Consistent-Teacher: Towards Reducing Inconsistent Pseudo-targets in Semi-supervised Object Detection

Xinjiang Wang[1*] Xingyi Yang[3*†] Shilong Zhang[2], Yijiang Li[1‡]
Litong Feng[1] Shijie Fang[4‡] Chengqi Lyu[2] Kai Chen[2] Wayne Zhang[1]
[1]SenseTime Research [2]Shanghai AI Laboratory [3]National University of Singapore [4]Peking University
wangxinjiang@sensetime.com, xyang@u.nus.edu

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 detector. It injects noise into the student's training, leading to severe overfitting problems. 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 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 47.7 mAP. Our code is available at* `https://github.com/Adamdad/ConsistentTeacher`.

1. Introduction

The goal of semi-supervised object detection (SSOD) [3, 5, 12, 12, 13, 17, 24, 25, 30, 36, 43, 44] 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 [35] or data augmentation [12, 30]. In addition, to improve pseudo-label quality, the teacher model is updated as a moving average [24, 36, 44] of the student parameters.

In this study, we point out that the performance of semi-supervised detectors is still largely hindered by the inconsistency in pseudo-targets. **Inconsistency** means that the pseudo boxes may be highly inaccurate and vary greatly at different stages of training. As a consequence, inconsistent oscillating bounding boxes (bbox) bias SSOD predictions with accumulated error. Different from semi-supervised classification, SSOD has one extra step of assigning a set of pseudo-bboxes to each RoI/anchor as dense supervision. Common two-stage [24, 30, 36] and single-stage [4, 42] SSOD networks adopt static criteria for anchor assignment, *e.g.* IoU score or centerness. It is observed that the static assignment is sensitive to noise in the bounding boxes predicted by the teacher, as a small perturbation in the pseudo-bboxes might greatly affect the assignment results. It thus leads to severe overfitting on unlabeled images.

To verify this phenomenon, 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, causing erroneous targets to be associated with nearby objects under static IoU-based assignment. This is because some inactivated anchors are falsely assigned positive in the student network. Consequently, the network overfits as it produces inconsistent labels for neighboring objects. The overfitting is also observed in the classification loss curve on unlabeled images[1].

*Equally contributed.
‡Work done during internship at SenseTime.
†Work done during internship at Shanghai AI Laboratory.

