a'_l\}$. The A-IOU is computed as the intersection-of-union between A and A' . The higher A-IOU score suggests the assignment is more robust to label noise.

We testify the assignment consistency under two scenario. First, we calculate the assignment IOU with different degrees of noise ratio $\rho \in \{0.1, 0.2, \dots, 0.5\}$ using the final model. Second, we would like to investigate how the assignment consistency change through training. We report the A-IOU at different time of training with a constant $\rho = 0.1$. We compare our ASA with IOU-based assigner [25, 28, 29] and ATSS assigner [24] with Mean Teacher RetinaNet baseline on COCO 10%. All modules except for the assignment are kept the same to provide a fair comparison. For both evaluations, we randomly select 1000 images from val2017 to compute the A-IOU. Figure 9 visualize the mean \pm std A-IOU between clean and noisy label at different training time and different noise ratio ρ . In Figure 9(a), both ATSS and our ASA provides higher A-IOU compared with the broadly applied IOU-based assignment. However, ATSS is still based on heuristic matching rule between label and anchor boxes. ASA, instead, steadily improves itself as the detector becomes more accurate. In Figure 9(b), we see that IOU-based assignment fails to maintain the initial assignment when the large magnitude of noise is introduced in the labels. Given the noisy nature of pseudo label in SSOD, our experiment suggests that IOU-based assignment is incapable of maintaining the assignment consistency in SSOD. In contrast, our ASA strategy still performs well under sever noise scenario. This experiment supports our argument that the proposed consistent assignment strategy is robust to label noise in SSOD. [b]

Table 6: Classification and Regression inconsistency analysis using IOU-Confidence linear regression (LR) error. We also provide the Mean Teacher IOU-Confidence plot on the right.



Classification and Regression Inconsistency. We unveil the regression and classification mismatch problem in SSOD by identifying the mismatch between the high-score and high-IOU predictions. We obtain the confidence-IOU pairs on val2017 using **Consistent-Teacher** and Mean Teacher RetinaNet when trained on COCO 10% data, and analyze the correlation between the two variables. We apply linear regression and measures the standard error to reflect the correlation between confidences and IOUs. Smaller error indicates higher correlation.

Table 6 provides the LR standard error for **Consistent-Teacher** and Mean Teacher RetinaNet. The right scatter figure displays the confidence-IOU of Mean Teacher. We observe clear cls-reg misalignment on semi-supervised detectors: numerous low-confident predictions possess high IOU score. It indicates that classification confidence does not provide a strong enough clue for an accurate regression result, which give rise to erroneous pseudo-label noise during training. The high LR error of 0.109 with Mean teacher also demonstrates this point. On the contrary, our **Consistent-Teacher** largely eliminates the mismatch between the two tasks with a lower LR error of 0.080. It supports our arguments that **Consistent-Teacher** can align the classification and regression sub-tasks and reduce the mismatch in SSOD.

3 Semi-supervised detection results visualization

3.0.1 Qualitative comparison with Baseline.

We further compare the baseline Mean Teacher RetinaNet with our **Consistent-Teacher** by visualizing the predicted bounding boxes on val2017 under the COCO 10% protocol. In Figure 10, we plot the predicted and ground-truth bounding boxes in **Violet** and **Orange** respectively, alongside with the false positive bboxes highlighted in **Red**.

There are 3 general properties that we could observed in our demonstration.

1. First, **Consistent-Teacher** fits the situation of crowded object localization better, whereas Mean Teacher often mistakes the intersection of two overlapped objects as a new instance. For example, in the scenes of zebras or sheep, Mean Teacher often gives a false positive output in the overlapping area of the two objects, while **Consistent-Teacher** largely resolves the inaccurate positioning problem through the adaptive anchor selection mechanism.
2. Secondly, we see that under the semi-supervised setting, the Mean Teacher RetinaNet would either predict the wrong class for the correct location or regress an inaccurate bounding box despite its high classification confidence. For example, birds are sometimes misidentified as airplanes even when the localization is accurate. It is mainly attributed to the inconsistency of classification and regression tasks, i.e. the features required for regression may not be optimal for classification. **Consistent-Teacher** effectively discriminates similar categories using the FAM-3D to select the features dynamically.
3. Third, **Consistent-Teacher** embraces higher recall since it is capable of detecting small or crowded instances which Mean Teacher fails to point out. For example, **Consistent-Teacher** discovers most of the hot dogs on the grill while Mean Teacher neglects most of them.

