Diverse Learner: Exploring Diverse Supervision for Semi-supervised Object Detection

Linfeng Li[1,2*†], Minyue Jiang[1*], Yue Yu[1*], Wei Zhang[1], Xiangru Lin[1], Yingying Li[1], Xiao Tan[1‡], Jingdong Wang[1], and Errui Ding[1]

[1] Baidu Inc
{jiangminyue,yuyue15,zhangwei99,liyingying05,dingerrui}@baidu.com
{xrlin.me,tanxchong}@gmail.com, {wangjingdong}@outlook.com
[2] National University of Singapore
{e0724289}@u.nus.edu

Abstract. Current state-of-the-art semi-supervised object detection methods (SSOD) typically adopt the teacher-student framework featured with pseudo labeling and Exponential Moving Average (EMA). Although the performance is desirable, many remaining issues still need to be resolved, for example: (1) the teacher updated by the student using EMA tends to lose its distinctiveness and hence generates similar predictions comparing with student and causes potential noise accumulation as the training proceeds; (2) the exploitation of pseudo labels still has much room for improvement. We present a diverse learner semi-supervised object detection framework to tackle these issues. Concretely, to maintain distinctiveness between teachers and students, our framework consists of two paired teacher-student models with diverse supervision strategy. In addition, we argue that the pseudo labels which are typically regarded as unreliable and obsoleted by many existing methods are of great value. A particular training strategy consisting of Multi-threshold Classification Loss (MTC) and Pseudo Label-Aware Erasing (PLAE) is hence designed to well explore the full set of all pseudo labels. Extensive experimental results show that our diverse learner framework outperforms the previous state-of-the-art method on the MS-COCO dataset by 2.10%, 1.50% and 0.83% when training with only 1%, 5% and 10% labeled data, demonstrating the effectiveness of our proposed framework. Moreover, our approach also performs well with larger amount of data, e.g. using full COCO training set and 123K unlabeled images from COCO, reaching a new state-of-the-art performance of 44.86% mAP.

Keywords: Semi-supervised Object Detection, Diverse Learner, Multi-threshold Loss, Pseudo Label-Aware Erasing

---

* These authors contributed equally to this work.
† This work was done when Linfeng Li was an intern at Baidu Inc.
‡ Corresponding author.

1 Introduction

Machine vision systems have witnessed a remarkable progress over the last decades in the wave of deep neural networks, including image classification [5, 6], object detection [16, 17], and image segmentation [18, 3], etc. Recent years, object detection task is dominated by the deep neural network based approaches [14, 30, 11] which require a large amount of labeled training data. However, obtaining large-scale labeled object detection data is laborious and time-consuming. To mitigate this issue, semi-supervised object detection (SSOD) is proposed [7, 20], where it exploits theoretically unlimited and cost-free unlabeled data to boost the performance of the fully-supervised object detector. Current state-of-the-art SSOD methods typically follow the teacher-student framework featured with pseudo labeling [31, 15] and exponential moving average [22]. In most existing teacher-student framework, reliable pseudo labels of the unlabeled data are selected, e.g. by thresholding the outputs of the teacher, and then they are used to train the student model. Afterwards, an EMA form strategy is employed to update the teacher model for temporally ensembling the student models in different time steps, which alleviates the detrimental effect caused by the imbalanced and noisy pseudo labels. Although the performance of this popular framework is competitive, according to our observation, there are two unresolved problems: (1) existing teacher-student frameworks suffer from erroneous pseudo labels especially in the late of training stage when the teacher and the student models become nearly identical and lose their distinctiveness. (2) the exploitation of pseudo labels is naive and more sophisticated methods are preferable. For example, STAC [20] and Unbiased Teacher [13] only exploit a single-thresholding method to pick some reliable pseudo labels and disregard all the rest.

To illustrate the first problem, we delve into the EMA updating equation, which is defined as follows,

$$\theta_{tea} = \alpha\theta_{tea} + (1 - \alpha)\theta_{stu}, \tag{1}$$

where θ_{tea} and θ_{stu} are the parameters of the teacher and the student respectively. α is the blending hyper-parameter balancing the historical teacher's parameters and current student's parameters. Normally, α is set to 0.999, accumulating more historical information for model stability concerns. However, given traditional one pair of teacher-student model, since EMA updates the teacher model in each training iteration, the model weights of the teacher becomes extremely similar to those of the student especially when the learning rate is small at the last phase of the training process. This will lead to the fact that the prediction of the teacher model and the student model become nearly identical, which hinders the teacher model from digging information from the unlabeled data to supervise the student. To address this problem, we propose a diverse learner framework consisting of two-paired teacher-student models to maintain the distinctiveness of the teacher against the student. DL introduces diverse supervision for each learner from the counterpart which is important to alleviate the less informative teacher problem during the later training process.