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

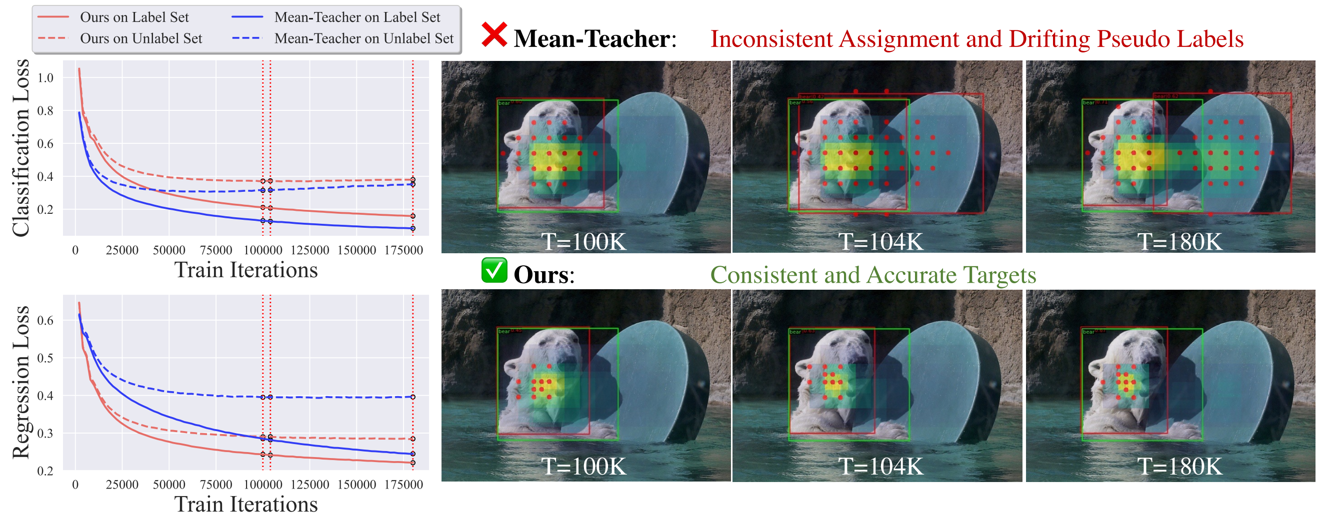

Figure 1. Illustration of inconsistency problem in SSOD on COCO 10 % evaluation. (Left) We compare the training losses between the Mean-Teacher and our Consistent-Teacher . In Mean-Teacher, inconsistent pseudo targets lead to overfitting on the classification branch, while regression losses become difficult 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, respectively, 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). A nearby board is finally misclassified as a polar bear in the baseline while our adaptive assignment prevents overfitting.

Through dedicated investigation, We find that one important factor that leads to the drifting pseudo-label is the mismatch between classification and regression tasks. Typically, only the classification score is used to filter pseudo-bboxes in SSOD. However, confidence does not always indicate the quality of the bbox [36]. Two anchors with similar scores, as a result, can have significantly different predicted pseudo-bboxes, leading to more false predictions and label drifting. Such phenomenon is illustrated in Fig. (1) with the varying pseudo-bboxes of the MeanTeacher around $T = 104K$. Therefore, the mismatch between the quality of a bbox and its confidence score would result in noisy pseudo-bboxes, which in turn exacerbates the label drifting.

The widely-employed hard threshold scheme also causes threshold inconsistencies in pseudo labels. Traditional SSOD methods [24, 30, 36] utilize a static threshold on confidence score for student training. However, the threshold serves as a hyper-parameter, which not only needs to be carefully tuned but should also be dynamically adjusted in accordance with the model's capability at different time steps. In the Mean-Teacher [32] paradigm, the number of pseudo-bboxes may increase from too few to too many under a hard threshold scheme, which incurs inefficient and biased supervision 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) [10, 11] greatly alleviates the effect of 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 the 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.

Then, we calibrate the classification and regression tasks so that the teacher's classification confidence provides a better proxy of the bbox quality. It produces consistent pseudo-bboxes for anchors of similar confidence scores, and thus the oscillation in pseudo-bbox boundaries is reduced. Inspired by TOOD [9], we propose a 3-D feature alignment module (FAM-3D) that allows classification features to sense and adopt the best feature in its neighborhood for regression. Different from the single scale searching, FAM-3D reorders the features pyramid for regression across scales as well. 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 threshold inconsistency in pseudo-bboxes, we apply Gaussian Mixture Model (GMM) to generate an adaptive threshold for each category during training. 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 will be 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. Our approach reaches 40.0 mAP with 10% of labeled data on MS-COCO, which is $\tilde{3}$ mAP ahead of the state-of-the-art [43]. When using the 100% labels together with extra unlabeled MS-COCO data, the performance is further boosted to 47.7 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 of the inconsistent target problem in SSOD, which incurs severe overfitting issues.
- We introduce an adaptive sample assignment to stabilize the matching between noisy pseudo-bboxes and anchors, leading to robust training for the student.
- We develop a 3-D feature alignment module (FAM-3D) to calibrate the classification confidence and regression quality, which improves the quality of pseudo-bboxes.
- We adopt GMM to flexibly determine the threshold for each class during training. The adaptive threshold evolves through time and reduces the threshold inconsistencies 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 predictions on augmented input samples [12, 18, 24, 30, 31, 34, 36, 38, 44]. Two-stage detectors [12, 24, 36] have been dominant in traditional SSOD methods while single-stage detectors have also shown the advantages for their simplicity and higher performance [4, 42, 43]. In this study, we adopt a single-stage SSOD framework [4, 43] and focus on the inconsistency problem. 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 [40] plays a substantial role in object detection. Typical Anchor-based or anchor-free detectors either adopt hard IoU thresholding [1,6,19,20,23,27,28,37] or the centerness prior [16, 26, 33] as the assigning criterion. In contrast, modern detectors have been shifting to adaptive assignment strategies. [10,14,15,41,45] For example, PAA [15] adaptively differentiates the positive anchors and negative ones by fitting the anchor scores distribution. OTA [10] treats the label assignment as an optimal transport problem so that the assignment cost is minimized.