3.0.2 Good cases and Failure cases.

We provides more examples to showcase the good and failure examples produced by **Consistent-Teacher** on COCO val2017 in Figure 11 and Figure 12. Although our proposed method achieved gratifying performance on a series of SSOD benchmarks, we can still point out its deficiencies in Figure 12. First, the trained detector lacks robustness to

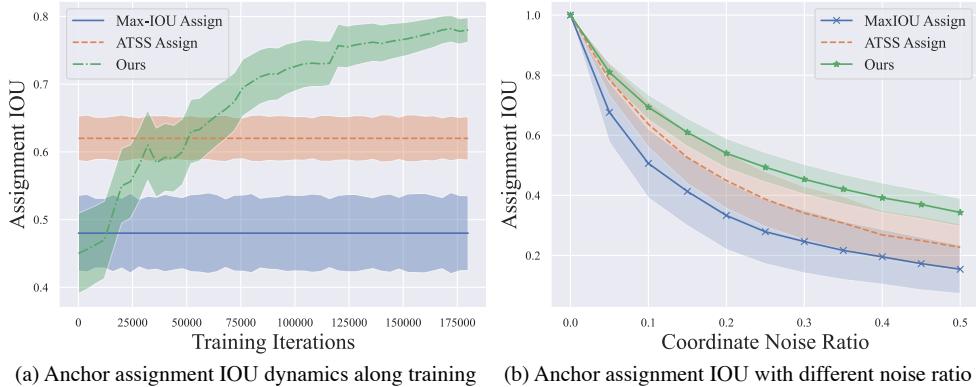


Figure 9: Assignment IOU score between ground-truth and the noisy bounding boxes (a) at different time of training and (b) using different noise ratio.



Figure 10: Qualitative comparison on the COCO%10 evaluation. The bounding boxes in **Orange** is the ground-truth, and **Violet** refers to the prediction. **Red** highlights the false positive predictions.

