De-biased Teacher: Rethinking IoU Matching for Semi-supervised Object Detection

Kuo Wang[1]*, Jingyu Zhuang[1]*, Guanbin Li[1]†, Chaowei Fang[2], Lechao Cheng[3], Liang Lin[1], Fan Zhou[1]†

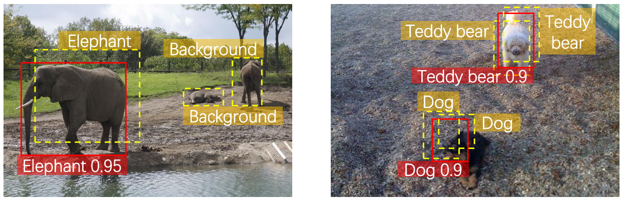
[1]School of Computer Science and Engineering, Research Institute of Sun Yat-sen University in Shenzhen, Sun Yat-sen University, Guangzhou, China
[2]School of Artificial Intelligence, Xidian University, Xi'an, China
[3]Zhejiang Lab, Zhejiang, China
{wangk229, zhuangjy6}@mail2.sysu.edu.cn, liguanbin@mail.sysu.edu.cn, chaoweifang@outlook.com,
chenglc@zhejianglab.com, linliang@ieee.org, isszf@mail.sysu.edu.cn

Abstract

Most of the recent research in semi-supervised object detection follows the pseudo-labeling paradigm evolved from the semi-supervised image classification task. However, the training paradigm of the two-stage object detector inevitably makes the pseudo-label learning process for unlabeled images full of bias. Specifically, the IoU matching scheme used for selecting and labeling candidate boxes is based on the assumption that the matching source (ground truth) is accurate enough in terms of the number of objects, object position and object category. Obviously, pseudo-labels generated for unlabeled images cannot satisfy such a strong assumption, which makes the produced training proposals extremely unreliable and thus severely spoil the follow-up training. To de-bias the training proposals generated by the pseudo-label-based IoU matching, we propose a general framework – De-biased Teacher, which abandons both the IoU matching and pseudo labeling processes by directly generating favorable training proposals for consistency regularization between the weak/strong augmented image pairs. Moreover, a distribution-based refinement scheme is designed to eliminate the scattered class predictions of significantly low values for higher efficiency. Extensive experiments demonstrate that the proposed De-biased Teacher consistently outperforms other state-of-the-art methods on the MS-COCO and PASCAL VOC benchmarks. Source codes are available at https://github.com/wkfdb/De-biased-Teracher.

Introduction

Benefiting from the availability of large-scale annotated datasets, supervised learning has achieved astounding performance on various computer vision tasks. However, large amounts of annotations are expensive and time-consuming to collect, particularly for object detection. To alleviate this issue, Semi-Supervised Learning (SSL) which is designed to fully exploit the potential of unlabeled data to facilitate model learning has received much attention. Yet, the majority of the advanced SSL methods come from classification tasks (Sohn et al. 2020a), and there are still many unsolved problems in generalizing them to object detection.



(a) Low recall rate of pseudo labels leads to many objects in candidates be incorrectly labeled as background by IoU matching.

(b) The noisy information (incorrect position or category) in one single pseudo label will be propagated to all the matched candidates by IoU matching.

Figure 1: Possible detrimental analysis of IoU matching with the pseudo label. The red solid boxes are the pseudo-labels adopted by the model during training, and the yellow dashed boxes are the training proposals generated during IoU matching.

Recent research on semi-supervised object detection (SSOD) basically follows the paradigm of pseudo-labeling, that is, by generating pseudo-labels for weakly-augmented unlabeled images to supervise the training of their strongly augmented version. However, a simple generalization of this scheme in the field of object detection fails to achieve stunning performance comparable to classification tasks (Sohn et al. 2020a). In this work, we deeply dissect the training design of the two-stage object detectors and discover IoU matching, the culprit that inevitably causes biased training proposals based on pseudo labels of unlabeled images.

According to the default training design of Faster-RCNN, a number of candidate boxes from RPN are sent to the ROI head while the model does not use all of them for training. Instead, the process of IoU matching is applied to select training proposals from those candidates according to their overlapping with the ground truth. However, by treating pseudo labels as ground truth for unlabeled images, the IoU matching process will inevitably assign noisy category

*These authors contributed equally.
†Corresponding authors.

labels to the candidates and ultimately lead to extremely biased training proposals, which severely hinders the subsequent training on the classification branch of object detection.

The bias caused by IoU matching on unlabeled images is multifaceted. First of all, IoU matching produces biased background proposals because of the low recall rate of pseudo labels, as shown in Figure 1(a). Second, IoU matching aggravates the noisy information contained in pseudo labels and eventually leads to biased foreground proposals, as shown in Figure 1(b). Third, based on the wrongly labeled candidates, the selection for training proposals is further biased. One of the most obvious problems is that the training proposal tends to only focus on relatively easy objects while ignoring the other hard positive samples (marked as background) in unlabeled images. After all, **the IoU matching scheme is designed based on the assumption that the ground truth is completely accurate, which is inappropriate for the classification learning branch from the pseudo-labeled "ground truth"**.

