Semi-supervised Object Detection with Adaptive Class-Rebalancing Self-Training

Fangyuan Zhang,[1,2] **Tianxiang Pan,** [1,2] **Bin Wang** [1,2*]

[1] School of Software, Tsinghua Unviersity
[2] Beijing National Research Center for Information Science and Technology
zhangfy19@mails.tsinghua.edu.cn, ptx9363@gmail.com, wangbins@tsinghua.edu.cn

Abstract

While self-training achieves state-of-the-art results in semi-supervised object detection (SSOD), it severely suffers from foreground-background and foreground-foreground imbalances in SSOD. In this paper, we propose an Adaptive Class-Rebalancing Self-Training (ACRST) with a novel memory module called CropBank to alleviate these imbalances and generate unbiased pseudo-labels. Besides, we observe that both self-training and data-rebalancing procedures suffer from noisy pseudo-labels in SSOD. Therefore, we contribute a simple yet effective two-stage pseudo-label filtering scheme to obtain accurate supervision. Our method achieves competitive performance on MS-COCO and VOC benchmarks. When using only 1% labeled data of MS-COCO, our method achieves 17.02 mAP improvement over the supervised method and 5.32 mAP gains compared with state-of-the-arts.

Introduction

Recently, significant progress has been witnessed in deep-learning-based object detection (Ren et al. 2015; Zhu et al. 2021; Tian et al. 2019). However, this success heavily relies on large datasets with bounding-box annotations, which are prohibitively time-consuming and expensive to collect.

Therefore, a surge of attention has been dedicated to semi-supervised object detection (SSOD), which uses a small amount of labeled data and a large amount of unlabeled data to obtain an accurate detector. In this regard, state-of-the-art SSOD performance has been achieved by the self-training paradigm (Liu et al. 2021; Zhou et al. 2021; Sohn et al. 2020), in which pseudo-labels of unlabeled data are generated as extra supervisions.

Motivations. Despite the promising results, the majority of SSOD approaches are inherited directly from advanced self-training algorithms (Tarvainen and Valpola 2017; Xie et al. 2020b; Laine and Aila 2017), which are designed specifically for classification tasks under a class-balanced data distribution. However, most real-world detection datasets have biased class distributions where few classes occupy the majority of instances, i.e. foreground-foreground imbalance as shown in Fig.1 (a). And, to obtain accurate pseudo-labels, self-training adopts a high confidence threshold. This scheme leads to sparse foreground instances distribution in detection data, i.e. foreground-background imbalance (see Fig.1 (b)).

The above two types of imbalance yield biased pseudo-labels during self-training. Subsequent training on biased supervisions further intensifies the class imbalance, thereby aggravating the performance of the final model. Unfortunately, this problem is largely overlooked in current solutions and hinders further improvements in SSOD.

To address the preceding issues, using data-rebalancing algorithms in classification tasks (Pang et al. 2019; Ouyang et al. 2016; Ren et al. 2015) is an intuitive solution. However, this idea is impeded by entanglements of foreground instances and background in detection data. Besides, directly redistributing class distributions without prior information on unlabeled data is insufficient in previous researches.

Contributions. In this work, we introduce a simple yet effective Adaptive Class-Rebalancing Self-Training (ACRST) method to redistribute pseudo-labels. ACRST consists of two detection-specific data-rebalancing algorithms: foreground-background rebalancing (FBR) and adaptive foreground-foreground rebalancing (AFFR).

Before handling class imbalance, we design a memory module called CropBank to decouple instance entanglements in detection data. CropBank stores classification and localization information of foreground instances, according to ground-truths and pseudo-labels during training. As far as we know, CropBank is the first method to allow distribution rebalancing at instance-level instead of image-level. Besides, we contribute a selective supervision scheme to reduce noise in inaccurate regression with CropBank.

We first propose FBR to address the foreground-background imbalance in SSOD. FBR samples foreground instances from CropBank and injects them into other images to produce unbiased data. In this regard, FBR directly adjusts the proportion of foreground instances in self-training and alleviates the foreground-background imbalance.

We then design AFFR based on FBR to handle the foreground-foreground imbalance. Specifically, a simple yet effective criterion called Pseudo Recall is proposed to judge which class is neglected or over-focused during training. Consequently, pseudo-labels of neglected classes are sampled more frequently because of higher negative confidence,

*corresponding author.