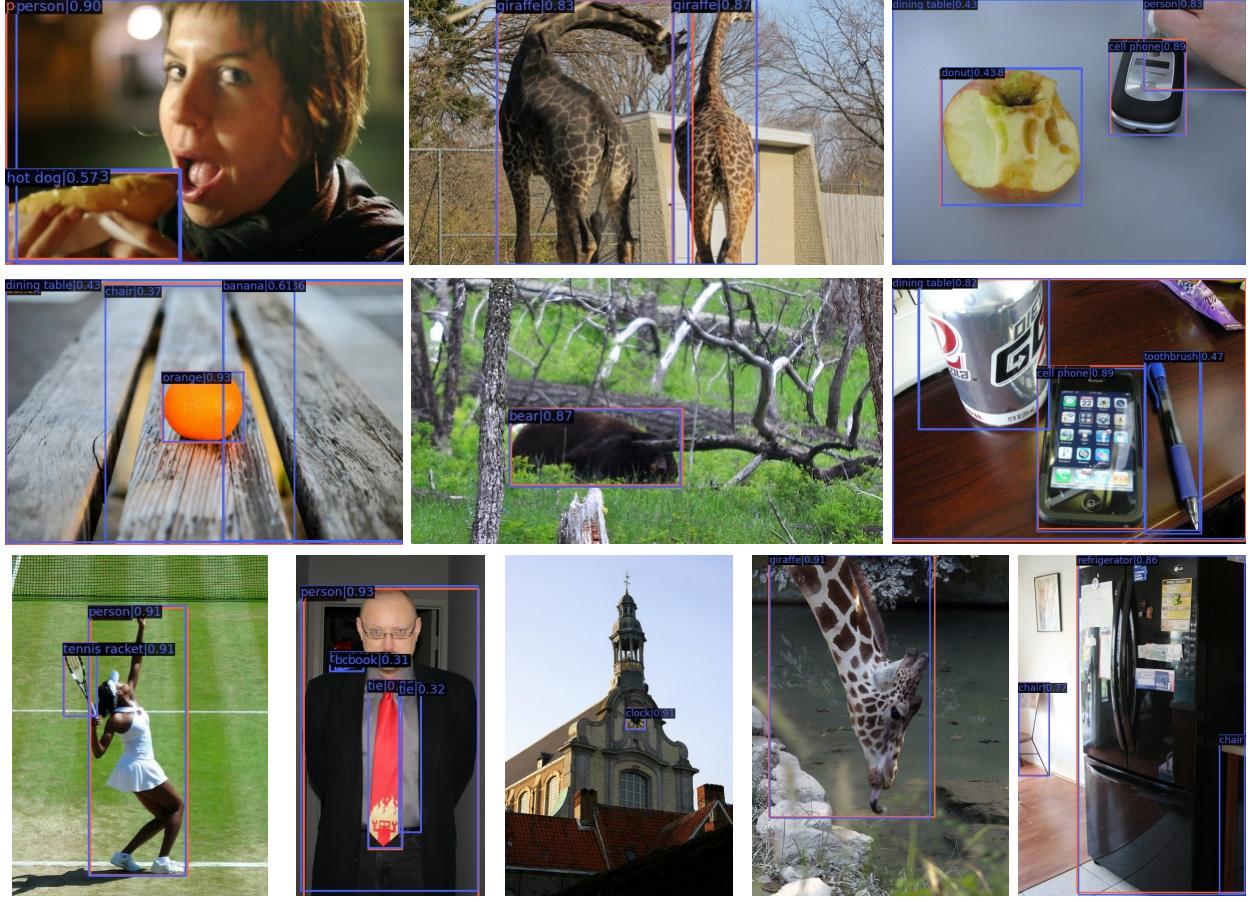


Figure 11: Good detection results for the COCO%10 evaluation. The bounding boxes in **Orange** is the ground-truth, and **Violet** refers to the prediction.

some out-of-distribution samples, for example, cartoon characters on street signs are recognized as real people, and reflections in mirrors are recognized as objects. Second, our detection performance is poor for some classes with small sizes, such as toothbrushes, hair dryers, etc. Third, *Consistent-Teacher* also tends to treat parts of the object as a whole, such as the head of the giant panda as a separate animal (in the lower left corner), and the dial of a clock as the entire clock (on the right of the panda).

4 Experiment and Hyper-parameter settings

4.1 Datasets and data prepossessing.

4.1.1 MS-COCO 2017.

The Microsoft Common Objects in Context (MS-COCO) is a large-scale object detection, segmentation, key-point detection, and captioning dataset. We use COCO2017 in our experiments for SSOD, which includes 118K training and 5K validation images along with bounding boxes of 80 object categories.

4.1.2 PASCAL VOC 2007-2012.

The PASCAL Visual Object Classes (VOC) dataset contains 20 object categories alongside with pixel-level segmentation annotations, bounding box annotations, and object class annotations. The official VOC 2007 `trainval` set is adopted as the labeled set with 5011 images and the 11540 images from VOC 2012 `trainval` set is used as unlabeled data in this study. We evaluate on the VOC 2007 test set.