To alleviate the bias in the collected training proposals, existing SSOD methods employ various strategies to improve the quality of pseudo-labels. However, all of the existing solutions are not aware of the inappropriateness of the IoU matching on unlabeled images, resulting in very limited performance gains. To fundamentally eliminate the bias caused by IoU matching, we propose a novel framework: *De-biased Teacher*, which **deserts the IoU matching and pseudo labeling processes** on the classification learning branch for unlabeled images by **directly generating favorable training boxes for consistency regularization** between the weak and strong image augmentation pairs. Moreover, as the softmax function inevitably produces long-tailed distributions which are hard to fit, we further design a distribution refinement scheme to cut off the tails in the target distribution for higher efficiency.

In summary, our main contributions are as follows:

- Through a thorough analysis of the prevalent two-stage object detectors, we conclude that IoU matching process is inappropriate within the semi-supervised setting, which is the essential reason for the biased training proposals on unlabeled images.
- We propose a simple yet effective semi-supervised object detection framework, *De-biased Teacher*, which eliminates the IoU matching bias by directly generating training proposals for consistency regularization, combined with a distribution refinement mechanism.
- Extensive experiments demonstrate that *De-biased Teacher* consistently outperforms other state-of-the-art methods on both MS-COCO and PASCAL VOC benchmarks. Moreover, we additionally evaluate the *De-biased Teacher* on the open scene setting, and the results verified the robustness and effectiveness of our method.

Related Work

Semi-Supervised Learning

Semi-supervised learning (SSL) is targeted at exploiting the potential of unlabeled data during the model learning procedure. Recently great progress has been made in SSL for image classification. Among them, two main principles are followed, namely pseudo labeling and consistency regularization. Pseudo labeling (also named self-training) methods (Xie et al. 2020b; Iscen et al. 2019) aim to improve the performance of SSL by generating high-quality pseudo labels for unlabeled data. Consistency regularization (Bachman, Alsharif, and Precup 2014; Miyato et al. 2018; Tarvainen and Valpola 2017) methods incentivize the model to produce consistent predictions on different perturbations of the same image. The ways to implement the disturbance span perturbing the model (Bachman, Alsharif, and Precup 2014), augmenting the images (Sajjadi, Javanmardi, and Tasdizen 2016), and adversarial training (Miyato et al. 2018). Recently, data augmentations have proven effective for boosting SSL on image classification, such as Mixmatch (Berthelot et al. 2019) and UDA (Xie et al. 2020a), which prompted the model to generate consistent predictions on multiple views, and Fixmatch (Sohn et al. 2020a), which trains the model by using one-hot high-confidence pseudo-labels generated from weakly augmented images to supervise strongly augmented ones. Our De-biased Teacher also adopts different data augmentations to achieve consistent regularization while focusing on instance-level consistency, instead of the whole image version.

Semi-Supervised Object Detection

As annotations for object detection are more expensive to obtain, semi-supervised object detection (SSOD) has gained increasing attention. Existing SSOD methods are also mainly based on pseudo-labeling methods (Li et al. 2020; Wang et al. 2018). Recently, STAC (Sohn et al. 2020b) follows Fixmatch (Sohn et al. 2020a) to generate pseudo labels at the image level using weakly augmented images and then train the model on strongly augmented ones. This pseudo-labeling method then becomes the mainstream paradigm for handling SSOD tasks and STAC has become a classic benchmark. After STAC, many methods (Xu et al. 2021; Liu et al. 2021; Li, Yuan, and Li 2022; Zheng et al. 2022; Mi et al. 2022) are proposed to improve the quality of the pseudo labels. Among them, Unbiased Teacher (Liu et al. 2021) utilizes focal loss to alleviate the data imbalance problem, Soft Teacher (Xu et al. 2021) generates pseudo-labels with more accurate positions to improve the box regression performance, Active Teacher (Mi et al. 2022) filters unlabeled data and picks appropriate images to generate pseudo-labels, Li et al (Li, Yuan, and Li 2022) generates better pseudo labels from the perspective of multi-view by jointly using prediction results and prototypes of each category. However, all these existing SSOD methods suffer from the detrimental effects of the unreliable proposals filtered by IoU matching. In our devised *De-biased Teacher*, the biased caused by IoU matching is fundamentally eliminated by directly performing instance-level consistency-regularization between the weak-strong image pairs.