On the other hand, the exploitation of pseudo labeling in existing teacher-student framework is severely underexplored. Previous works [20, 13] typically use a single high value threshold to generate high confident pseudo labels and the performance depends heavily on the choice of the threshold. Even worse, the quality of the pseudo labels produced by the teacher is misaligned with the image-level erasing operator [28], a typical operation of strong augmentation used in existing weak-strong augmentation module [29, 24] of SSOD methods. Specifically, the operator may erase the entire foreground object due to the lack of ground-truth foreground information on the unlabeled images.

To make full use of pseudo labels, we divide the pseudo labels into certain and uncertain categories and propose a multi-threshold classification loss, which uses hard labels for certain pseudo labels and soft labels for uncertain pseudo labels. This ensures the high quality of pseudo labels, meanwhile, increases the number of available foreground pseudo labels. Therefore, the recall of foreground objects is enhanced without sacrificing precision. Additionally, to take full advantage of the high-quality certain foreground pseudo labels, we devise a pseudo label-aware erasing module by masking the contents of certain foreground objects in the unlabeled images according to the bounding box coordinates of the high-quality certain pseudo labels, guiding the erasing operator to become more focused on the foreground objects, which leads to a more generalized model and shows superior performance according to our experiments.

To conclude, this paper has the following contributions:

- We investigate the defects of existing EMA mechanism in SSOD and propose a diverse learner framework with diverse supervision that maintains the distinctiveness of the teacher against the student as the training proceeds.
- We introduce a more favorable pseudo labeling strategy. Specifically, we divide the pseudo labels into two different categories and propose a multi-threshold classification loss to smoothly combine high quality pseudo labels with potential foreground pseudo labels. Thanks to this strategy, our approach is capable to achieve a much higher recall rate of foreground objects at the same precision against existing SOTA methods .
- We extend the exploitation of pseudo labeling. Concretely, we introduce a simple yet efficient pseudo label-aware erasing module that guides the image-level erasing operator to become more focused on the foreground objects.
- Extensive experiments show that our method outperforms all previous state-of-the-art methods by clear margins under various SSOD settings on the MS COCO benchmark dataset.

2 Related Works

2.1 Semi-Supervised Image Classification.

Recent semi-supervised image classification methods can be roughly divided into two categories: consistency based methods [22, 9, 23] and pseudo labeling based

methods [19, 2, 1]. The consistency based methods are predicated on the assumption that modest data disturbances should not change the predictions of images. There are several ways to implement perturbations. UDA [23] proposes image augmentations on unlabeled images to boost the performance of model. Temporal Ensembling [9] introduces an exponential moving average of label predictions on each training example. Mean Teacher [22] develops Temporal Ensembling [9] by averaging student model's parameters instead of predictions to obtain superior teacher models. The pseudo labeling based approaches annotate unlabeled data by generating pseudo labels with a strict threshold. Mixmatch and Remixmatch [1, 2] apply stochastic data augmentation to unlabeled images and obtain pseudo labels by averaging the corresponding predictions. Fixmatch [19] generates pseudo labels on weakly-augmented unlabeled images and then trained to predict the pseudo-label when fed a strongly-augmented version of the same image. Flexmatch [26] sets a flexible thresholds for different classes at each time step to let pass informative unlabeled data and their pseudo labels. However, due to the complexity of object detection task, these image classification methods can not be directly applied to semi-supervised target detection field.

2.2 Semi-Supervised Object Detection.

Similar to semi-supervised image classification task, consistency based and pseudo labeling based methods are widely utilized in semi-supervised object detection methods [20, 13, 7, 8, 29, 25, 24, 21]. Consistency based methods enforce models to generate consistent predictions on augmented images. CSD [7] is typical of consistency based methods, which constrains the consistency of features between original images and horizontal flip images. ISD [8] further proposes a mixup data augmentation method specially designed for semi-supervised object detection to create data perturbations.

