PseCo: Pseudo Labeling and Consistency Training for Semi-Supervised Object Detection

Gang Li[1,2], Xiang Li[1]*, Yujie Wang[2], Yichao Wu[2], Ding Liang[2], and Shanshan Zhang[1]*

[1] Nanjing University of Science and Technology
[2] SenseTime Research
{gang.li, xiang.li.implus, shanshan.zhang}@njust.edu.cn
{wangyujie,wuyichao,liangding}@sensetime.com

Abstract. In this paper, we delve into two key techniques in Semi-Supervised Object Detection (SSOD), namely pseudo labeling and consistency training. We observe that these two techniques currently neglect some important properties of object detection, hindering efficient learning on unlabeled data. Specifically, for pseudo labeling, existing works only focus on the classification score yet fail to guarantee the localization precision of pseudo boxes; For consistency training, the widely adopted random-resize training only considers the label-level consistency but misses the feature-level one, which also plays an important role in ensuring the scale invariance. To address the problems incurred by noisy pseudo boxes, we design Noisy Pseudo box Learning (NPL) that includes Prediction-guided Label Assignment (PLA) and Positive-proposal Consistency Voting (PCV). PLA relies on model predictions to assign labels and makes it robust to even coarse pseudo boxes; while PCV leverages the regression consistency of positive proposals to reflect the localization quality of pseudo boxes. Furthermore, in consistency training, we propose Multi-view Scale-invariant Learning (MSL) that includes mechanisms of both label- and feature-level consistency, where feature consistency is achieved by aligning shifted feature pyramids between two images with identical content but varied scales. On COCO benchmark, our method, termed PSEudo labeling and COnsistency training (PseCo), outperforms the SOTA (Soft Teacher) by 2.0, 1.8, 2.0 points under 1%, 5%, and 10% labelling ratios, respectively. It also significantly improves the learning efficiency for SSOD, e.g., PseCo halves the training time of the SOTA approach but achieves even better performance. Code is available at https://github.com/ligang-cs/PseCo.

Keywords: Semi-supervised Learning, Object Detection

1 Introduction

With the rapid development of deep learning, many computer vision tasks achieve significant improvements, such as image classification [3,2], object detection [16,1,10],
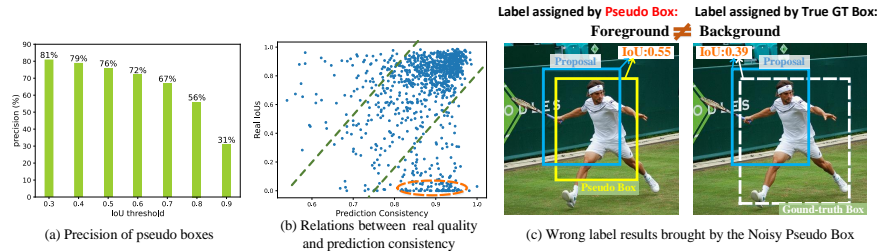
---

* Corresponding author.

Fig. 1: (a) The precision of pseudo boxes under various IoU thresholds. (b) The scatter diagram of the relation between the prediction consistency and their true localization quality. Some dots falling in the orange ellipse are caused by annotation errors. Examples are attached in *supplemental material*. (c) One specific example to demonstrate that noisy pseudo boxes will mislead label assignment.

etc. Behind these advances, plenty of annotated data plays an important role [24]. However, labeling accurate annotations for large-scale data is usually time-consuming and expensive, especially for object detection, which requires annotating precise bounding boxes for each instance, besides category labels. Therefore, employing easily accessible unlabeled data to facilitate the model training with limited annotated data is a promising direction, named Semi-Supervised Learning, where labeled data and unlabeled data are combined together as training examples.

Semi-Supervised for Image Classification (SSIC) has been widely investigated in previous literature, and the learning paradigm on unlabeled data can be roughly divided into two categories: pseudo labeling [8,19] and consistency training [25,23], each of which receives much attention. Recently, some works (e.g., FixMatch [20], FlexMatch [29]) attempt to combine these two techniques into one framework and achieve state-of-the-art performance. In Semi-Supervised Object Detection (SSOD), some works borrow the key techniques (e.g. pseudo labeling, consistency training) from SSIC, and directly apply them to SSOD. Although these works [31,27] obtain gains from unlabeled data, they neglect some important properties of object detection, resulting in sub-optimal results. On the one hand, compared with image classification, pseudo labels of object detection are more complicated, containing both category and location information. On the other hand, object detection is required to capture stronger scale-invariant ability than image classification, as it needs to carefully deal with the targets with rich scales. In this work, we present a SSOD framework, termed PSEudo labeling and COnsistency training (PseCo), to integrate object detection properties into SSOD, making pseudo labeling and consistency training work better for object detection tasks.