Approach

Our goal is to address the semi-supervised object detection task. Existing SSOD methods rely on IoU matching when

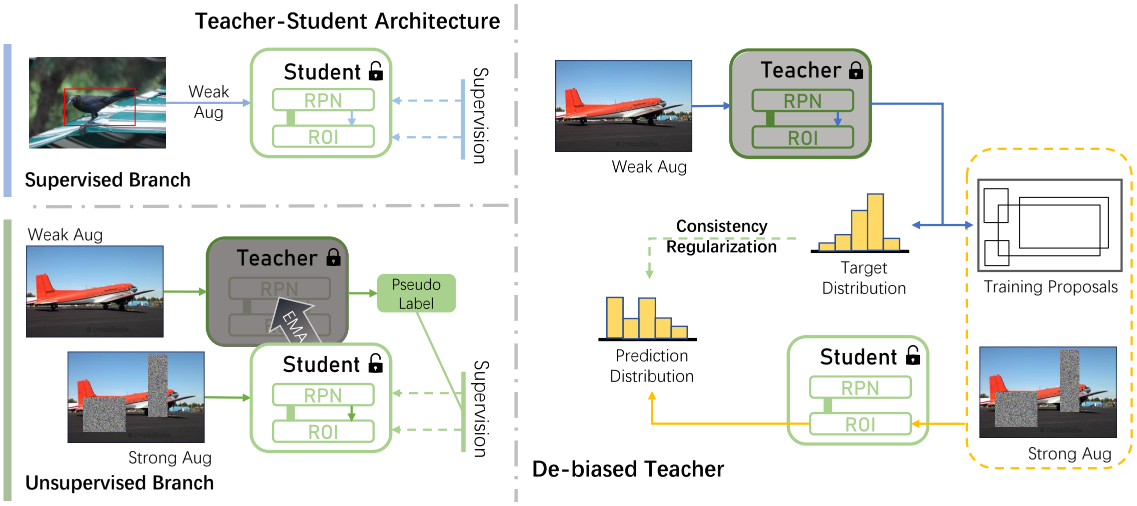


Figure 2: The framework of De-biased Teacher, which follows the paradigm of Teacher-Student architecture. The pseudo labels are only used in RPN loss and ROI regression loss. De-biased Teacher abandons the conventional IoU matching mechanism to eliminate the bias in ROI classification. In detail, the Teacher model will directly generate favorable training proposals with prediction distribution on the weakly augmented images, and then send them to the Student for consistency regularization on the corresponding strongly augmented version.

generating training proposals from the candidates produced by RPN. This greatly limits the model's exploration of unlabeled data, which may not only miss many high-quality positive samples but also introduce biased labels. We propose a novel SSOD framework named De-biased Teacher, which directly infers suitable training proposals with soft labels for consistency regularization from the candidate boxes extracted on unlabeled images, avoiding the detrimental bias caused by the conventional IoU matching. Details of the proposed method are introduced in the following sections.

Problem Description

Semi-supervised object detection aims to train detection models by leveraging a large unlabeled dataset $\mathcal{D}_u = \{(\mathbf{x}_i^u)\}_{i=1}^{n_u}$ alongside a small labeled dataset $\mathcal{D}_l = \{(\mathbf{x}_i^l, y_i^l)\}_{i=1}^{n_l}$, where n_u and n_l are the number of unlabeled and labeled data. For each labeled image $\mathbf{x}_i^l$, the annotation y_i^l contains the class labels and bounding box coordinates of all objects in the image. The crucial problem remains in the exploration of unlabeled data.

Overall Framework

Following the paradigm of existing SSOD methods (Xu et al. 2021; Liu et al. 2021), *De-biased Teacher* is designed based on the Teacher-Student learning framework (Tarvainen and Valpola 2017) and adopts Faster-RCNN as the detector. In detail, the framework is composed of two independent models, including a *Teacher* model and a *Student* model. In each iteration, a batch of labeled and unlabeled images are randomly selected from the dataset $\mathcal{D}_l$ and $\mathcal{D}_u$ respectively. Among them, the labeled images are directly used to train the *Student* model in a supervised manner with weak augmentations applied. For unlabeled images, the *Teacher* model first generates training targets with the weakly augmented images, which are used for training the *Student* model on the strongly augmented images. Denote the parameters of the student model as θ_s. θ_s is updated by optimizing the training objective function. The *Teacher* model's parameters (denoted by θ_t) are updated via the exponential moving average (EMA) (Tarvainen and Valpola 2017):

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

where α is a constant indicating the ensemble ratio between the *Teacher* model and the *Student* model.

For training the detection model on labeled images, we adopt the conventional supervised loss function as follows,

$$\begin{aligned}\mathcal{L}_{sup} = \sum_i & \mathcal{L}_{cls}^{rpn}\left(\boldsymbol{x}_i^l, y_i^l\right) + \mathcal{L}_{reg}^{rpn}\left(\boldsymbol{x}_i^l, y_i^l\right) + \\ & \mathcal{L}_{cls}^{roi}\left(\boldsymbol{x}_i^l, y_i^l\right) + \mathcal{L}_{reg}^{roi}\left(\boldsymbol{x}_i^l, y_i^l\right).\end{aligned} \tag{2}$$

Here, $\mathcal{L}_{cls}^{rpn}$, $\mathcal{L}_{reg}^{rpn}$, $\mathcal{L}_{cls}^{roi}$, and $\mathcal{L}_{reg}^{roi}$ denotes the RPN classification loss , the RPN regression loss , the ROI classification loss, and the ROI regression loss, respectively.