For its excellent performance, the mainstream method in semi-supervised object detection is pseudo labeling based method and our method also belongs to this category. STAC [20] firstly proposes a teacher-student framework in semi-supervised object detection task which uses weak augmented images for teacher model to generate pseudo labels and trains student model to match the respective pseudo labels. Many other works [24, 13, 21, 25] further improve the performance based on STAC [20]. In Unbiased Teacher [13], focal loss is introduced to solve the class imbalance issue. Soft Teacher [24] assesses the uncertainty of bounding boxes by box jittering and selects certain bounding boxes for regression. However, all these pseudo labeling based methods inevitably suffer from minor updates in the late training process since the teacher model is deeply related to the student model due to EMA mechanism. In contrast, we design a diverse learner framework with a diverse supervision strategy to keep discrepancy between teacher and student model which is beneficial for training process. Recently, data augmentations have proven to be an effective strategy for boosting model performance in semi-supervised object detection [4, 28]. Some works [13, 21, 24, 20] apply random erasing in strong augmentation, Instant Teaching [29]

further combines Mixup and Mosaic augmentations to increase data perturbations. However, these augmentations operate on image level and neglect the information of pseudo labels. By introducing pseudo box location information, we propose a pseudo label-aware erasing module that encourages the random erasing operator to concentrate on the foreground objects.

3 Methodology

Preliminary. We follow the conventional setting of the semi-supervised object detection task, where the training set consists of two types of images: labeled images $D_s = \{s_i, y_i\}_{i=1}^{N_s}$ and unlabeled images $D_u = \{u_i\}_{i=1}^{N_u}$, where N_s and N_u are the number of labeled images s and unlabeled images u respectively. y represents the annotations for s.

Previous semi-supervised object detection works [24, 13] mostly use Faster RCNN [17] and apply the pseudo labeling method in their framework, we also follow this setting. We denote the pseudo label of the j-th bounding box in image i as p_i^j. Specifically, p_i^j consists of bounding box locations $b_i^j \in \mathbb{R}^4$ and confidence $c_i^j \in \mathbb{R}$ which is the highest classification score in all categories. There are two types of training loss: unsupervised loss L_u and supervised loss L_s. Both L_s and L_u consist of the classification loss L_{cls} and regression loss L_{reg}. For more details, please refer to Faster RCNN [17]. The overall loss for semi-supervised object detection is defined as:

$$L = L_s + \lambda L_u, \tag{2}$$

$$L_s = \frac{1}{N_s}\left(\sum_{i=1}^{N_s}(L_{cls}(s_i, y_i) + (L_{reg}(s_i, y_i))\right), \tag{3}$$

$$L_u = \frac{1}{N_u}\left(\sum_{i=1}^{N_u}(L_{cls}(u_i, p_i) + (L_{reg}(u_i, p_i))\right), \tag{4}$$

where λ indicates unsupervised loss weight.

3.1 Diverse Learner

Existing most semi-supervised object detection methods adopt teacher-student framework. However, in the traditional framework teacher and student models tend to lose their distinctiveness in the late of training stage and this will cause a less informative teacher problem. We further illustrate this phenomenon in Figure 1 where the similarity metric is calculated by

$$Similarity = \frac{1}{N_c}\sum_{i=1}^{N_c}\frac{M_{tea}^c \bigcap M_{stu}^c}{M_{tea}^c \bigcup M_{stu}^c}, \tag{5}$$

where N_c is the number of classes. M_{tea}^c is obtained by aggregating all detection results of class c produced by the teacher. The aggregation process is performed

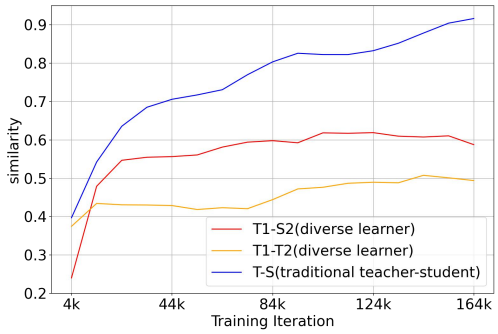


Fig. 1. Evaluation of models trained with 1% labeled images on *COCO-val* dataset. Similarity between predictions of two models. T1, T2, S2 denote teacher1, teacher2, student2 models respectively in our diverse learner framework. T and S represent teacher and student models in a traditional one paired teacher-student model correspondingly.

by setting the foreground area to 1 and background area to 0, thus generating a binary mask for each class. M^c_{stu} is generated by similar process. This similarity metric measures the prediction consistency between the teacher and the student. Obviously, as the training iteration increases, the predictions of the teacher and the student tend to be more similar, which manifests that the teacher becomes less informative and thus limits further performance improvement (the blue line in Figure 1).