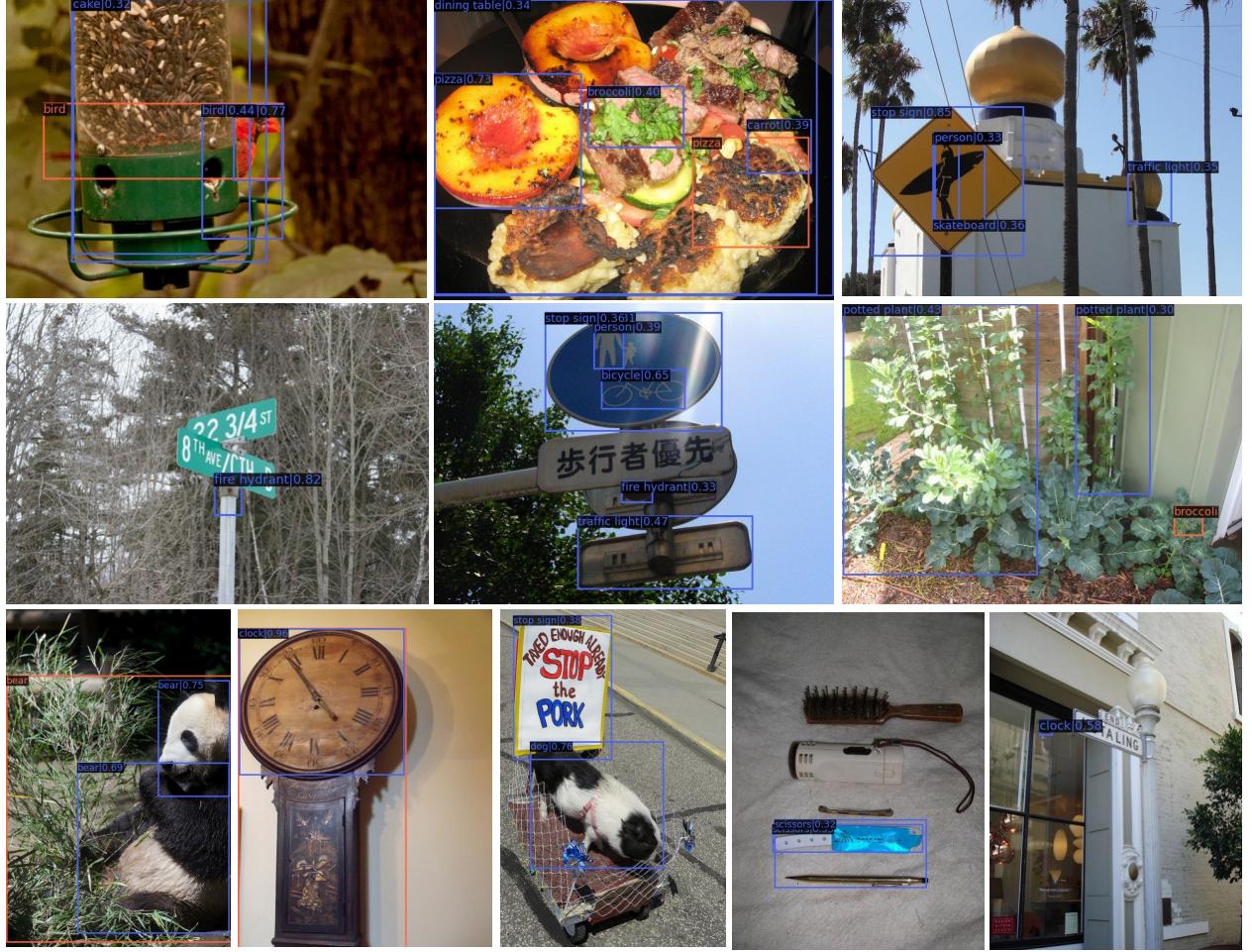


Figure 12: Failure detection results for the COCO%10 evaluation. The bounding boxes in **Orange** is the ground-truth, and **Violet** refers to the prediction.

4.1.3 Data Augmentations.

We use the same data augmentations as described in Soft Teacher [6], including a labeled data augmentation in Table 7, a weak unlabeled augmentation in Table 8 and a strong unlabeled augmentation in Table 9.

4.2 Implementation Details

We implement our **Consistent-Teacher** based on MMdetection⁴ framework with the data prepossessing code from the open-sourced Soft-Teacher⁵ and google ssl-detection⁶. We train our detectors on 8 NVIDIA Tesla V100 GPUs. It takes approximately 3 days for an 180K training. Each GPU contains 1 labeled image and 4 unlabeled images. The source code is attached in a separate zip file.