Although the existing assignment methods are effective, they are limited to fully-supervised settings. In our work, we observe that using static assignment in SSOD induces server inconsistency issues and accumulates errors. We show that a simple cost-ware assignment stabilizes the label noise and significantly improves the performance of SSOD.

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.

3.1. Baseline Semi-Supervised Detector

We adopt a general SSOD paradigm as our baseline, namely a Mean-Teacher [24, 32, 36] pipeline with a RetinaNet [20] detector. The teacher model is an exponential moving average [32] of a student detector. 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, the 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

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

where T and T' stands for weak and strong image transformations, $\mathbf{y} = \{y_l = (c_l, \mathrm{bbox}_l)\}_{l=1}^L$ is the ground truth (GT) 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 [20] and GIoU loss [29] are set for $\mathcal{L}_{cls}$ and $\mathcal{L}_{reg}$ for all models in this study.

3.2. Consistent Adaptive Sample Assignment

Each anchor in RetinaNet is assigned as positive only if its IoU with ground truth (GT) bbox is larger than a threshold. Such static label assignment breaks one important

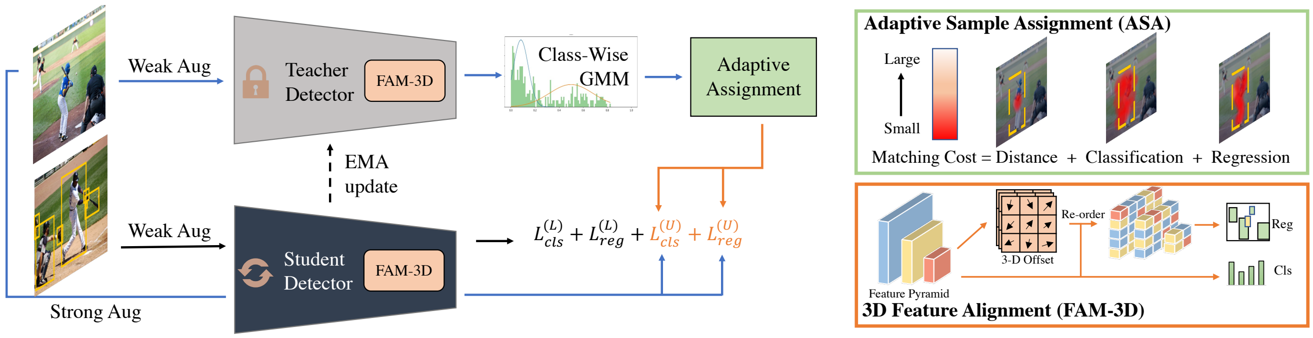

Figure 2. The pipeline of `Consistent-Teacher` . We design three modules to address the inconsistency in SSOD, where GMM dynamically determines the threshold; 3D feature alignment calibrates regression quality; Adaptive assignment assigns anchor based on matching cost.

property in semi-supervised learning. Take classification as an example, the instance-level pseudo-label satisfies

$$\hat{c} = \underset{c}{\operatorname{argmin}}\, \mathcal{L}(f_t(\mathbf{x}^u), c), \tag{2}$$

meaning that the pseudo-label $\hat{c}$ should align with its own prediction. However, this rule is broken when adopting static anchor assignment to SSOD. That is, the assigned labels for anchors sometimes contradict their own predictions, which is the root of the pseudo-label drifting phenomenon in Fig. 1. Therefore, we propose to assign pseudo-bboxes to anchors that minimize their loss

$$\min_{a_1,\cdots,a_N} \sum_{n}^{N} \Big[\mathcal{L}_{cls}\big(f_s(\mathbf{x}^u)_n, \hat{\mathbf{y}}^u_{a_n})\big) + \mathcal{L}_{reg}\big(f_s(\mathbf{x}^u)_n, \hat{\mathbf{y}}^u_{a_n}\big)\Big] \tag{3}$$

where n is the anchor index, and $a_n \in \{1, 2, \cdots, L+1\}$ stands for the assigned pseudo-bbox index from the L predicted bboxes, and the index $L+1$ represents the background label.