In pseudo labeling, the model produces one-hot pseudo labels on unlabeled data by itself, and only pseudo labels whose scores are above the predefined score threshold are retained. As for object detection, the pseudo label consists of both category labels and bounding boxes. Although category labels can be guaranteed

to be accurate via setting a high score threshold, the localization quality of pseudo box fails to be measured and guaranteed. It has been validated in previous works that the classification score is not strongly correlated with the precision of box localization [11,30,7,27]. In Fig. 1(a), we compute the precision of pseudo boxes under various Intersection-over-Union (IoU) thresholds, via comparing produced pseudo boxes with ground-truths. Under loose criterion (IoU=0.3), precision can reach 81%, but it will drop to 31% when we lift the IoU threshold to 0.9. This dramatic precision gap indicates coarse pseudo boxes whose IoUs belong to [0.3,0.9] account for 50%. If these noisy pseudo boxes are used as targets to train the detector, it must hinder the optimization, resulting in slow convergence and inefficient learning on unlabeled data. Furthermore, we analyze the negative effects brought by noisy pseudo boxes on classification and regression tasks as follows, respectively.

For the classification task, noisy pseudo boxes will mislead the label assignment, where labels are assigned based on IoUs between proposals and gt boxes (pseudo boxes in our case). As shown in Fig. 1(c), a background proposal is taken as foreground due to a large IoU value with a poorly localized pseudo box. As a result, the IoU-based label assignment will fail on unlabeled data and confuse decision boundaries between foreground and background. To address this issue, we design a prediction-guided label assignment strategy for unlabeled data, which assigns labels based on predictions of the teacher, instead of IoUs with pseudo boxes as before, making it robust for poorly localized pseudo boxes.

For the regression task, it is necessary to measure the localization quality of pseudo boxes. We propose a simple yet effective method to achieve this, named Positive-proposal Consistency Voting. We empirically find that regression consistency from positive proposals is capable of reflecting the localization quality of corresponding pseudo boxes. In Fig. 1(b), we visualize the relations between predicted consistency and their true IoUs, where their positive correlations can be found. Therefore, it is reasonable to employ the estimated localization quality (i.e., regression consistency from positive proposals) to re-weight the regression losses, making precise pseudo boxes contribute more to regression supervisions.

Apart from pseudo labeling, we also analyze the consistency training for SSOD. Consistency training enforces the model to generate similar predictions when fed with perturbed versions of the same image, where perturbations can be implemented by injecting various data augmentations. Through consistency training, models can be invariant to different input transformations. Current SSOD methods [31,27,15] only apply off-the-shelf, general data augmentations, most of which are borrowed from image classification. However, different from classification, object detection is an instance-based task, where object scales usually vary in a large range, and detectors are expected to handle all scale ranges. Therefore, learning strong scale-invariant ability via consistency training is important. In scale consistency, it should be allowed for the model to predict the same boxes for input images with identical contents but varied scales. To ensure label consistency, random-resizing is a common augmentation, which resizes input images and gt boxes according to a randomly generated resize ratio. Be-

sides label consistency, feature consistency also plays an important role in scale-invariant learning, but it is neglected in previous works. Thanks to the pyramid structure of popular backbone networks, feature alignment can be easily implemented by shifting feature pyramid levels according to the scale changes. Motivated by this, we introduce a brand new data augmentation technique, named Multi-view Scale-invariant Learning (MSL), to learn label-level and feature-level consistency simultaneously in a simple framework.

In summary, we delve into two key techniques of semi-supervised learning (e.g., pseudo labeling and consistency training) for SSOD, and integrate object detection properties into them. On COCO benchmarks, our PseCo outperforms the state-of-the-art methods by a large margin, for example, under 10% labelling ratio, it can improve a 26.9% mAP baseline to 36.1% mAP, surpassing previous methods by at least 2.0%. When labeled data is abundant, i.e., we use full COCO training set as labeled data and extra 123K unlabeled2017 as unlabeled data, our PseCo improves the 41.0% mAP baseline by +5.1%, reaching 46.1% mAP, establishing a new state of the art. Moreover, PseCo also significantly boosts the convergence speed, e.g. PseCo halves the training time of the SOTA (Soft Teacher [27]), but achieves even better performance.

2   Related Works

Semi-supervised learning in image classification. Semi-supervised learning can be categorized into two groups: pseudo labeling (also called self-training) and consistency training, and previous methods design learning paradigms based on one of them. Pseudo labeling [8,19,5,26] iteratively adds unlabeled data into the training procedure with pseudo labels annotated by an initially trained network. Here, only model predictions with high confidence will be transformed into the one-hot format and become pseudo labels. Noisy Student Training [26] injects noise into unlabeled data training, which equips the model with stronger generalization through training on the combination of labeled and unlabeled data. On the other hand, consistency training [23,25,1] relies on the assumption that the model should be invariant to small changes on input images or model hidden states. It enforces the model to make similar predictions on the perturbed versions of the same image, and perturbations can be implemented by injecting noise into images and hidden states. UDA [25] validates the advanced data augmentations play a crucial role in consistency training, and observes the strong augmentations found in supervised-learning can also lead to obvious improvements in semi-supervised learning.

Recently, some works [20,29] attempt to combine pseudo labeling and consistency training, achieving state-of-the-art performance. FixMatch [20] firstly applies the weak and strong augmentations to the same input image, respectively, to generate two versions, then uses the weakly-augmented version to generate hard pseudo labels. The model is trained on strongly-augmented versions to align predictions with pseudo labels. Based on FixMatch, FlexMatch [29] proposes to adjust score thresholds for different classes during the generation of pseudo

labels, based on curriculum learning. It has been widely validated that pseudo labeling and consistency training are two powerful techniques in semi-supervised image classification, hence in this work, we attempt to integrate object detection properties into them and make them work better for semi-supervised object detection.