For each unlabeled image, the weakly augmented image $\boldsymbol{x}_i^w$ is first fed into the Teacher branch, resulting in a set of training targets. The targets here are divided into three aspects: $\widehat{y}_i^{rpn}$, $\widehat{y}_i^{cls}$, $\widehat{y}_i^{reg}$, which are used to calculate RPN loss, ROI classification loss and ROI regression loss respectively. The unsupervised loss $\mathcal{L}_{unsup}$ is calculated as follows:

$$\begin{aligned}\mathcal{L}_{unsup} = \sum_i & \mathcal{L}_{cls}^{rpn}\left(\boldsymbol{x}_i^s, \widehat{y}_i^{rpn}\right) + \mathcal{L}_{reg}^{rpn}\left(\boldsymbol{x}_i^s, \widehat{y}_i^{rpn}\right) + \\ & \mathcal{L}_{cls}^{roi}\left(\boldsymbol{x}_i^s, \widehat{y}_i^{cls}\right) + \mathcal{L}_{reg}^{roi}\left(\boldsymbol{x}_i^s, \widehat{y}_i^{reg}\right).\end{aligned} \tag{3}$$

Note that $\boldsymbol{x}_i^s$ is the strongly augmented image. With the supervised loss $\mathcal{L}_{sup}$ and the unsupervised loss $\mathcal{L}_{unsup}$, the overall objective function is defined as the weighted sum:

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda\mathcal{L}_{unsup}, \tag{4}$$

where λ is a constant. The SGD optimizer is adopted to update *Student*'s parameter θ_s. After each optimization step, the *Teacher*'s parameter θ_t is updated via Eq. 1.

The training targets for RPN loss and ROI regression loss are generated by conventional pseudo labeling scheme. The primary novelty of our De-biased Teacher remains in the ROI classification loss on unlabeled images, where we desert the convention IoU matching and pseudo labeling scheme and adopt consistency regularization to learn from unlabeled data. Note that the detrimental effect of bias caused by IoU matching is mainly concentrated on ROI classification, with minimal impact on the function of RPN.

Consistency Regularization

As described in the introduction section, the bias caused by IoU matching on unlabeled images is reflected in three aspects: incorrectly labeling positive samples as background, wide propagation of noise in pseudo labels, and misleading selection of training proposals. To eliminate all those biases, we replace the ont-hot pseudo label with soft training proposals to directly train the Student without IoU matching. The detailed operation is described below.

The Teacher model will first infer the weak image for each unlabeled weak-strong augmented image pair to generate training targets for ROI classification loss. Based on the inference results of all the candidates, our De-biased Teacher directly filters favorable proposals to supervise the training of the Student model on the corresponding strongly augmented image. In detail, candidates with foreground score (maximum prediction score on the foreground categories) higher than a specific small threshold δ are considered as foreground boxes while the others are regarded as background. For each image pair, a fixed number (default as 512) of training proposals are selected, which are composed of foreground boxes and some randomly selected backgrounds. This selection process enables the model to make full use of unlabeled images, instead of only benefitting from the easy-to-detect objects.

The classification learning branch on unlabeled images of our De-biased Teacher is defined as the consistency regularization between the target distribution from the Teacher on weak image and the prediction distribution from the Student on strong image. By directly sending the training proposals to the Student model, the prediction distribution of these boxes on the strongly augmented image will be obtained. Combined with the target distribution, the ROI classification loss is calculated via the typical soft cross-entropy loss:

$$\hat{\mathcal{L}}^{roi}_{cls} = \sum_{i=1}^{N}\sum_{c=1}^{C} -p_{i,c}\log q_{i,c} \tag{5}$$

In Eq. 5, N is the fixed number of training proposals selected for each image, C is the total category (including background) number. p_i is the target distribution of the $i-st$ proposal generated by the Teacher on weak images and q_i is the corresponding prediction distribution produced by the Student on strong images.

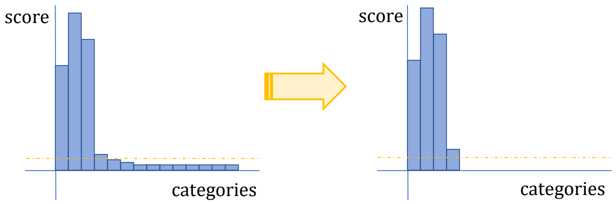


Figure 3: The distribution refinement mechanism adopts a pre-defined threshold to cut off the tails and then renormalize the distribution. The refined soft labels get rid of the irrelevant categories, which can speed up the convergence.