⁴<https://github.com/open-mmlab/mmsegmentation>

⁵<https://github.com/microsoft/SoftTeacher>

⁶https://github.com/google-research/ssl_detection/

Table 7: Data augmentation for labeled image training.

Transformation	Description	Parameter Setting
RandomResize	Resize the image to a height of h randomly sampled from $h \sim U(h_{min}, h_{max})$, while keeping the height-width ratio unchanged.	$h_{min} = 400, h_{max} = 1200$ in MS-COCO $h_{min} = 480, h_{max} = 800$ in PASCAL-VOC
RandomFlip	Randomly horizontally flip a image with probability of p .	$p = 0.5$
OneOf	Select one of the transformation in a transformation set T .	$T = \text{TransAppearance}$

Table 8: Weak data augmentation for unlabeled image.

Transformation	Description	Parameter Setting
RandomResize	Resize the image to a height of h randomly sampled from $h \sim U(h_{min}, h_{max})$, while keeping the height-width ratio unchanged.	$h_{min} = 400, h_{max} = 1200$ in MS-COCO $h_{min} = 480, h_{max} = 800$ in PASCAL-VOC
RandomFlip	Randomly horizontally flip a image with probability of p .	$p = 0.5$

Table 9: Strong data augmentation for unlabeled image.

Transformation	Description	Parameter Setting
RandomResize	Resize the image to a height of h randomly sampled from $h \sim U(h_{min}, h_{max})$, while keeping the height-width ratio unchanged.	$h_{min} = 400, h_{max} = 1200$ in MS-COCO $h_{min} = 480, h_{max} = 800$ in PASCAL-VOC
RandomFlip	Randomly horizontally flip a image with probability of p .	$p = 0.5$
OneOf	Select one of the transformation in a transformation set T .	$T = \text{TransAppearance}$
OneOf	Select one of the transformation in a transformation set T .	$T = \text{TransGeo}$
RandErase	Randomly selects K rectangle region of size $\lambda h \times \lambda w$ in an image and erases its pixels with random values, where (h, w) are height and width of the original image.	$K \in U(1, 5)$ $\lambda \in U(0, 0.2)$

Table 10: Appearance transformations, called TransAppearance.

Transformation	Description	Parameter Setting
Identity	Returns the original image.	
Autocontrast	Maximizes the image contrast by setting the darkest (lightest) pixel to black (white).	
Equalize	Equalizes the image histogram.	
RandSolarize	Invert all pixels above a threshold value T .	$T \in U(0, 1)$
RandColor	Adjust the color balance of image. $C = 0$ returns a black&white image, $C = 1$ returns the original image.	$C \in U(0.05, 0.95)$
RandContrast	Adjust the contrast of image. $C = 0$ returns a solid grey image, $C = 1$ returns the original image.	$C \in U(0.05, 0.95)$
RandBrightness	Adjust the brightness of image. $C = 0$ returns a black image, $C = 1$ returns the original image.	$C \in U(0.05, 0.95)$
RandSharpness	Adjust the sharpness of image. $C = 0$ returns a blurred image, $C = 1$ returns the original image.	$C \in U(0.05, 0.95)$
RandPolarize	Reduce each pixel to C bits.	$C \in U(4, 8)$

Table 11: Geometric transformations, called TransGeo.

Transformation	Description	Parameter Setting
RandTranslate X	Translate the image horizontally by $\lambda \times$ image width.	$\lambda \in U(-0.1, 0.1)$
RandTranslate Y	Translate the image vertically by $\lambda \times$ image height.	$\lambda \in U(-0.1, 0.1)$
RandRotate Y	Rotates the image by θ degrees.	$\theta \in U(-30^\circ, 30^\circ)$
RanShear X	Shears the image along the horizontal axis with rate R .	$R \in U(-0.480, 0.480)$
RanShear Y	Shears the image along the vertically axis with rate R .	$R \in U(-0.480, 0.480)$