Semi-supervised learning in object detection. STAC [21] is the first attempt to apply pseudo labeling and consistency training based on the strong data augmentations to semi-supervised object detection, however, it adopts two stages of training as Noisy Student Training [26], which prevents the pseudo labels from updating along with model training and limits the performance. After STAC, [27,31,22,28,15] borrow the idea of Exponential Moving Average (EMA) from Mean Teacher [23], and update the teacher model after each training iteration to generate instant pseudo labels, realizing the end-to-end framework. To pursue high quality of pseudo labels and overcome confirmation bias, Instant-Teaching [31] and ISMT [28] introduce model ensemble to aggregate predictions from multiple teacher models which are initialized differently; similarly, Humble Teacher [22] ensembles the teacher model predictions by taking both the image and its horizontally flipped version as input. Although these ensemble methods can promote the quality of pseudo labels, they also introduce considerable computation overhead. Unbiased Teacher [15] replaces traditional Cross-entropy loss with Focal loss [13] to alleviate the class-imbalanced pseudo-labeling issue, which shows strong performance when labeled data is scarce. Soft Teacher [27] uses teacher classification scores as classification loss weights, to suppress negative effects from underlying objects missed by pseudo labels. Different from previous methods, our work elaborately analyzes whether the pseudo labeling and consistency training can be directly applied to SSOD, but gets a negative answer. To integrate object detection properties into these two techniques, we introduce Noisy Pseudo box Learning and Multi-view Scale-invariant Learning, obtaining much better performance and faster convergence speed.

3   Method

We show the framework of our PseCo in Fig. 2. On the unlabeled data, PseCo consists of Noisy Pseudo box Learning (NPL) and Multi-view Scale-invariant Learning (MSL). In the following parts, we will introduce the basic framework, the proposed NPL and MSL, respectively.

3.1   The basic framework

At first, we directly apply standard pseudo labeling and consistency training to SSOD, building our basic framework. Following previous works [27,15,31], we also adopt Teacher-student training scheme, where the teacher model is built from the student model at every training iteration via Exponential Moving Average (EMA). We randomly sample labeled data and unlabeled data based on a sample
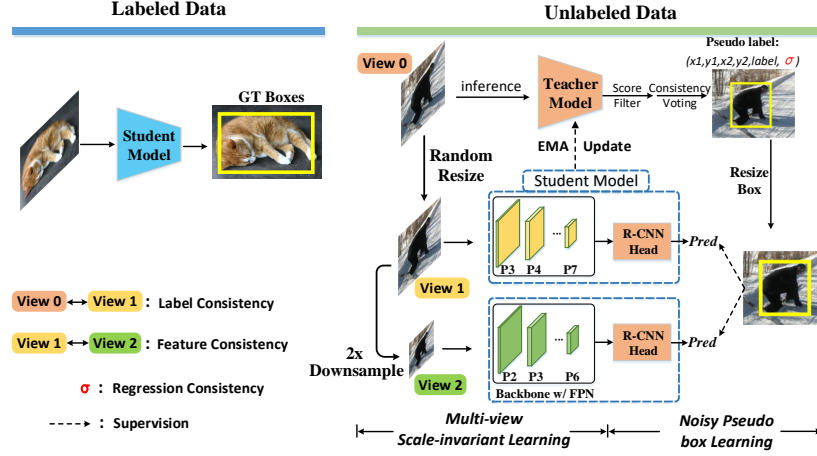
Fig. 2: The framework of our PseCo. Each training batch consists of both labeled and unlabeled images. On the unlabeled images, the student model trains on view V_1 and V_2 at the same time, taking the same pseudo boxes as supervisions. View V_0 refers to input images for the teacher model.

ratio to form the training batch. On the labeled data, the student model is trained in a regular manner, supervised by the ground-truth boxes:

$$\mathcal{L}^l = \mathcal{L}^l_{cls} + \mathcal{L}^l_{reg}. \tag{1}$$

On the unlabeled data, we firstly apply weak data augmentations (e.g. horizontal flip, random resizing) to input images, and then feed them to the teacher model for pseudo label generation. Considering the detection boxes tend to be dense even after NMS, we set a score threshold τ and only retain boxes with scores above τ as pseudo labels. After that, strong augmentations (e.g. cutout, rotation, brightness jitter)[3] will be performed on the input image to generate the training example for student model. Since high classification scores do not lead to precise localization, we abandon bounding box regression on unlabeled data, as done in [15]. Actually, applying the box regression loss on unlabeled data will cause unstable training in our experiments.