Consistency regularization of the same training proposals between strongly and weakly enhanced images fundamentally removes the bias caused by IoU matching. The training boxes can be assigned with tailored soft distributions and the noisy propagation problem of IoU matching no longer exists. Furthermore, the selection of training proposals is also less biased, where more positive objects will be minded from the unlabeled images to optimize the learning.

Distribution Refinement

Directly generating training proposals for consistency regularization eliminates the bias caused by IoU matching on unlabeled images, which bridges the gap between semi-supervised image classification and semi-supervised object detection. However, the target distribution inferred by the Teacher model for each training box may contain scattered class predictions of significantly low values, which is not conducive to classification learning. To make better use of the inference results, the distribution refinement mechanism is proposed.

The extremely low scores in target distributions are the side effect of softmax. To remove the side effects, the threshold δ (threshold for distinguishing foreground boxes from the candidates) is again used to cut off the irrelevant categories in the target distribution. After setting the scores lower than δ to zero, re-normalization is performed to generate the refined target distribution for each training box, as shown in Figure 3.

By applying distribution refinement, the easy-to-detect foreground objects and the background in the training boxes would get one-hot labels as target distribution, and the other objects with low confidence would get compact soft labels, which indicate both the correlation and non-correlation between the training box and the categories. After distribution refinement, the ROI classification loss in Eq.5 changes to the following loss:

$$\hat{\mathcal{L}}^{roi}_{cls} = \sum_{i=1}^{N}\sum_{c=1}^{C'} -p'_{i,c}\log q_{i,c} \tag{6}$$

where C' is the categories with a strong correlation with the feature and p' is the re-normalized scores, which means the

Methods	1% labeled COCO	5% labeled COCO	10% labeled COCO
Supervised	9.05±0.16	18.47±0.22	23.86±0.81
STAC (Sohn et al. 2020b)	13.97±0.35	24.38±0.12	28.64±0.21
Instant-Teaching (Zhou et al. 2021)	18.05±0.15	26.75±0.05	30.40±0.05
Humble-Teacher (Tang et al. 2021)	16.96±0.38	27.70±0.15	31.61±0.28
Unbiased-Teacher (Liu et al. 2021)	20.75±0.12	28.27±0.11	31.50±0.10
Soft-Teacher (Xu et al. 2021)	20.46±0.39	30.74±0.08	34.04±0.14
Active-Teacher (Mi et al. 2022)	22.20	30.07	32.58
Scale Equivalent (Guo et al. 2022)	-	29.01	34.02
DDT (Zheng et al. 2022)	18.62±0.42	29.24±0.16	32.80±0.22
MA-GCP (Li, Yuan, and Li 2022)	21.30±0.28	31.67±0.16	35.02±0.26
De-biased Teacher	**22.50**±0.23	**32.10**±0.15	**35.50**±0.20

Table 1: Comparison with existing SSOD methods using different percentages of labeled data.

Methods	mAP
Supervised	37.63
STAC (Sohn et al. 2020b)	39.20
Instant-Teaching (Zhou et al. 2021)	40.20
Humble-Teacher (Tang et al. 2021)	42.37
Unbiased-Teacher (Liu et al. 2021)	41.30
Scale Equivalent (Guo et al. 2022)	41.50
DDT (Zheng et al. 2022)	41.90
Soft-Teacher (Xu et al. 2021)	44.50
De-biased Teacher	**44.70**

Table 2: Comparison with existing SSOD methods on fully labeled COCO.

degree of correlation between the feature and categories.

Experiments

Datasets We evaluate our proposed approach on three object detection datasets, PASCAL VOC (Everingham et al. 2010), MS-COCO (Lin et al. 2014) and Object365 (Shao et al. 2019). Four benchmarks are established: 1) VOC: We use `VOC2007-trainval` as the labeled dataset, `VOC2012-trainval` as the unlabeled dataset and `VOC2007-test` as the evaluation set. 2) Partially labeled COCO: We randomly sample 1%/5%/10% images from `train2017` as the labeled dataset and use the remaining images as the unlabeled dataset. 3) Fully labeled COCO: The whole `train2017` is used as the labeled dataset, and the whole `unlabeled2017` is used as the unlabeled dataset. `COCO-val2017` is used as the evaluation set for both 2) and 3). 4) Open scene: We use the `COCO-train2017` as the labeled data and the `Object365` as the unlabeled data. `COCO-val2017` and `Object365-val` are both used to evaluate the approach.