We observe in Figure 1 that for two pairs of randomly initialized teacher-student models, the teachers exhibit a small similarity score at the beginning of the training (the orange line in Figure 1). This leads us to ponder: *can this discrepancy be maintained in the two paired teacher-student models where the teacher in one pair supervises the student in the other pair?* In this paper, we argue that the diverse supervision strategy does create distinctive teachers as the training proceeds (see the red and orange lines in Figure 1). The underlying working principle is three-fold: (1) our diverse learner framework creates two different learners of the same unlabeled input image through differently initialized teachers; (2) the evolution of one pair of teacher-student models receive diverse supervisory signal (pseudo labels) from the other pair, which alleviates the less informative teacher problem mentioned above; (3) the teacher in one pair is regularized by its corresponding student to maintain its distinctiveness, which prevents itself from overfitting to the supervision signal from the other pair.

Our proposed diverse learner adopts diverse supervision strategy, as shown in Figure 2(a), students are supervised by the counterpart teacher instead of the paired teacher. Specifically, we apply two randomly chosen weak augmentations W_1, W_2 to the unlabeled input images U and feed them to each teacher to obtain pseudo labels P_{tea1}, P_{tea2} correspondingly. P_{tea1}, P_{tea2} act as diverse supervising signal for the counterpart student student2, student1. Formally:

$$P_{tea1} = f_t\left(W_1(U), \theta_{tea1}\right), \tag{6}$$

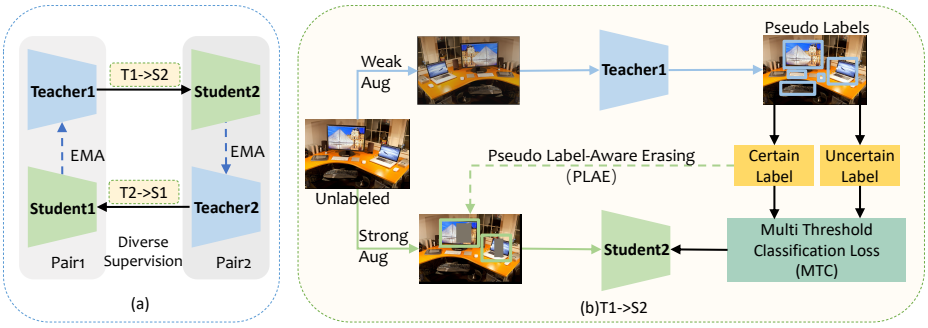


Fig. 2. Overview of our proposed **Diverse Learner** (DL) framework. (a) **Framework Abstraction**. It contains two learners, each consists of a pair of teacher-student models, where the teacher is updated by the student using EMA. In order to maintain the distinctiveness of the teacher against the student in each learner, each student receives diverse supervision from the teacher of the counterpart learner in each training iteration. All these models are randomly initialized. (b) **Detailed Framework and Training Procedure**. For simplicity, here we only illustrate supervision process of teacher1 and student2 on the unlabeled data in each training iteration, and the process of teacher2 and student1 is similar. Specifically, weak-augmented unlabeled images are fed to teacher1 to generate pseudo labels. Then we divide pseudo labels into certain labels and uncertain labels. On one hand, we devise a Pseudo Label-Aware Erasing (PLAE) module to enhance the strong augmentation for the training samples of the student2 model. On the other hand, we propose to calculate a Multi-Threshold Classification Loss (MTC) which treats certain labels and uncertain labels differently for supervising student2.

$$P_{tea2} = f_t\left(W_2(U), \theta_{tea2}\right), \tag{7}$$

where $f_t(*)$ represents the inference process of teacher, θ_{tea1}, θ_{tea2} are the network parameters of teacher1 and teacher2, respectively. In order to further introduce diversity between two learners, both θ_{tea1} and θ_{tea2} are randomly initialized.

Similarly, students' inputs are generated by two randomly chosen strong augmentation operations S_1, S_2. We define the unsupervised loss as:

$$L_u = L_u^{stu1} + L_u^{stu2}, \tag{8}$$

$$L_u^{stu1} = L_u(f_s(S_1(U), \theta_{stu1}), P_{tea2}), \tag{9}$$

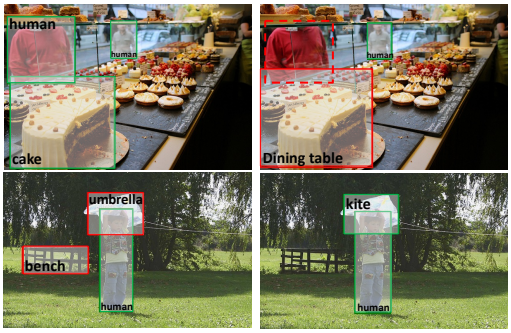
$$L_u^{stu2} = L_u(f_s(S_2(U), \theta_{stu2}), P_{tea1}), \tag{10}$$

where $f_s(*)$ represents the prediction process of student, θ_{stu1}, θ_{stu2} are the network parameters of the two students. In each training iteration, the teacher is updated by the student in the same pair learner via EMA,

$$\theta_{tea1} \leftarrow \alpha\theta_{tea1} + (1-\alpha)\theta_{stu1}, \tag{11}$$

$$\theta_{tea2} \leftarrow \alpha\theta_{tea2} + (1-\alpha)\theta_{stu2}, \tag{12}$$