Foreground-background imbalance [9,13] is an intrinsic issue in object detection, and it gets worse under the semi-supervised setting. A high score threshold τ is usually adopted to guarantee the precision of pseudo labels, but it also results in scarcity of pseudo labels, aggravating the imbalance of foreground/background. Moreover, there also exists foreground-foreground imbalance, exactly, training examples from some specific categories can be limited when labeled data is scarce, which makes the model prone to predict the dominant classes, causing biased prediction. To alleviate these imbalance issues, we

---

[3] We adopt the same data augmentations as Soft Teacher [27], please refer to [27] for more augmentation details.

follow the practice of Unbiased Teacher [15], and replace the standard cross-entropy loss with focal loss [13]:

$$\mathcal{L}_{cls}^{u} = -\alpha_t(1-p_t)^{\gamma}log(p_t), \ p_t = \begin{cases} p, & if \ y = 1, \\ 1-p, & otherwise, \end{cases} \tag{2}$$

where parameters α_t and γ adopt default settings in original focal loss paper [13]. The overall loss function is formulated as:

$$\mathcal{L} = \mathcal{L}^{l} + \beta\mathcal{L}^{u}, \tag{3}$$

where β is used to control the contribution of unlabeled data. In theory, our proposed method is independent of the detection framework and can be applied on both one-stage and two-stage detectors. However, considering all previous methods are based on Faster R-CNN [18] detection framework, for a fair comparison with them, we also adopt Faster R-CNN as the default detection framework.

3.2 Noisy Pseudo Box Learning

In SSOD, pseudo labels contain both category and location. Since the score of pseudo labels can only indicate the confidence of pseudo box categories, the localization quality of pseudo boxes is not guaranteed. Imprecise pseudo boxes will mislead the label assignment and regression task, making learning on unlabeled data inefficient. Motivated by this, we introduce Prediction-guided Label Assignment and Positive-proposal Consistency Voting to reduce negative effects on the label assignment and regression task, respectively.

Prediction-guided Label Assignment. The standard label assignment strategy in Faster R-CNN [18] only takes the IoUs between proposals and gt boxes (pseudo boxes in our case) into consideration and assigns foreground to those proposals, whose IoUs are above a pre-defined threshold t (0.5 as default). This strategy relies on the assumption that gt boxes are precise, however, this assumption does not hold for unlabeled data obviously. As a result, some low-quality proposals will be mistakenly assigned as positive, confusing the classification boundaries between foreground and background. One specific example is shown in Fig. 1(c), where a proposal with the true IoU as 0.39 is mistakenly assigned as positive.

To address this problem, we propose Prediction-guided Label Assignment (PLA), which takes teacher predictions as auxiliary information and reduces dependency on IoUs. In Teacher-student training scheme, not only can the detection results (after NMS) of teacher perform as pseudo labels, but also teacher's dense predictions (before NMS) are able to provide guidance for student model training. We share the proposals generated by the teacher RPN with the student, so that teacher predictions on these proposals can be easily transferred to student. To measure the proposal quality (q) comprehensively, the classification confidence and localization precision of teacher predictions are jointly employed, concretely, $q = s^{\alpha} \times u^{1-\alpha}$, where s and u denote a foreground score and an IoU

value between the regressed box and the ground truth, respectively. α controls the contribution of s and u in the overall quality. On unlabeled data, we first construct a candidate bag for each ground truth g by the traditional IoU-based strategy, where the IoU threshold t is set to a relatively low value, e.g., 0.4 as default, to contain more proposals. Within each candidate bag, the proposals are firstly sorted by their quality q, then top-$\mathcal{N}$ proposals are adopted as positive samples and the rest are negatives. The number $\mathcal{N}$ is decided by *the dynamic k estimation* strategy proposed in OTA [4], specifically, the IoU values over the candidate bag is summed up to represent the number of positive samples. The proposed PLA gets rid of strong dependencies on IoUs and alleviates negative effects from poorly localized pseudo boxes, leading to clearer classification boundaries. Furthermore, our label assign strategy integrates more teacher knowledge into student model training, realizing better knowledge distillation.

Positive-proposal Consistency Voting. Considering the classification score fails to indicate localization quality, we introduce a simple yet effective method to measure the localization quality, named Positive-proposal Consistency Voting (PCV). Assigning multiple proposals to each gt box (or pseudo box) is a common practice in CNN-based detectors [18,11,30], and we observe that the consistency of regression results from these proposals is capable of reflecting the localization quality of the corresponding pseudo box. Regression consistency σ^j for pseudo box (indexed by j) is formulated as:

$$\sigma^j = \frac{\sum_{i=1}^{N} u_i^j}{N},$$

(4)

where u denotes an IoU value between the predicted box and the pseudo box, as defined above; N denotes the number of positive proposals, assigned to the pseudo box j. After obtaining σ^j, we employ it as the instance-wise regression loss weight:

$$\mathcal{L}_{reg}^u = \frac{1}{MN} \sum_{j=1}^{M} \sigma^j \sum_{i=1}^{N} |reg_i^j - r\hat{e}g_i^j|,$$

(5)

where reg and $r\hat{e}g$ refer to the regression output and ground-truth, respectively. In Fig. 1(b), we depict the scatter diagram of the relation between prediction consistency σ of pseudo boxes and their true IoUs. It is obvious that σ is positively correlated with true IoUs. Note that, some dots falling in the orange ellipse are mainly caused by annotation errors. We visualize some examples in *supplemental material*, where the pseudo boxes accurately detect some objects, which are missed by the ground truths.