A simple solution to Eq. 3 is to assign anchors of lowest losses as positive for a pseudo-bbox. In practice, a matching cost between each anchor[2] and pseudo-bbox is calculated, and the anchors with the lowest costs are considered positive. Given an anchor n, the cost between each pseudo-bbox y_l and the prediction p_n from the anchor is calculated as

$$C_{nl} = \mathcal{L}_{cls}(p_n, y_l) + \lambda_{reg}\mathcal{L}_{reg}(p_n, y_l) + \lambda_{dist}C_{dist}, \tag{4}$$

where λ_{reg} and λ_{dist} are weighting parameters. C_{dist} calculates the distance between the center of anchor n and pseudo-bbox y_l, serving as a center prior with a small weighting value ($\lambda_{dist} \sim 0.001$) to stabilize the training. With the matching cost for each pseudo-bbox, 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 then have a negligible impact on the feature points assignment.

We are aware that a similar anchor assignment is adopted in supervised object detection [2, 10, 11], and thus we adopt a unified assignment for both labeled and unlabeled images. Despite their similar form, our ASA module addresses the unique pseudo-label shifting issue instead of catering for object variations in supervised settings [10].

3.3. BBox Consistency via 3-D Feature Alignment

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 [36]. It again contributes to the noise in the pseudo-bbox. Therefore, inspired by TOOD [9], we introduce a 3-D Feature Alignment Module (FAM-3D) to calibrate the bbox localization with classification confidence. It allows each classification feature to adaptively locate the optimal feature for the regression task.

Assuming the feature pyramid is $\mathbf{P}$ with $P(i,j,l)$ indicating the spatial location (i,j) at the l^{th} pyramid level, 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, so that $\mathbf{P}'$ better aligns with the classification features. Different from the single-scale feature re-sampling in [9], 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 [22].

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 $\textsc{Conv}_{3\times3}(\textsc{ReLU}(\textsc{Conv}_{1\times1}))$ layer at different FPN levels and estimate an offset vector $\mathbf{d} = (d_0, d_1, d_2) \in \mathbb{R}^3$ for each prediction. $\mathbf{P}$ is then re-ordered using the predicted offsets in two steps

$$P'(i,j,l) \leftarrow P(i+d_0, j+d_1, l) \tag{5}$$
$$P'(i,j,l) \leftarrow P'(i', j', l+d_2), \tag{6}$$

[2]Our anchor definition generalizes to anchor points in anchor-free and anchor boxes in anchor-based detectors.

where Eq. 5 is to conduct feature offset in a 2-D space and Eq. 6 is the offset across different scales. In Eq. 6, i' and j' are the rescaled coordinates of i and j at different FPN levels. Eq. 5 is realized by a bilinear interpolation, and Eq. 6 is conducted by a resizing of $P'(:, :, l + \lfloor d_2 \rfloor + 1)$ followed by a weighted average with $P'(:, :, l + \lfloor d_2 \rfloor)$ for a decimal number d_2, where $\lfloor \cdot \rfloor$ is the floor function. Notably, the extra CONV layers increase the computational cost slightly ($\sim 1\%$), but significantly improve the performance.

3.4. Thresholding with Gaussian Mixture Model

Previous works [24,30] 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 targets and has a profound effect on performance [4]. 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 for 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)

$$\mathcal{P}(s^c) = w_n^c \mathcal{N}(s^c|\mu_n^c, (\sigma_n^c)^2) + w_p^c \mathcal{N}(s^c|\mu_p^c, (\sigma_p^c)^2), \quad (7)$$

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. The Expectation-Maximization (EM) algorithm is then used to infer the posterior $\mathcal{P}(pos|s^c, \mu_p^c, (\sigma_p^c)^2)$ which is the probability that detection should be set as the pseudo-target for the student, and the adaptive score threshold is determined as

$$\tau^c = \underset{s^c}{\operatorname{argmax}}\, \mathcal{P}(pos|s^c, \mu_p^c, (\sigma_p^c)^2) \quad (8)$$

In practice, we maintain a prediction queue of size N ($N \sim 100$) 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 $\sim 10\%$ 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 component through extensive ablation studies.

Datasets and Evaluation Setup. we conduct comprehensive experiment on the MS-COCO 2017 [21] benchmark and PASCAL VOC datasets [8].