Besides, when only one single pair of teacher-student is applied, once the teacher produces a wrong pseudo label for the unlabeled data, there is little



(a) (b)

Fig. 3. Examples of the pseudo labels generated by diverse learner. Red boxes denote false predictions while green boxes stand for correct predictions. (a) Predictions of teacher1; (b) Predictions of teacher2.

chance that the noise can eliminated in the subsequent training process. However, our proposed diverse learner framework makes this rectification possible. As shown in Figure 3, although some wrong pseudo labels such as the red boxes exist in one teacher model, another teacher still keeps the opportunity for generating correct labels.

3.2 Multi-Threshold Classification Loss

A common way [20, 13, 24] to ensure the precision of pseudo labels is setting a high threshold to filter unreliable labels, while this strategy brings another problem that only few pseudo labels can remain after the filtering so that many correct labels are treated as background by mistake. Figure 4 shows that in standard teacher-student framework, high precision of pseudo labels comes with the cost of low recall. On the contrary, when confidence threshold decreases, although recall of pseudo labels increases, precision declines significantly. Thus, it is not feasible to simply set a threshold to determine whether the pseudo label is foreground or background.

Based on such observation, we propose a Multi-Threshold Classification Loss, which deals with the pseudo labels differently according to the classification score. Specifically, we denote a lower bound threshold as δ_l and an upper bound threshold as δ_u. Then, we divide the pseudo labels p^j into two categories: uncertain pseudo labels $p^j_{uncertain}$ and certain pseudo labels $p^j_{certain}$:

$$p^j = \begin{cases} p^j_{uncertain} & \delta_l \leq c^j \leq \delta_u \\ p^j_{certain} & otherwise \end{cases} \tag{13}$$

For the pseudo labels whose c^j are higher than δ_u or lower than δ_l, we believe that these $p^j_{certain}$ are reliable thus we follow the standard object detection task's setting, which adopts one-hot label y^j_{hard} and cross entropy loss.

For those $p^j_{uncertain}$ whose c^j are in the uncertain interval $[\delta_l,\delta_u]$, we use the classification score of $p^j_{uncertain}$ as soft label y^j_{soft} instead of one hot label y^j_{hard}

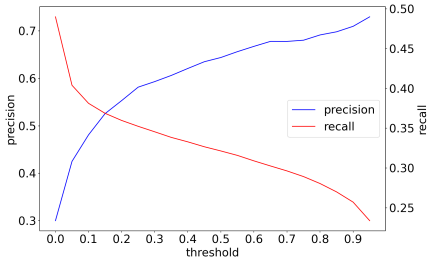


Fig. 4. precision and recall under different threshold in traditional teacher-student framework

since soft labels retain the estimations over all categories thus tolerate noisy predictions well in the uncertain interval. Next, to deal with these uncertain pseudo labels, we mainly propose following three methods:

1. **Neglect loss**: neglect all $P_{uncertain}$, i.e. loss function is not calculated.
2. **Binary cross entropy loss** : neglect the specific class of $P_{uncertain}$, only compute the binary cross entropy loss of foreground/background.
3. **KL divergence loss**: compute the KL divergence loss of $P_{uncertain}$.

We find that the KL divergence loss achieves the best performance (see Table 5), thus the final Multi-Threshold Classification Loss for unlabeled data is defined as follows:

$$L^j_{MTC} = \begin{cases} KL(y^j_{soft}||p^j_{uncertain}), & \delta_l \leq c^j \leq \delta_u \\ CE(y^j_{hard}, p^j_{certain}), & otherwise \end{cases} \tag{14}$$

where KL and CE stand for KL divergence loss and cross entropy loss, respectively. By replacing L_{cls} with L_{MTC} , the final unsupervised loss L_U is:

$$L_u = \frac{1}{N_u}(\sum_{i=1}^{N_u}(L_{MTC}(u_i, p_i) + (L_{reg}(u_i, p_i))) \tag{15}$$

3.3 Pseudo Label-Aware Erasing

In addition to employing Multi-Threshold Classification loss in classification branch, in order to further make full utilization of certain foreground pseudo labels, we introduce pseudo label-aware erasing which randomly erases some content inside these high confidence pseudo boxes.