3.3   Multi-view Scale-invariant Learning

Different from image classification, in object detection, object scales vary in a large range and detectors hardly show comparable performance on all scales. Therefore, learning scale-invariant representations from unlabeled data is considerably important for SSOD. In consistency training, strong data augmentations
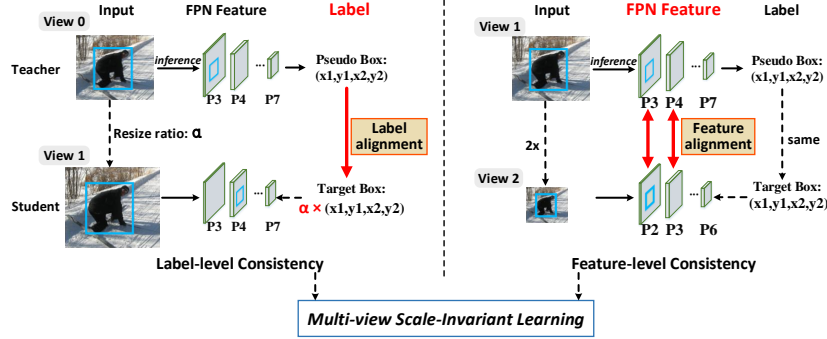
Fig. 3: Comparisons between label-level consistency learning and feature-level consistency learning. For label consistency, labels are aligned according to the resize ratio α; for feature consistency, features are aligned by shifting the feature pyramid level.

play a crucial role [25,26] in achieving competitive performance. Through injecting the perturbations into the input images, data augmentations equip the model with robustness to various transformations. From the perspective of scale invariance, we regard the common data augmentation strategy (e.g. random-resizing) as label-level consistency since it resizes the label according to the scale changes of input images. Unfortunately, existing works only involve the widely adopted label-level consistency but fail to consider the feature-level one. Since detection network usually has designs of rich feature pyramids, feature-level consistency is easy to implement across paired inputs [17] and should be considered seriously. In this paper, we propose Multi-view Scale-invariant Learning (MSL) that combines both label- and feature-level consistency into a simple framework, where feature-level consistency is realized by aligning shifted pyramid features between two images with identical content but different scales.

To be specific, two views, namely V_1 and V_2, are used for student training in MSL. We denote the input image for the teacher model as V_0. Views V_1 and V_2 are constructed to learn label- and feature-level consistency, respectively. Among them, V_1 is implemented by vanilla random resizing, which rescales the input V_0 and pseudo boxes according to a resize ratio α randomly sampled from the range $[\alpha_{min}, \alpha_{max}]$ ([0.8, 1.3] as default). For feature consistency learning, we firstly downsample V_1 by even number times (2x as default) to produce V_2, then combine V_1 and V_2 into image pairs. Upsampling is also certainly permitted, but we only perform downsampling here for GPU memory restriction. Because the spatial sizes of adjacent FPN layers always differ by 2x, the P3-P7 layers[4] of V_1 can align well with P2-P6 layers of V_2 in the spatial dimension. Through feature alignment, the same pseudo boxes can supervise the student model training on both V_1 and V_2. Integrating label consistency and feature consistency into consistency learning leads to stronger scale-invariant learning and significantly

---

[4] P_x refers to the FPN layer whose feature maps are downsampled by 2^x times.

accelerates model convergence, as we will show later in the experiments. Comparisons between label consistency and feature consistency are shown in Fig. 3.

Learning scale-invariant representation from unlabeled data is also explored by SoCo [24]. However, we claim there are two intrinsic differences between MSL and SoCo: (1) MSL models scale invariance from both label consistency and image feature consistency, while SoCo only considers object feature consistency. Through aligning **dense image features** of shifted pyramids between paired images, our MSL can provide more comprehensive and dense supervisory signals than the SoCo, which only performs consistency on **sparse objects**. (2) SoCo implements feature consistency via contrastive learning, which is designed for the pretraining; in contrast, our MSL uses bounding box supervision to implement consistency learning and can be integrated into the detection task.

4   Experiments

4.1   Dataset and Evaluation Protocol

In this section, we conduct extensive experiments to verify the effectiveness of PseCo on MS COCO benchmark [14]. There are two training sets, namely the train2017 set, containing 118k labeled images, and the unlabeled2017 set, containing 123k unlabeled images. The val2017 with 5k images is used as validation set, and we report all experiment results on val2017. The performance is measured by COCO average prevision (denoted as mAP). Following the common practice of SSOD [21], there are two experimental settings: **Partially Labeled Data** and **Fully Labeled Data**, which are described as follows:
Partially Labeled Data. We randomly sample 1, 2, 5, and 10% data from train2017 as labeled data, and use the rest as unlabeled. Under each labelling ratio, we report the mean and standard deviation over 5 different data folds.
Fully Labeled Data. Under this setting, we take train2017 as the training labeled set and unlabeled2017 as the training unlabeled set.

4.2   Implementation Details

For a fair comparison, we adopt Faster R-CNN [18] with FPN [12] as the detection framework, and ResNet-50 [6] as the backbone. The confidence threshold τ is set to 0.5, empirically. We set β as 4.0 to control contributions of unlabeled data in the overall losses. The performance is evaluated on the Teacher model. Training details for **Partially Labeled Data** and **Fully Labeled Data** are described below:
Partially Labeled Data. All models are trained for 180k iterations on 8 GPUs. The initial learning rate is set as 0.01 and divided by 10 at 120k and 160k iterations. The training batch in each GPU includes 5 images, where the sample ratio between unlabeled data and labeled data is set to 4:1.
Fully Labeled Data. All models are trained for 720k iterations on 8 GPUs. Mini-batch in each GPU is 8 with the sample ratio between unlabeled and labeled data as 1:1. The learning rate is initialized to 0.01 and divided by 10 at 480k and 680k iterations.