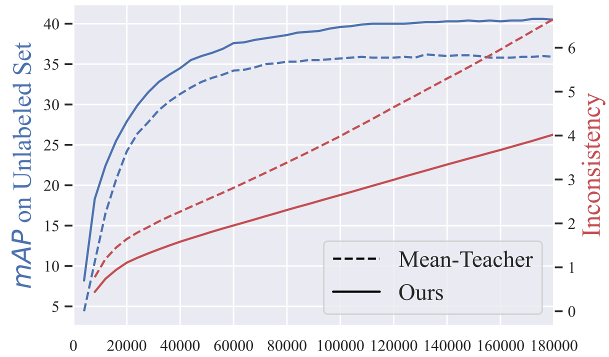

Figure 3. `Consistent-Teacher` improves the training consistency in SSOD. (Left axis) mAP on the unlabeled set at different times. (Right axis) The inconsistency of pseudo labels.

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 `unlabel2017` 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 verified on `VOC2007` test set using both AP_{50} and $AP_{50:95}$ following [30]. Additionally, we evaluate the model improvements on the standard fully-supervised COCO-1x training [20] to compare the relative benefits of our proposed method on both 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 [36]. The detectors are optimized using SGD with a constant learning rate of 0.01, a momentum of 0.9, and a 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 [36]. We adopt RetinaNet [20] with ResNet-50 [19] backbone as our baseline. ImageNet [7]-pretrained model is used as initialization.

We compare our `Consistent-Teacher` with numerous prevailing SSOD approaches including CSD [12], STAC [30], Instant Teaching [44], Humble Teacher [31], Unbiased Teacher v1 and v2 [24, 25], Soft Teacher [36], ACRST [39], DSL [4], S4OD [42], Dense Teacher [43] and

[3] $AP_{50:95}$ is interchangable with mAP in this study.

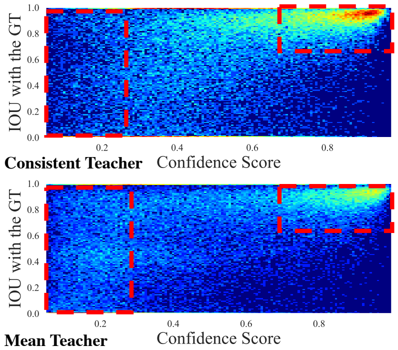

Figure 4. Heatmap of predicted bboxes confidence and its IoU score with GTs.

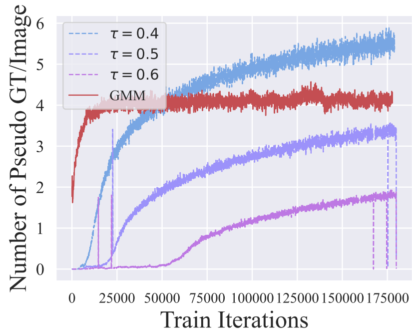

Figure 5. Number of pseudo labels/image with threshold schedules on COCO 10%.

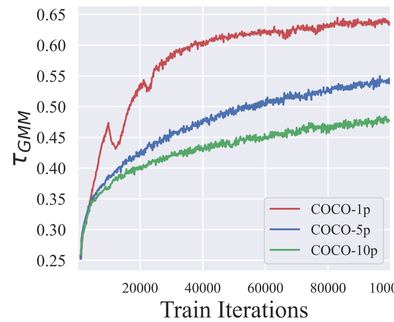

Figure 6. Average GMM thresholds across different classes along with the training.

PseCo [17]. In addition, we implement a baseline method where students are trained using labeled and pseudo-labeled data, and the teacher is updated through a moving average of the student. We name it the Mean-Teacher baseline [32]. The default confidence threshold is set as 0.4.

4.1. Troubleshooting the Inconsistencies in SSOD

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

Inconsistency Leading 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 of 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 supplementary for the full formulation.

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

Inconsistency Caused by Classification-Regression Misalignment. It is a well-known problem in object detection that, the classification score may not fully reflect the regression quality [36, 40]. It deters the essence of SSOD since we rely heavily on the prediction score to filter labels. Figure 4 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 and high-IoU regions. `Consistent-Teacher` gives rise to more calibrated predictions.