In previous works [20, 24, 21], random erasing [28] on the whole image are widely utilized as a way of strong augmentation. We observe two major drawbacks when the random erasing is performed on the whole image. Firstly, the erased areas are likely to locate in the background, as shown in the first row of figure 5(c). This is ineffective for training object detection models since the appearance of foreground objects are unchanged. Secondly, for object detection task, image level random erasing is possible to obscure objects completely, as

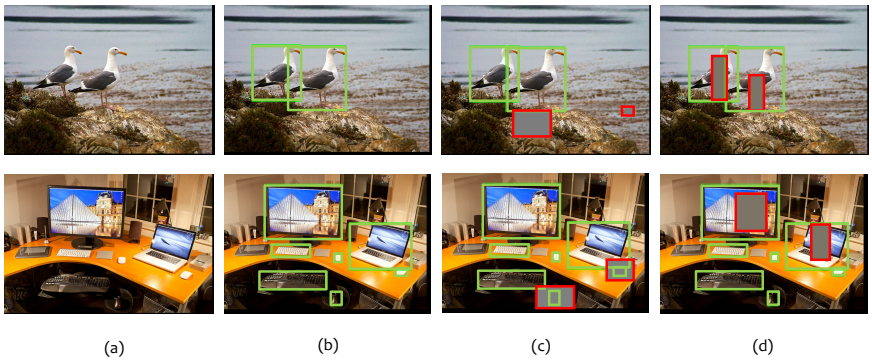


Fig. 5. (a)the original images; (b)green boxes mean pseudo labels; (c)the effect of random erasing; (d)the effect of our pseudo label-aware erasing. Red boxes indicate erasing regions.

shown in the second row in Figure 5(c). This will definitely mislead students and harms the whole training process.

Figure 5(d) shows our proposed pseudo label-aware erasing strategy on the unlabeled data. We take the pseudo label's location information into account and only random erase the objects inside the pseudo bounding box according to a certain proportion. Comparing with image level random erasing method, this erasing strategy pays more attention to foreground area and does not have the risk of dispelling the objects completely.

4 Experiments

4.1 Experiments Setting

Datasets. We validate the efficacy of our method on the MS-COCO dataset [12]. The original *COCO-standard* set contains 118K labeled images, *COCO-additional* set contains 123K unlabeled images and *COCO-val* set contains 5K images. Following the previous works [13, 24], two experimental settings are used: (1)Partially labeled data: we randomly sample 1%, 5% and 10% of the labeled training data from *COCO-standard* as a labeled training set and form the rest data into the unlabeled training data. (2)Fully Labeled data: we utilize the full labeled data in *COCO-standard* as training data set and *COCO-additional* as the unlabeled data set. We analyze the above settings on *COCO-val* set using mean average precision(mAP) as the evaluation metrics.

Implementation Detail. For fair comparison, we follow the previous works [20, 13, 24], using Faster-RCNN [17] with FPN [10] as our detection framework. We initialize the parameters of backbones in four models with the Resnet-50 [5] pre-trained on ImageNet and the parameters of detection heads randomly. In regression branch, we use box-jittering strategy mentioned in Soft Teacher [24].

For partially labeled data, we train models for 180k iterations, and set unsupervised loss weight λ to 4.0, batch size to 40, unlabeled data sampling ratio to 0.2. For fully labeled data, we train models for 720k iterations, and set unsupervised loss weight λ to 2.0, batch size to 64, unlabeled data sampling ratio to 0.5. And we set EMA update parameter α to 0.999 in both partially labeled data and fully labeled data setting.

For the multi-threshold classification loss, we set lower bound confidence threshold $\delta_l = 0.8$ and upper bound confidence threshold $\delta_u = 0.9$. We apply pseudo label-aware erasing strategy to bounding boxes with confidence score higher than 0.9 since 80000 iteration to meet the demand of accurate bounding box locations. Besides, we utilize random resize and horizontal flip as weak augmentation and strong augmentation contains random erasing, rotation, color jittering, etc.

In inference, since two teacher models both achieve high performance, we report the performance of the better teacher model.

4.2 Results

Partially labeled data. We first compare our method with previous state-of-the-art methods with 1%, 5% and 10% labeled data from MS-COCO. As shown in Table 1, our method achieves the SOTA performance under all three settings. Diverse learner outperforms the latest best method Soft Teacher [24] by 2.10%, 1.50% and 0.83% under 1%, 5% and 10% setting respectively. It is worth mentioning that diverse learner outperforms other methods especially when labeled data is extremely rare.