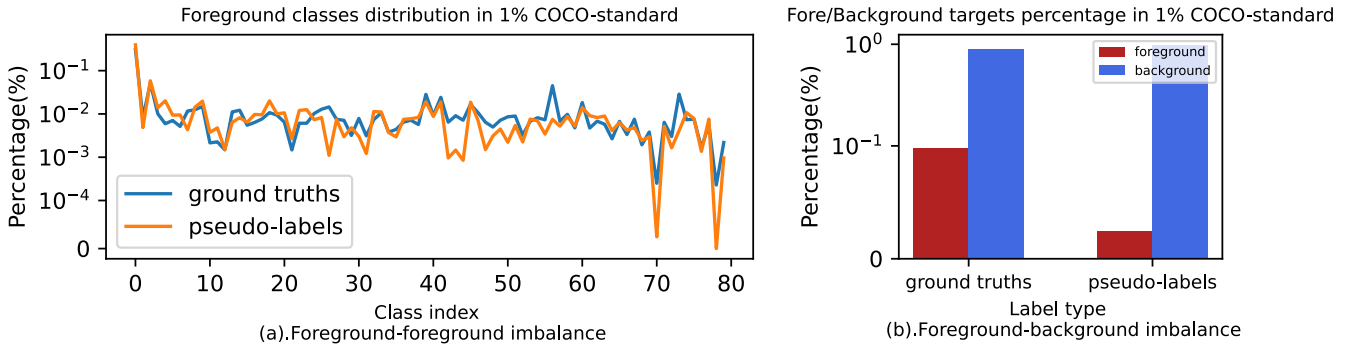

Figure 1: Class imbalance in SSOD on 1% COCO-standard. Ground truths are true labels of labeled data and pseudo-labels are generated by the teacher model.

and the class distribution is adaptively redistributed according to learning states, thus leading to a minimally biased detector in the subsequent self-training.

While FBR and AFFR are simple and effective in addressing the class imbalance in SSOD, inaccurate pseudo-labels (see Fig.2) severely hinder their effectiveness. To obtain accurate pseudo-labels, we get a free lunch from a semi-supervised multi-label learning (SSMLL) module, which provides image-level constraints complementary with the original detection confidence threshold. Thereafter, we design a two-stage filtering scheme to remove pseudo-labels that activate negative in detection confidences or multi-label predictions.

Our proposed method is simple, generic, and efficient, which can be seamlessly incorporated into other self-training pipelines for SSOD. Albeit simple, our method outperforms previous state-of-the-art results on MS-COCO and VOC benchmarks by significant margins. When using only 1% labeled COCO-standard (Lin et al. 2014), our method obtains 5.32 mAP improvement over other competitive methods. When using VOC07 (Everingham et al. 2010) as labeled data, our method outperforms state-of-the-arts by 1.26 mAP improvement.

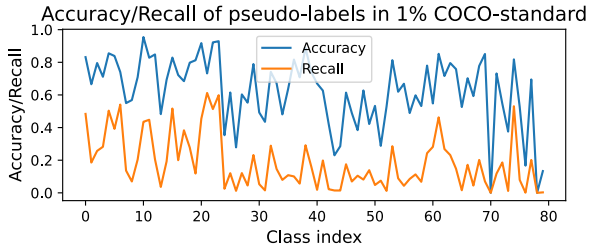

Figure 2: Accuracy and Recall of pseudo-labels in 1% COCO-standard.

Related Work

Supervised Object Detection

Existing object detection frameworks include one- and two-stage detectors. One-stage detectors (Redmon et al. 2016; Lin et al. 2017; Law and Deng 2018; Duan et al. 2019) detect directly instances on dense grids, and two-stage detectors (Ren et al. 2015; He et al. 2017; Girshick et al. 2014; Girshick 2015) first generate regions of interest (RoIs) and perform refinement on RoIs for the final predictions. We choose Faster-RCNN (Ren et al. 2015) in our experiments for a fair comparison with previous works.

Semi-supervised Learning

Recently, semi-supervised learning has achieved remarkable progress. Typical examples of SSL typically fall onto two types. One is consistency-regularization (Berthelot et al. 2019b,a; Xie et al. 2020a; Takeru et al. 2018; Sajjadi, Javanmardi, and Tasdizen 2016), enforcing variant predictions for the input under various perturbations. The other is self-training (Tarvainen and Valpola 2017; Bachman, Alsharif, and Precup 2014; Arazo et al. 2019; Iscen et al. 2019), exploiting high-quality pseudo-labels of unlabeled data as extra supervisions.