A demo video is attached in the Supplementary Material to illustrate that cls-reg misalignment leads to shifting and noisy targets. Our FAM-3D largely prevents low-quality, but high-score noise predictions thus reducing inconsistency.

Inconsistency Caused by Hard Score Threshold. Figure 5 plots the number of pseudo GTs per image on the unlabeled data using different thresholding schedules. Notably, it reveals a critical problem that, with static confidence thresholds $\tau = 0.4, 0.5, 0.6$, the number of pseudo labels keeps going up as the detector becomes more confident. GMM-based approach, on the other hand, adaptively adjusts the best threshold according to the model capacity, with a nearly constant number of GTs, which reduces temporal inconsistency. In Figure 6, we plot the estimated threshold curve obtained by GMM on COCO 1%/5%/10%. The value steadily increases as training proceeds. Furthermore, with fewer labeled samples, GMM sets a higher confidence 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.

4.2. Semi-supervised Object Detection

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

COCO-PARTIAL Results. Table 1 systematically compares the 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 a 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 the 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	10.51	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	26.07	28.69	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	33.01	37.13
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**

Table 2. COCO-ADDITION experimental results on `val2017` with `unlabel2017` as unlabeled set. Note that 1× represents 90K training iterations, and N× represents N×90K iterations.

Method	$AP_{50:95}$
CSD(3×)	$40.20 \xrightarrow{-1.38} 38.82$
STAC(6×)	$39.48 \xrightarrow{-0.27} 39.21$
Unbiased Teacher(3×)	$40.20 \xrightarrow{+1.10} 41.30$
ACRST(3×)	$40.20 \xrightarrow{+2.59} 42.79$
Soft Teacher(16×)	$40.90 \xrightarrow{+3.70} 44.50$
DSL(2×)	$40.20 \xrightarrow{+3.60} 43.80$
PseCo(8×)	$41.00 \xrightarrow{+5.10} 46.10$
Dense Teacher(8×)	$41.24 \xrightarrow{+4.88} 46.12$
`Consistent-Teacher` (8×)	$40.50 \xrightarrow{\mathbf{+7.20}} \mathbf{47.70}$

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 [4, 25].

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 `unlabel2017` as unlabeled data. As shown in Table 2, in the case of COCO-ADDITION, our model achieves 47.7 mAP, surpassing all previous state-of-the-art works.

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	81.29	56.87
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	$AP^{1\times}_{50:95}$	$AP^{10\%}_{50:95}$
IoU-based	38.4	35.50
our ASA	40.1 (+1.7)	38.50 (+3.0)

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 a fixed confidence threshold of 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 applica-

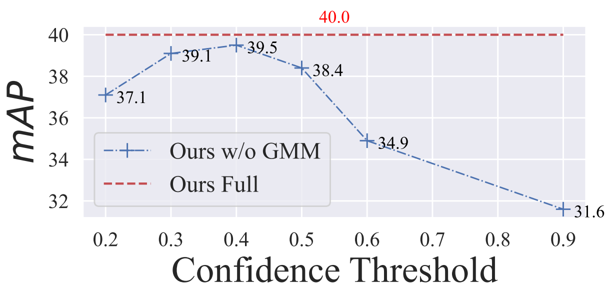

Figure 7. Ablative study of GMM-based pseudo-label filtering. Each value represents the mAP score on COCO 10% data.

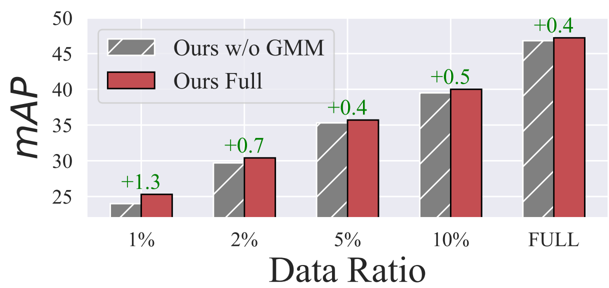

Figure 8. Ablation of GMM at different data ratio on COCO. Models are compared to baselines with a hard threshold 0.4.

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 1280×800.