Table 1: Comparisons with the state-of-the-art methods on val2017 set under the **Partially Labeled Data** and **Fully Labeled Data** settings.

Method	Partially Labeled Data				Fully Labeled Data
	1%	2%	5%	10%	
Supervised baseline	12.20±0.29	16.53±0.12	21.17±0.17	26.90±0.08	41.0
STAC [21]	13.97±0.35	18.25±0.25	24.38±0.12	28.64±0.21	39.5 $\xrightarrow{-0.3}$ 39.2
Humble Teacher [22]	16.96±0.35	21.74±0.24	27.70±0.15	31.61±0.28	37.6 $\xrightarrow{+4.8}$ 42.4
ISMT [28]	18.88±0.74	22.43±0.56	26.37±0.24	30.53±0.52	37.8 $\xrightarrow{+1.8}$ 39.6
Instant-Teaching [31]	18.05±0.15	22.45±0.15	26.75±0.05	30.40±0.05	37.6 $\xrightarrow{+2.6}$ 40.2
Unbiased Teacher [15]	20.75±0.12	24.30±0.07	28.27±0.11	31.50±0.10	40.2 $\xrightarrow{+1.1}$ 41.3
Soft Teacher [27]	20.46±0.39	-	30.74±0.08	34.04±0.14	40.9 $\xrightarrow{+3.6}$ 44.5
PseCo (ours)	**22.43**±0.36	**27.77**±0.18	**32.50**±0.08	**36.06**±0.24	41.0 $\xrightarrow{+5.1}$ **46.1**

4.3   Comparison with State-of-the-Art Methods

We compare the proposed PseCo with other state-of-the-art methods on COCO val2017 set. Comparisons under the **Partially Labeled Data** setting are first conducted, with results reported in Tab. 1. When labeled data is scarce (i.e., under 1% and 2% labelling ratios), our method surpasses the state-of-the-art method, Unbiased Teacher [15], by 1.7% and 3.5%, reaching 22.4 and 27.8 mAP, respectively. When more labeled data is accessible, the SOTA method is transferred to Soft Teacher [27]. Our method still outperforms it by 1.8% and 2.0% under 5% and 10% labelling ratios, respectively. Therefore, the proposed method outperforms the SOTAs by a large margin, at least 1.7%, under all labelling ratios. Compared with the supervised baseline, PseCo obtains even better performance with only 2% labeled data than the baseline with 10% labeled data, demonstrating the effectiveness of proposed semi-supervised learning techniques.

Moreover, we also compare the convergence speed with the previous best method (Soft Teacher [27]) in Fig. 4, where convergence curves are depicted under 10% and 5% labelling ratios. It is obvious that our method has a faster convergence speed, specifically, our method uses only 2/5 and 1/4 iterations of Soft Teacher to achieve the same performance under 10% and 5% labelling ratios respectively. Although we employ an extra view (V_2) to learn feature-level consistency, it only increases the training time of each iteration by 25% (from 0.72 sec/iter to 0.91 sec/iter), due to the low input resolution of V_2. In summary, we halve the training time of SOTA approach but achieve even better performance, which validates the superior learning efficiency of our method on unlabeled data.

The experimental results under the **Fully Labeled Data** setting are reported in Tab. 1, where both results of comparison methods and their supervised baseline are listed. Following the practice in Soft Teacher [27], we also apply weak augmentations to the labeled data and obtain a strong supervised baseline, 41.0 mAP. Although with a such strong baseline, PseCo still achieves larger improvements (+5.1%) than others and reaches 46.1 mAP, building a new state of the art. Some qualitative results are attached in *supplemental material*.
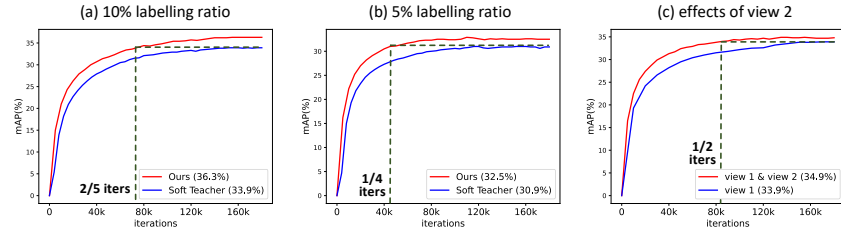
Fig. 4: Comparison of model convergence speed. In (a) and (b), we compare PseCo against Soft Teacher [27]. Here, we reproduce Soft Teacher using their source codes. (c) depicts the comparison between V_1 and $V_1\&V_2$. In legend, the numbers in brackets refer to mAP. Performance is evaluated on the teacher.

Table 2: Ablation studies on each component of our method. MSL represents Multi-view Scale-invariant Learning; NPL represents Noisy Pseudo box Learning. In MSL, V_1 and V_2 are constructed for label- and feature-level consistency, respectively. In NPL, PCV and PLA stand for Positive-proposal Consistency Voting and Prediction-guided Label Assignment, respectively.