Implementation Details Our De-biased Teacher and all the experimental settings are implemented based on MMDetection (Chen et al. 2019). For fair comparisons, we utilize Faster-RCNN as our detector and use Resnet-50-FPN as the backbone. Following existing works (Xu et al. 2021), we set the EMA updating rate $\alpha = 0.999$. For the coefficient of unsupervised loss, we set $\lambda = 4$ on partially labeled COCO and $\lambda = 2$ on fully labeled COCO, VOC and open scene. The similar weak-strong data augmentation schemes in (Liu et al. 2021) are utilized. For RPN and regression loss, the pseudo label is generated by the conventional thresholding method with $\sigma_{RPN} = 0.7$ and $\sigma_{reg} = 0.9$. For ROI classification loss, we set threshold $\delta = 0.05$ for selecting foregrounds and distribution refinement. The batch size of labeled and unlabeled data is (12,24) for VOC, (8,32) for partially labeled COCO, (32,32) for fully labeled COCO and open scene. The training iteration is 90k for VOC, 180k for partially labeled COCO and 720k for fully labeled COCO and open scene.

Results on MS-COCO

We compare our method with other existing SSOD methods on MS-COCO dataset, where partially or fully labeled data are leveraged as the labeled dataset.

On partially labeled COCO, the size of the labeled data is limited, which means that the pseudo-labels generated by the model are extremely unstable, resulting in more biased training proposals due to IoU matching. Existing works have designed various complex approaches to improve the quality of pseudo-labels, however, this cannot overcome the fatal flaw of IoU matching when dealing with unlabeled samples. On the contrary, our De-biased Teacher replaced the pseudo-label-based IoU matching by directly selecting favorable training boxes for consistency regularization, which fundamentally eliminates the bias in the selection and labeling of training boxes. It's simple while effective, as shown in Table 1, our method consistently outperforms existing SSOD algorithms.

On fully labeled MS-COCO, the large amount of labeled data enhances the capabilities of the model, making pseudo-labels particularly reliable. Under such conditions, our method still achieves optimal performance, as shown in Table 2. It is worth noting that the other methods may additionally improved the regression learning (box jittering in Soft Teacher (Xu et al. 2021)) while our method did nothing on the regression branch. The results in Table 1 and Table 2 indicate that our proposed method can stably replace the

Methods	AP50
Supervised	76.30
STAC (Sohn et al. 2020b)	77.45
Multi-Phase (Wang et al. 2021)	78.60
Instant-Teaching (Zhou et al. 2021)	79.20
Humble-Teacher (Tang et al. 2021)	80.94
Unbiased-Teacher (Liu et al. 2021)	77.37
Scale Equivalent (Guo et al. 2022)	80.60
De-biased Teacher	**81.50**

Table 3: Comparison with existing SSOD methods on PASCAL-VOC.

Method	mAP on Object365 validation	mAP on COCO val2017
Supervised	23.1	37.5
STAC*	23.8(+0.7)	38.7
Unbiased Teacher*	24.2(+1.1)	40.7
Soft Teacher	25.0(+1.9)	42.8
De-biased Teacher	**28.1**(+**5.0**)	**43.1**

Table 4: Results on open scene SSOD task. The labeled data is COCO-train2017 and the unlabeled data is Object365. We record the mAP of 80 COCO categories on COCO-val2017 set and Object365 validation set. *: Our implementation on MMDetection.

conventional IoU matching and improve the performance of the model under various conditions.

Results on PASCAL VOC

The comparison results between our De-biased Teacher with other existing SSOD methods on VOC dataset are shown in Table 3. The proportion of data in this experimental protocol is around 1:2, and the number of categories contained in this dataset is relatively small compared with MS-COCO. On such a small-scale dataset, it is less difficult to generate more reliable pseudo-labels, which means that the IoU matching causes less bias. Even though, our method still outperforms other methods. This indicates that even on simple small-scale datasets, the problems caused by IoU matching on unlabeled images are still non-negligible.

Results on Open Scene SSOD

To verify the robustness of our method, we additionally conduct SSOD experiments in open scenarios. The Object365 dataset contains plenty of open set samples, benefiting from open unlabeled data is more realistic and difficult. As shown in Table 4, on open scene SSOD tasks, our De-biased Teacher still consistently outperforms other mainstream SSOD methods, especially on open validation set, where huge improvement is achieved. The reason is that conventional SSOD methods only extract high-confidence objects from unlabeled data for optimization while our method makes full use of foreground objects in unlabeled data to boost the detection ability. This allows the model to detect more close-set objects on the noisy open validation set. At the same time, although our method introduces many open objects during training, the results on COCO-val2017 demonstrate that consistency regularization on open-set samples does not hurt the model's ability to detect known objects. Reuslts from Table 1– 4 demonstrate that our method is both robust and effective.

	label-type	Precision	Recall
pseudo label	one-hot	0.892	0.441
foreground boxes (IoU Matching)	one-hot	0.775	0.467
foreground boxes (De-biased Teacher)	soft	-	0.782

Table 5: Analysis of foreground boxes generated by conventional IoU matching and De-biased Teacher. All the other boxes are one-hot labeled as background.