Table 1. Comparsion with CSD [7], STAC [20], Unbiased Teacher [13], Humble Teacher [21], Instant Teaching [29] and Soft Teacher [24] on MS-COCO dataset with partially labeled data setting.

Method	1%COCO	5%COCO	10%COCO
supervised	10.0	20.92	26.94
CSD [7]	10.51	18.63	22.46
STAC [20]	13.97	21.18	26.18
Humble Teacher [21]	16.98	27.70	31.61
Instant Teaching [29]	18.05	26.75	30.40
Unbiased Teacher [13]	20.75	28.27	31.50
Soft Teacher [24]*	21.62	30.42	33.78
Ours	**23.72**	**31.92**	**34.61**

Fully labeled data. Aside from the excellent performance on the partially labeled dataset, we also show that our method can surpass other methods

* Metrics reported on Soft Teacher's official Github repo.

Table 2. Results on fully labeled data comparison with with CSD [7], STAC [20], Unbiased Teacher [13], Humble Teacher [21], Instant Teaching [29] and Soft Teacher [24]

Method	mAP
supervised	40.89
CSD [7]	38.82
STAC [20]	39.21
Humble Teacher [21]	42.37
Instant Teaching [29]	40.20
Unbiased Teacher [13]	41.30
Soft Teacher [24]*	44.05
Ours	**44.86**

trained on fully labeled dataset. As shown in Table 2, our method exceeds Soft Teacher [24] by 0.81% and reaches 44.86%, demonstrating the effectiveness of diverse learner in case of large amount of labeled data and unlabeled data.

4.3 Ablation Study

In this section, in order to validate our key designs, we conduct extensive ablation experiments using 1% labeled MS-COCO dataset. We choose the popular STAC [20] framework as our baseline method which contains one pair of teacher-student models with EMA strategy. Additionally, we apply the box jittering techniques [24] to enhance the performance of the regression branch. Based on the baseline method, we gradually integrate our proposed key designs and ablate the effectiveness.

Table 3. Effect of all the key designs, we denote multi-threshold classification loss as MTC, pseudo-label aware erasing as PLAE, diverse learner as DL and mutual learning as ML.

No.	MTC	PLAE	DL	ML	mAP
1					20.6
2	√				22.1 (+1.5)
3	√	√			22.7 (+2.1)
4	√	√		√	19.0 (-1.6)
5	√	√	√		**23.7** (+3.1)

Effectiveness of MTC. Result NO.1 and NO.2 in Table 3 illustrate that training with MTC surpasses the baseline method over 1.5 points. For a better understanding of the effectiveness of MTC, we analyze the precision and recall values

* Metrics reported on Soft Teacher's official Github repo.

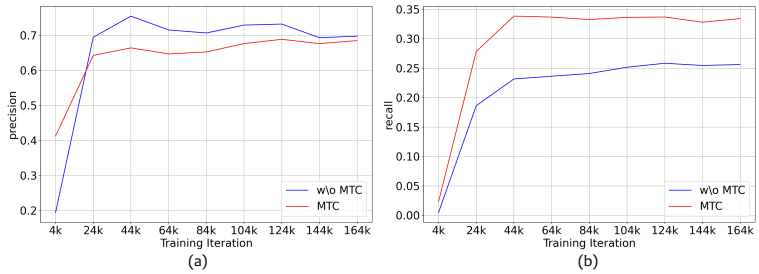


Fig. 6. Evaluation on *COCO-val* dataset under iou threshold=0.5. (a)Diverse learner with multi-threshold classification loss(MTC) achieves almost the same precision as diverse learner without MTC. (b)Diverse learner with MTC achieves much higher recall than diverse learner without MTC.

during the training progress. As shown in Figure 6, when MTC is applied, the recall increases significantly while the precision is maintained comparing with the baseline.

Effectiveness of PLAE. As demonstrated by the result NO.2 and NO.3 in Table 3, a further improvement of 0.6% mAP is achieved when PLAE strategy is applied.

Effectiveness of Diverse Learner. Here we first ablate the effect of incorporating our proposed DL framework. As shown in result NO.5 in Table 3, DL achieves another 1% improvement in mAP, reaching 23.7% mAP, which is 3.1% better than our baseline. Secondly, in order to further ablate the effects of using a teacher-student pair instead of one single model in each learner, we conduct experiment of integrating the conventional mutual learning method [27]. As shown in Figure 7, directly integrating mutual learning strategy (green line) leads to an unstable training process thus harms the performance. We observe a severe drop (1.6%) in mAP, as shown in the result NO.4 in Table 3. On the contrary, our proposed DL framework enjoys merits from the mutual learning strategy while successfully stabilizes the training using the teacher-student pair with the EMA updating mechanism in each learner.