MSL		NPL		mAP	AP_{50}	AP_{75}
V_1	V_2	PCV	PLA			
				26.8	44.9	28.4
✓				33.9(+7.1)	55.2	36.0
✓	✓			34.9(+8.1)	56.3	37.1
✓		✓		34.8(+8.0)	55.1	37.4
✓		✓	✓	35.7(+8.9)	56.4	38.4
✓	✓	✓		36.0(+9.2)	56.9	38.7
✓	✓	✓	✓	**36.3**(+9.5)	**57.2**	**39.2**

4.4   Ablation Study

We conduct detailed ablation studies to verify key designs. All ablation studies are conducted on a single data fold from the 10% labelling ratio.

Effect of individual component. In Tab. 2, we show effectiveness of each component step by step. When only using 10% labeled data as training examples, it obtains 26.8 mAP. Next, we construct the semi-supervised baseline by applying V_1 on unlabeled data for label-level consistency learning. The baseline does not consider any adverse effects incurred by coarse pseudo boxes and obtains 33.9 mAP. Furthermore, by leveraging additional view V_2, the feature-level scale-invariant learning is enabled, and an improvement of +1.0 mAP is found. On the other hand, to alleviate the issue of coarse pseudo boxes, we introduce PCV to suppress the inaccurate regression signals, improving the baseline from 33.9 to 34.8 mAP. After that, we replace the traditional IoU-based label assignment strategy with the PLA and enjoy another +0.9 mAP gain. Finally, when combing MSL and NPL together, it achieves the best performance, 36.3 mAP.

Table 3: Analysis of Multi-view Scale-invariant learning, which contains both the label- and feature-level consistency.

(a) Study on label consistency.

method	mAP	AP_S	AP_M	AP_L
single-scale training	32.7	19.0	36.0	42.5
label consistency	**33.9**	**19.1**	**37.2**	**44.4**

(b) Study on feature consistency.

method	mAP	AP_S	AP_M	AP_L
vanilla multi-view training	33.9	20.9	37.2	43.0
feature consistency	**34.9**	**22.1**	**38.2**	**43.6**

Table 4: Ablation studies related to Positive-proposal Consistency Voting (PCV) and Prediction-guided Label Assignment (PLA).

(a) Comparison between our PCV and other regression methods.

method	mAP	AP_{50}	AP_{75}
abandon reg [15]	33.9	55.2	36.0
reg consistency [22]	34.2	55.1	36.5
box jittering [27]	34.5	54.9	36.9
PCV (ours)	**34.8**	**55.1**	**37.4**

(b) Study on hyper-parameter α.

α	mAP	AP_{50}	AP_{75}
0	35.2	56.1	37.8
0.5	**35.7**	**56.4**	**38.4**
1.0	35.4	55.7	38.4

(c) Study on IoU threshold t.

t	mAP	AP_{50}	AP_{75}
0.3	35.7	56.2	**38.6**
0.4	**35.7**	**56.4**	38.4
0.5	35.5	56.1	38.3

Comparison with other regression methods. Scores of pseudo boxes can only indicate the confidence of predicted object category, thus they fail to reflect localization quality [11,15]. Naive confidence thresholding will introduce some coarse bounding boxes for regression tasks. To alleviate this issue, Unbiased Teacher [15] abandons regression losses on unlabeled data (denoted as "abandon reg"); Humble Teacher [22] aligns the regression predictions between the teacher and student on selected top-$\mathcal{N}$ proposals (dubbed "reg consistency"); Soft Teacher [27] introduces the box jittering to calculate prediction variance on jittered pseudo boxes, which is used to filter out poorly localized pseudo boxes. In Tab. 4a, we compare our Positive-proposal Consistency Voting (PCV) with these methods. PCV obtains the best performance, concretely, on AP_{75}, PCV surpasses two competitors, reg consistency and box jittering, by 0.9% and 0.5%, respectively. Although both PCV and box jittering [27] rely on prediction variance, there exist great differences. Firstly, PCV produces localization quality by intrinsic proposals, thus it avoids extra network forward on jittered boxes, enjoying higher training efficiency. Moreover, unlike the box jittering, which meticulously tunes the variance threshold, PCV is free of hyper-parameters.

Study on different hyper-parameters of PLA. We first investigate the performance using different α in PLA, which balances the influence of classification score (s) and localization precision (u) in the proposal quality. Through a coarse search shown in Tab. 4b, we find that combining s and u yields better performance than using them individually. We then carry out experiments to study the robustness of the IoU threshold t, which is used to build the candidate bag. From the Tab. 4c, using lower t to construct a bigger candidate bag is preferred.