In this work, we focus on self-training, which normally assumes balanced class distributions in unlabeled datasets. Recently, cReST (Wei et al. 2021) reveals that such assumption is irrational in real-world datasets and previous researches degrade heavily on biased distributions. Concurrently, cReST introduces an effective rebalancing method, which relies on prior knowledge on the unlabeled class distribution and can not be extended to SSOD due to entangled semantics in detection tasks. In contrast, our method is simple yet efficient to handle class imbalances in detection datasets without any prior information.

Semi-supervised Object Detection

Following standard SSL settings, semi-supervised object detection has a rapid development recently. Consistency based methods, e.g. CSD (Jeong et al. 2019) and ISD (Jeong et al. 2021), impose consistency-regularization on inputs under various permutations.

Recently, self-training based methods are frequently revisited. Inherited from Noisy Student (Xie et al. 2020b), STAC (Sohn et al. 2020) introduces detection-specific data augmentations for weakly- and strongly-augmented views generation. Instant Teaching (Zhou et al. 2021) enforces

consistency between the mixed predictions and predictions of mixed inputs with MixUp (Zhang et al. 2018) and Mosaic (Bochkovskiy, Wang, and Liao 2020). Humble Teacher (Tang et al. 2021) mines more information from soft pseudo-labels. Soft Teacher (Xu et al. 2021) focuses on accurate pseudo-labels generation with uncertainty in classification and regression.

While these studies improve the detector against the supervised baseline, they lack considerations into serious class imbalances in real-world detection tasks and generate biased predictions. Recently, Unbiased-Teacher (Liu et al. 2021) applies focal loss (Lin et al. 2017) to implicitly balance the classification predictions. However, this work fails to model detection-specific imbalance in SSOD, and detectors with focal loss easily overfit in noisy pseudo-labels. To address the preceding issues, we propose ACRST to explicitly handle the class imbalance. We also contribute a two-stage pseudo-label filtering algorithm to assist ACRST and alleviate the noise in self-training.

Method

Overview

In SSOD, detectors are trained with a small labeled dataset D_l and a large unlabeled dataset D_u, where $D_l = \{x_i^l, y_i^l\}_{i=1}^{N_l}$ with bounding-box annotations y^l, and $D_u = \{x_i^u\}_{i=1}^{N_u}$. For fair comparisons, we choose Mean Teacher (MT) (Tarvainen and Valpola 2017) as the SSOD framework, and represent the overview of our framework in Fig.3. The corresponding training steps consisted of pre-training and mutual learning are clarified as follows.

Pre-training. The student model is first pre-trained with D_l via gradient back-propagation, and then the teacher model resumes from the student model. Pre-training generates noisy-less pseudo-labels, thereby facilitating the subsequent mutual training.

Teacher-Student Mutual Learning. In the mutual learning stage, the student model is trained with ground truths and pseudo-labels. The student model is updated via the gradient back-propagation, and the teacher model is updated via exponential moving average (EMA):

$$\theta_s \leftarrow \theta_s + \frac{\partial \mathcal{L}}{\partial \theta_s}, \tag{1}$$

$$\theta_t \leftarrow \lambda_{ema}\theta_t + (1-\lambda_{ema})\theta_s, \tag{2}$$

where θ_s/θ_t represents the model parameters of the student/teacher model, and λ_{ema} is the parameter for EMA. $\mathcal{L}$ represents the total SSOD losses, i.e. a combination of losses on labeled data $\mathcal{L}_{sup}$ and unlabeled data $\mathcal{L}_{unsup}$:

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda_{unsup}\mathcal{L}_{unsup}, \tag{3}$$

$$\begin{aligned}\mathcal{L}_{sup} = \Sigma_i \mathcal{L}_{cls}^{rpn}(x_i^l, y_i^l) + \mathcal{L}_{reg}^{rpn}(x_i^l, y_i^l) \\ + \mathcal{L}_{cls}^{roi}(x_i^l, y_i^l) + \mathcal{L}_{reg}^{roi}(x_i^l, y_i^l),\end{aligned} \tag{4}$$

$$\mathcal{L}_{unsup} = \Sigma_i \mathcal{L}_{cls}^{rpn}(x_i^u, \widetilde{y}_i^u) + \mathcal{L}_{cls}^{roi}(x_i^u, \widetilde{y}_i^u), \tag{5}$$