We calculate the average layer-wise cosine similarity of random-initialized layers' parameters between the teacher and student to measure the similarity of the teacher student pair. Results show that our proposed DL successfully reduces the similarity of the teacher and student from 0.9987 (traditional one paired teacher student) to 0.1498 (teacher1 and student2 in diverse learner).

Table 4. The comparison of choosing different lower threshold in MTC. The upper threshold is fixed to 0.9.

	mAP@δ_l=0.6	mAP@δ_l=0.7	mAP@δ_l=0.8
Ours (w\o DL)	22.6	23.1	22.7
Ours	23.2	23.4	**23.7**

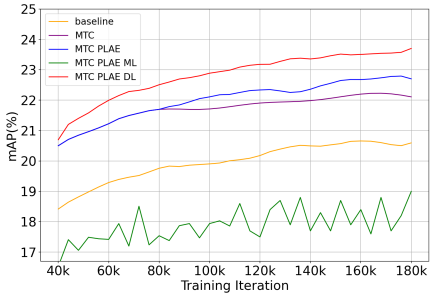


Fig. 7. The curves of mAP values corresponding to each setting in Table 3 during the training stage.

Threshold choice in MTC. We simply set upper threshold δ_u to 0.9 in our experiment and explore the choice of lower threshold δ_l from 0.6 to 0.8 (shown in Table 4) and 0.8 is the most suitable value for diverse learner.

Loss type for uncertain pseudo labels. As mentioned in Section 3.2, we experiment three types of loss functions to deal with the uncertain pseudo labels: neglect loss, binary cross entropy loss and KL divergence loss. The results of applying different loss functions on the baseline with PLAE module are shown in Table 5, KL divergence loss obtains the best performance.

Table 5. Effect of different types of loss functions for the uncertain pseudo labels.

Loss type	mAP
neglect	21.2
binary cross entropy loss	22.2
KL divergence loss	**22.7**

5 Conclusions

We present a diverse learner framework with diverse supervision that could maintain the distinctiveness of the teacher against the student. We also introduce a multi-threshold classification loss for a better utilization of both high-quality pseudo labels and potential uncertain pseudo labels and devise a simple yet efficient pseudo label-aware erasing strategy. Extensive experiments demonstrate the superiority of our method on the MS-COCO benchmark dataset. We will extend diverse learner to more learners in the future work, and study more elaborate supervision signals between multiple learners.

References

1. Berthelot, D., Carlini, N., Cubuk, E.D., Kurakin, A., Sohn, K., Zhang, H., Raffel, C.: Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. arXiv preprint arXiv:1911.09785 (2019)
2. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.: Mixmatch: A holistic approach to semi-supervised learning. arXiv preprint arXiv:1905.02249 (2019)
3. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
4. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
5. 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)
6. Iandola, F., Moskewicz, M., Karayev, S., Girshick, R., Darrell, T., Keutzer, K.: Densenet: Implementing efficient convnet descriptor pyramids. arXiv preprint arXiv:1404.1869 (2014)
7. Jeong, J., Lee, S., Kim, J., Kwak, N.: Consistency-based semi-supervised learning for object detection. Advances in neural information processing systems **32**, 10759–10768 (2019)
8. Jeong, J., Verma, V., Hyun, M., Kannala, J., Kwak, N.: Interpolation-based semi-supervised learning for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11602–11611 (2021)
9. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242 (2016)
10. 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)
11. 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)
12. 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)
13. 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)
14. 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)
15. Radosavovic, I., Dollár, P., Girshick, R., Gkioxari, G., He, K.: Data distillation: Towards omni-supervised learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4119–4128 (2018)
16. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
17. 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**, 91–99 (2015)

18. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
19. Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. arXiv preprint arXiv:2001.07685 (2020)
20. 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)
21. 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)
22. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv preprint arXiv:1703.01780 (2017)
23. Xie, Q., Dai, Z., Hovy, E., Luong, M.T., Le, Q.V.: Unsupervised data augmentation for consistency training. arXiv preprint arXiv:1904.12848 (2019)
24. 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. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
25. 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)
26. 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)
27. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4320–4328 (2018)
28. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 13001–13008 (2020)
29. 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)
30. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)
31. Zoph, B., Ghiasi, G., Lin, T.Y., Cui, Y., Liu, H., Cubuk, E.D., Le, Q.: Rethinking pre-training and self-training. Advances in neural information processing systems **33**, 3833–3845 (2020)