Analysis of Multi-view Scale-invariant Learning. We propose the MSL to model scale invariance from the aspects of both label- and feature-level consistency. The studies on them are reported in Tab. 3. At first, we construct a

single-scale training baseline without scale variance, where the input images for the teacher and student are kept on the same scale. It obtains 32.7 mAP. Next, we apply the different scale jitter on the teacher and student to implement label-level consistency, which surpasses the single-scale training by 1.2 mAP. Based on the label consistency, we further introduce the view V_2 to perform feature consistency learning. It obtains +1.0% improvements, reaching 34.9 mAP. Apart from performance gains, the feature consistency can also significantly boost the convergence speed as depicted in Fig. 4(c). To validate the improvements introduced by the V_2 come from comprehensive scale-invariant learning, instead of vanilla multi-view training, we also add an extra view V_2' besides the V_1, where V_2' is downsampled from V_1 by 2x and performs label consistency as V_1. From the Tab. 3b, vanilla multi-view training with only label consistency hardly brings improvements against the single V_1 (33.9 vs 33.9%).

Effect of Focal Loss. In Tab. 5, we compare the Cross Entropy (CE) Loss and Focal Loss. Thanks to the Focal Loss, an improvement of +0.6 mAP is achieved against the CE Loss. On the other hand, even with the CE Loss, our PseCo still surpasses the Soft Teacher by a large margin, i.e., 1.7 mAP.

Table 5: Ablation study on Focal Loss.

method	mAP	AP_{50}	AP_{75}
PseCo w/ CE Loss	35.7	55.6	38.9
PseCo w/ Focal Loss	**36.3**	**57.2**	**39.2**

5   Conclusion

In this work, we elaborately analyze two key techniques of semi-supervised object detection (e.g. pseudo labeling and consistency training), and observe these two techniques currently neglect some important properties of object detection. Motivated by this, we propose a new SSOD framework, PseCo, to integrate object detection properties into SSOD. PseCo consists of Noisy Pseudo box Learning (NPL) and Multi-view Scale-invariant Learning (MSL). In NPL, prediction-guided label assignment and positive-proposal consistency voting are proposed to perform the robust label assignment and regression task using noisy pseudo boxes, respectively. Based on the common label-level consistency, MSL additionally designs a novel feature-level scale-invariant learning, which is neglected in prior works. To validate the effectiveness of our method, extensive experiments are conducted on COCO benchmark. Experimental results validate PseCo surpasses the SOTAs by a large margin both in accuracy and efficiency.

References

1. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: A holistic approach to semi-supervised learning. Advances in Neural Information Processing Systems **32** (2019)
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Ge, Z., Liu, S., Li, Z., Yoshie, O., Sun, J.: Ota: Optimal transport assignment for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 303–312 (2021)
5. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. Advances in neural information processing systems **17** (2004)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
7. Jiang, B., Luo, R., Mao, J., Xiao, T., Jiang, Y.: Acquisition of localization confidence for accurate object detection. In: Proceedings of the European conference on computer vision (ECCV). pp. 784–799 (2018)
8. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. vol. 3, p. 896 (2013)
9. Li, B., Liu, Y., Wang, X.: Gradient harmonized single-stage detector. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8577–8584 (2019)
10. Li, G., Li, X., Wang, Y., Zhang, S., Wu, Y., Liang, D.: Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1306–1313 (2022)
11. Li, X., Lv, C., Wang, W., Li, G., Yang, L., Yang, J.: Generalized focal loss: Towards efficient representation learning for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
12. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
13. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
14. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
15. Liu, Y.C., Ma, C.Y., He, Z., Kuo, C.W., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P.: Unbiased teacher for semi-supervised object detection. arXiv preprint arXiv:2102.09480 (2021)
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)

17. Qi, L., Kuen, J., Gu, J., Lin, Z., Wang, Y., Chen, Y., Li, Y., Jia, J.: Multi-scale aligned distillation for low-resolution detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14443–14453 (2021)
18. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
19. Scudder, H.: Probability of error of some adaptive pattern-recognition machines. IEEE Transactions on Information Theory **11**(3), 363–371 (1965)
20. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in Neural Information Processing Systems **33**, 596–608 (2020)
21. Sohn, K., Zhang, Z., Li, C.L., Zhang, H., Lee, C.Y., Pfister, T.: A simple semi-supervised learning framework for object detection. arXiv preprint arXiv:2005.04757 (2020)
22. Tang, Y., Chen, W., Luo, Y., Zhang, Y.: Humble teachers teach better students for semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3132–3141 (2021)
23. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
24. Wei, F., Gao, Y., Wu, Z., Hu, H., Lin, S.: Aligning pretraining for detection via object-level contrastive learning. Advances in Neural Information Processing Systems **34** (2021)
25. Xie, Q., Dai, Z., Hovy, E., Luong, T., Le, Q.: Unsupervised data augmentation for consistency training. Advances in Neural Information Processing Systems **33**, 6256–6268 (2020)
26. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10687–10698 (2020)
27. Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., Bai, X., Liu, Z.: End-to-end semi-supervised object detection with soft teacher. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3060–3069 (2021)
28. Yang, Q., Wei, X., Wang, B., Hua, X.S., Zhang, L.: Interactive self-training with mean teachers for semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5941–5950 (2021)
29. Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., Shinozaki, T.: Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. Advances in Neural Information Processing Systems **34** (2021)
30. Zhang, H., Wang, Y., Dayoub, F., Sunderhauf, N.: Varifocalnet: An iou-aware dense object detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8514–8523 (2021)
31. Zhou, Q., Yu, C., Wang, Z., Qian, Q., Li, H.: Instant-teaching: An end-to-end semi-supervised object detection framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4081–4090 (2021)