Method	FLOPs (G)	$AP^{1\times}_{50:95}$	$AP^{10\%}_{50:95}$
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)

ble 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, a 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 benefit 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 to the effectiveness of FAM, we first replace the FAM-3D as a 2-D counterpart, which is adopted in [9]. Table 5 provides the ablative study of our method with different FAM structures. 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 is better calibrated to identify high-quality pseudo-labels. It is also noted that our FAM-3D brings much more gains under a semi-supervised setting than that in fully-supervised learning, validating its extra benefit in reducing the noises in SSOD.

GMM-based thresholding. We testify to the detector's performance with or without the GMM-based pseudo-labeling. 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 detector is highly sensitive to the 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 not only frees us from the tedious threshold tuning process but also allows for a gained accuracy and stabler 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 a hard threshold (0.4) at different ratios of labeled data as well, as illustrated in Figure 8.

5. Limitations and Future Work

Despite the effectiveness of `Consistent-Teacher`, it is currently mainly developed on traditional single-stage detectors. Its application to two-stage detectors and recent DETR-based [2] detectors is to be verified. Moreover, semi-supervised learning with pseudo-labels can accumulate errors due to inaccurate priors and human heuristics during the self-recurrent process. Our adaptive sample assignment strategy has replaced some human heuristics, such as anchor-based assignments, resulting in additional benefits for SSOD. It is believed that exploring more end-to-end approaches to semi-supervised learning could also bring similar advantages, which is an avenue for future research.

6. Conclusion

This paper offers a systematic investigation of the inconsistency issues that arise in SSOD, and proposes a straightforward yet effective semi-supervised object detector called `Consistent-Teacher` as a solution. The proposed method employs adaptive anchor assignment, which identifies the positive anchor with the lowest matching costs, and FAM, which aligns classification and regression tasks by regressing the 3-D feature pyramid offsets. To address the threshold inconsistency problem in pseudo-bboxes, GMM is utilized to dynamically adjust the threshold for self-training. By integrating these three modules, our `Consistent-Teacher` achieves a significant performance improvement over state-of-the-art methods on various SSOD benchmarks, demonstrating robust anchor assignment and consistent pseudo-bboxes.

References

[1] 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. 3

[2] 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. 4, 8

[3] Binbin Chen, Weijie Chen, Shicai Yang, Yunyi Xuan, Jie Song, Di Xie, Shiliang Pu, Mingli Song, and Yueting Zhuang. Label matching semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14381–14390, 2022. 1

[4] 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. 1, 3, 5, 7

[5] 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. 1

[6] 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. 3

[7] 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. 5

[8] 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. 5

[9] 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. 2, 4, 8

[10] 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. 2, 3, 4

[11] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 2, 4

[12] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. *Advances in neural information processing systems*, 32, 2019. 1, 3, 5

[13] 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. 1

[14] 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. 3

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

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

[17] 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. 1, 6

[18] 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. 3

[19] 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. 3, 5

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3, 5

[21] 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. 5

[22] 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. 4

[23] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 3

[24] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*, 2021. 1, 2, 3, 5

[25] Yen-Cheng Liu, Chih-Yao Ma, and Zsolt Kira. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9819–9828, 2022. 1, 5, 7

[26] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object de-

tection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 3

[27] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 3

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3

[29] 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. 3

[30] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 1, 2, 3, 5

[31] 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. 3, 5

[32] 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. 2, 3, 6

[33] 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. 3

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

[35] 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. 1

[36] 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. 1, 2, 3, 4, 5, 6

[37] 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. 3

[38] Xingyi Yang, Jingwen Ye, and Xinchao Wang. Factorizing knowledge in neural networks. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pages 73–91. Springer, 2022. 3

[39] Fangyuan Zhang, Tianxiang Pan, and Bin Wang. Semi-supervised object detection with adaptive class-rebalancing self-training. *arXiv preprint arXiv:2107.05031*, 2021. 5

[40] 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. 3, 6

[41] 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. 3

[42] 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. 1, 3, 5

[43] 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. 1, 3, 5

[44] Qiang Zhou, Chaohui Yu, Zhibin Wang, Qi Qian, and Hao Li. Instant-teaching: An end-to-end semi-supervised object detection framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4081–4090, 2021. 1, 3, 5

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