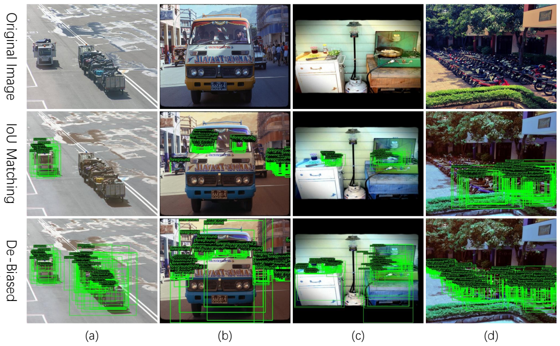


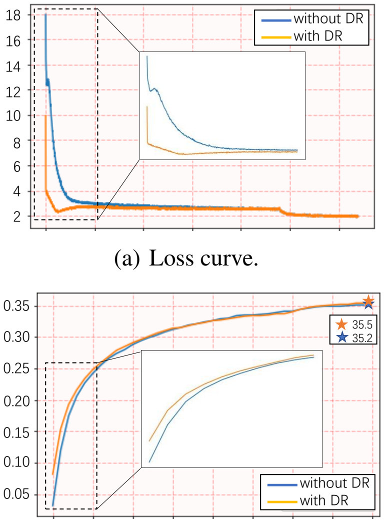
Figure 4: The qualitative comparison results of foreground boxes generated by conventional IoU matching and our De-biased Teacher.

Ablation Studies

We conduct experiments to verify the efficacy of critical components in our method. Without specification, 10% labeled COCO is adopted for network optimization.

Analysis of the Training Boxes

Our De-biased Teacher fundamentally removed the bias caused by pseudo-label-based IoU matching. To better demonstrate the effectiveness of our method, we visualize the quality of the training boxes based on the converged Soft Teacher model (Xu et al. 2021) under 10% labeled COCO protocol. Following the default setting of Soft Teacher, pseudo labels are filtered by threshold 0.9 and the results on 8000 random images are shown in Table 5. First, the precision of pseudo label is 89%, based on it, the precision of foreground boxes generated by IoU matching decreased to 77%. This proved the noisy aggravation problem of IoU matching, where the noisy information (incorrect position or category) in one single pseudo label was propagated to all the matched boxes. Second, the recall of the IoU



(a) Loss curve.

(b) mAP curve.

Figure 5: The effects of Distribution Refinement.

matching produced foreground boxes is only 46%, which means more than 50% ground truth objects are wrong labeled as background. Finally, the training boxes generated by pseudo-label-based IoU matching can only utilize less than 50% ground truth objects for optimization while the precision of the training label is less than 80%. By removing IoU matching, our De-biased Teacher can improve the model's utilization of ground truth objects by about 30%, and the training targets no longer suffer from the noisy aggravation problem of IoU matching. Figure 4 shows some qualitative results of our method, which demonstrates that our method can mine more foreground objects from unlabeled data to boost the model's detection ability.

Analysis of Distribution Refinement

The Teacher's prediction results are typical long-tailed distributions, which are hard for the Student to fit. The distribution refinement mechanism utilizes threshold δ to cut off the tails in the predicted distribution, note that δ is the threshold used for filtering foreground boxes. By doing so, background boxes will get one-hot labels while foreground boxes get refined soft labels. The effect of distribution refinement on the training process of the model is shown in Figure 5. In the early stage of training, the distribution refinement can more clearly express the correlation and non-correlation between the detection frame features and categories. The model only computes the classification loss on the strongly correlated categories. This greatly reduces the amount of computation, thereby speeding up the convergence of the model, and finally achieves the improvement of +0.3 mAP.

Effects of Threshold δ

The method we designed is very simple, requiring only one hyperparameter δ to filter foreground detection boxes and cut off the tails in the distribution. We conduct experiments to explore the effect of different thresholds on the performance of the algorithm, and the results are shown in Table 6. Best results are achieved when $\delta = 0.05$. Increasing the threshold δ leads to a decrease in the model's ability to utilize foreground objects in unlabeled data, thus impairing the performance of the algorithm. After setting δ to 0.01, the recall of the foreground frame is improved, but the final effect of the model is slightly reduced. This is because by default, the model uses 0.05 as the detection threshold during testing, and objects with predicted scores below 0.05 cannot be detected. Therefore, the additional foreground objects brought by $\delta = 0.01$ cannot improve the detection ability of the model on the validation set.

Threshold	Foreground Recall	mAP
0.01	0.840	35.1
0.05	0.782	35.5
0.1	0.742	35.4

Table 6: Effects of different threshold δ.

Conclusion

In this paper, we revisit the architecture design of object detectors and identify the fatal flaw of IoU matching when dealing with unlabeled data, which is the culprit of biased training boxes. To essentially prevent the detrimental effects of IoU matching, we propose De-biased Teacher, which replaces IoU matching by directly generating soft-labeled training boxes for consistency regularization. The selection of training boxes of our method enables the model to break away from the limitation of pseudo-label-based IoU matching and utilizes almost all the possible foregrounds within the entire image. The target distribution filtered by the distribution refinement mechanism is also more reliable than the one-hot labels generated by conventional IoU matching. Extensive experiments show that our method can stably replace the conventional IoU matching mechanism and improve the performance under various conditions.

Acknowledgments

This work was supported in part by the Shenzhen Science and Technology Program (No. JCYJ20200109142612234, NO. JCYJ20220530141211024), in part by the Guangdong Basic and Applied Basic Research Foundation (NO. 2020B1515020048), in part by the National Natural Science Foundation of China (NO. 61976250, NO. U1811463, NO. 62003256, NO. 62106235), in part by the Fundamental Research Funds for the Central Universities under Grant 22lgqb25. This work was also supported by Open Research Projects of Zhejiang Lab (NO. 2019KD0AD01/017), the Exploratory Research Project of Zhejiang Lab(2022PG0AN01), and MindSpore which is a new deep learning computing framework[1].

[1] https://www.mindspore.cn/

References

Bachman, P.; Alsharif, O.; and Precup, D. 2014. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27.

Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. 2019. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 5049–5059.

Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; Zhang, Z.; Cheng, D.; Zhu, C.; Cheng, T.; Zhao, Q.; Li, B.; Lu, X.; Zhu, R.; Wu, Y.; Dai, J.; Wang, J.; Shi, J.; Ouyang, W.; Loy, C. C.; and Lin, D. 2019. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv preprint arXiv:1906.07155*.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338.

Guo, Q.; Mu, Y.; Chen, J.; Wang, T.; Yu, Y.; and Luo, P. 2022. Scale-Equivalent Distillation for Semi-Supervised Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14522–14531.

Iscen, A.; Tolias, G.; Avrithis, Y.; and Chum, O. 2019. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5070–5079.

Li, A.; Yuan, P.; and Li, Z. 2022. Semi-Supervised Object Detection via Multi-Instance Alignment With Global Class Prototypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9809–9818.

Li, Y.; Huang, D.; Qin, D.; Wang, L.; and Gong, B. 2020. Improving object detection with selective self-supervised self-training. In *European Conference on Computer Vision*, 589–607. Springer.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, Y.-C.; Ma, C.-Y.; He, Z.; Kuo, C.-W.; Chen, K.; Zhang, P.; Wu, B.; Kira, Z.; and Vajda, P. 2021. Unbiased Teacher for Semi-Supervised Object Detection. In *ICLR*.

Mi, P.; Lin, J.; Zhou, Y.; Shen, Y.; Luo, G.; Sun, X.; Cao, L.; Fu, R.; Xu, Q.; and Ji, R. 2022. Active Teacher for Semi-Supervised Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14482–14491.

Miyato, T.; Maeda, S.-i.; Koyama, M.; and Ishii, S. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8): 1979–1993.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28: 91–99.

Sajjadi, M.; Javanmardi, M.; and Tasdizen, T. 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29.

Shao, S.; Li, Z.; Zhang, T.; Peng, C.; Yu, G.; Zhang, X.; Li, J.; and Sun, J. 2019. Objects365: A Large-Scale, High-Quality Dataset for Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Sohn, K.; Berthelot, D.; Li, C.-L.; Zhang, Z.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Zhang, H.; and Raffel, C. 2020a. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*.

Sohn, K.; Zhang, Z.; Li, C.-L.; Zhang, H.; Lee, C.-Y.; and Pfister, T. 2020b. A Simple Semi-Supervised Learning Framework for Object Detection. In *arXiv:2005.04757*.

Tang, Y.; Chen, W.; Luo, Y.; and Zhang, Y. 2021. Humble Teachers Teach Better Students for Semi-Supervised Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3132–3141.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 1195–1204.

Wang, K.; Yan, X.; Zhang, D.; Zhang, L.; and Lin, L. 2018. Towards human-machine cooperation: Self-supervised sample mining for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1605–1613.

Wang, Z.; Li, Y.; Guo, Y.; Fang, L.; and Wang, S. 2021. Data-Uncertainty Guided Multi-Phase Learning for Semi-Supervised Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4568–4577.

Xie, Q.; Dai, Z.; Hovy, E.; Luong, M.-T.; and Le, Q. V. 2020a. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*.

Xie, Q.; Luong, M.-T.; Hovy, E.; and Le, Q. V. 2020b. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10687–10698.

Xu, M.; Zhang, Z.; Hu, H.; Wang, J.; Wang, L.; Wei, F.; Bai, X.; and Liu, Z. 2021. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3060–3069.

Zheng, S.; Chen, C.; Cai, X.; Ye, T.; and Tan, W. 2022. Dual Decoupling Training for Semi-supervised Object Detection with Noise-Bypass Head. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3): 3526–3534.

Zhou, Q.; Yu, C.; Wang, Z.; Qian, Q.; and Li, H. 2021. Instant-Teaching: An End-to-End Semi-Supervised Object Detection Framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4081–4090.