where $\mathcal{L}_{cls}^{rpn}$, $\mathcal{L}_{reg}^{rpn}$, $\mathcal{L}_{cls}^{roi}$, $\mathcal{L}_{reg}^{roi}$ respectively represent loss functions of RPN classification, RPN regression, ROI classification and ROI regression. y_i^l represents the annotation of the labeled image x_i^l, and $\widetilde{y}_i^u$ represents the pseudo-labels of unlabeled image x_i^u. λ_{unsup} is used to balance the supervised and unsupervised losses. Note that regression losses are removed in L_{unsup} in previous studies for denoising.

In the following, we first introduce the CropBank for semantic disentanglement. Then, we elaborate on our proposed ACRST consisted of FBR and AFFR. Subsequently, we clarify the two-stage pseudo-label filtering algorithm to obtain accurate supervisions. Lastly, we introduce the selective supervision scheme for regression learning.

CropBank

Despite the effectiveness of data-resampling algorithms in distribution rebalancing, they are heavily hindered by strong entanglements between foreground instances and background in detection data. To decouple such interconnections, we propose a novel memory module called CropBank, which incorporates two sub-banks. One is Labeled CropBank $\Phi_L = \{y_i^l\}_{i=1}^{N_L}$, absorbing N_L ground truths from labeled images. The other is Pseudo CropBank $\Phi_U = \{\widetilde{y}_i^u\}_{i=1}^{N_U}$, accumulating N_U pseudo-labels generated by the teacher model.

In the implementation, the CropBank brings negligible memory and time consumption for only storing instance-level annotations. In the self-training, Φ_L is fixed once generated, while Φ_U is updated periodically with improved pseudo-labels in mutual training. CropBank supports the data resampling at the instance-level, based on which we design adaptive class-rebalancing self-training (ACRST) to handle the class imbalance in SSOD.

Adaptive Class-Rebalancing Self-Training

While self-training is an ideal solution to alleviate the lack of human annotations, it is hindered by the inherent class imbalance in real-world detection datasets. To handle the class imbalance in SSOD, we propose Adaptive Class-Rebalancing Self-Training (ACRST), which consists of foreground-background rebalancing (FBR) and adaptive foreground–foreground rebalancing (AFFR).

Foreground-Background Rebalancing. Models trained on foreground-background imbalanced data often overfit in background instances (Lin et al. 2017). While various solutions (Lin et al. 2017; Ren et al. 2015) have been proposed, they heavily rely on ground truths to redistribute training data. In contrast, we use abundant instance-level annotations with few ground truths and lots of pseudo-labels in CropBank to rebalance the foreground-background distribution.

Given a training data $\{x_i, y_i\}$, we fetch a set of foreground instances $F = \{c_j, y_j\}_{j=1}^{N_C}$ from the CropBank Φ_L and Φ_U for image x_i following a sample distribution P, where c_j is a foreground instance cropped from original image according to annotation y_j, and $N_C \in$ sample range $[N_{min}, N_{max}]$. Then, the new training data $\{x_i^{mix}, y_i^{mix}\}$ is generated as follows:

$$x_i^{mix} = \alpha x_i + (1-\alpha)c_j, \tag{6}$$

$$y_i^{mix} = merge(y_i, y_j), \tag{7}$$

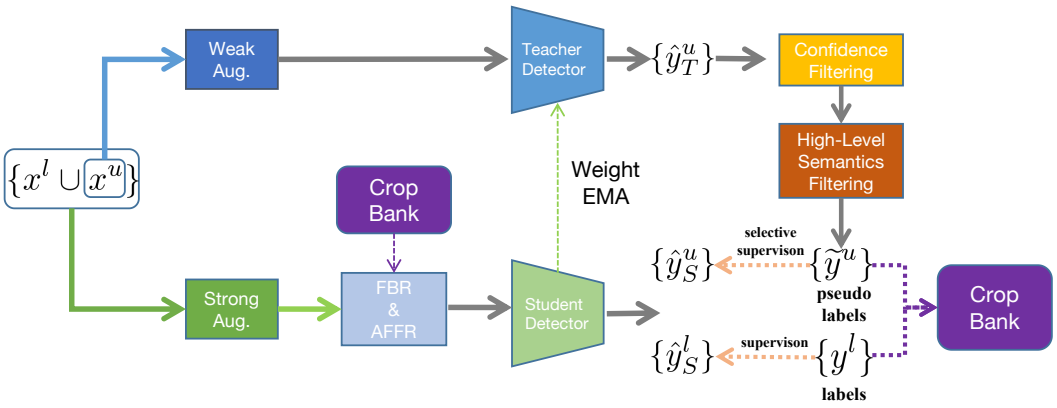

Figure 3: An overview of our semi-supervised object detection framework. The teacher model generates pseudo-labels from weakly-augmented unlabeled data and the student model is trained on strongly-augmented data with a combination of ground truths and pseudo-labels. To alleviate the class imbalance in SSOD, we first design a memory module called CropBank. Then, the foreground-background rebalancing (FBR) and adaptive foreground-foreground rebalancing (AFFR) are applied for adaptive class-rebalancing self-training (ACRST) based on CropBank. We also contribute a two-stage pseudo-label filtering (TPF) method and a selective supervision scheme to assist ACRST and generate accurate pseudo-labels.

where α is a binary mask of c_j, and y_i^{mix} denotes the new annotations, in which fully occluded instances are removed from the new image x_i^{mix}. During training, c_j is augmented and pasted to random locations of x_i. This combination directly increases the ratio of foreground instances, thereby rebalancing foreground-background distribution.

In the implementation, the data-rebalancing is seamlessly incorporated into strong data augmentations and brings no restriction on the SSOD framework. Besides, as discussed in the following section, such a crop-paste operation reduces noise in pseudo-labels and enables accurate regression with selective supervision.

Adaptive Foreground-Foreground Rebalancing. FBR adequately alleviates the foreground-background imbalance with considerable attention to foreground instances. However, sampling randomly or uniformly foreground instances from the CropBank fails to handle the foreground-foreground imbalance. Hence, we contribute an adaptive sample strategy, in which samples in neglected classes during self-training are selected more frequently.

To measure the neglected degree of each class, we propose a novel criterion Pseudo Recall (PR). For each category k, we empirically use a low threshold (0.1) to filter noisy predictions. Then detection confidences from Teacher Detector for each foreground instance are accumulated to calculate PR_k:

$$PR_k = \Sigma_{i=1}^{N_k} s_i^k, \tag{8}$$

where s_i^k is the detection confidence for i-th pseudo-label.

PR defines how neglected one class is in SSOD. High PR_k indicates that the detector is certain even over-confident on class k. Consequently, lower sample probabilities should be allocated to samples in class k to avoid overfitting. And, low PR_k implies that the detector lacks confidence for detecting instances of class k. Therefore, these instances should be selected more frequently in subsequent training. When categories are similarly neglected, lower PR is adaptively assigned to tail categories and raises increasing attention on them. Besides, unlike cReST (Wei et al. 2021), the definition of PR does not rely on any prior information on unlabeled data.

With PR, we design an adaptive sample strategy:

$$\mu_k = \frac{(1/PR_k)^\beta}{\Sigma_{j=1}^{K} (1/PR_j)^\beta}, \tag{9}$$

where μ_k is the probability of choosing instances of class k, and K is the number of categories. β is used to adjust the sample probability. This mechanism adaptively allocates higher/lower sample rates to neglected/over-focused instances. Note that AFFR performs FBR simultaneously.

Two-stage Pseudo-label Filtering

While proposed ACRST considerably alleviates the class imbalance in SSOD. However, its effectiveness is heavily affected by the quality of pseudo-labels. Once noise in the CropBank is selected improperly, it will be amplified undesirably in self-training. While a high threshold (0.9) is usually used in semi-supervised classification/segmentation (Berthelot et al. 2019b) to select accurate pseudo-labels, it is necessary to adopt a relative low threshold (0.7) in SSOD (Liu et al. 2021; Zhou et al. 2021) to ensure enough yet noisy pseudo-labels, which are unfriendly to ACRST. To alleviate the above issues, we propose a semi-supervised multi-label classification module to provide high-level semantics (i.e., image-level pseudo-labels) for two-stage pseudo-label filtering.

Semi-supervised Multi-label Learning. The proposed semi-supervised multi-label learning (SSMLL) module is devised based on ResNet50-based CTran (Lanchantin et al. 2021) following Mean Teacher pipeline. For each image x_i, we predict its image-level pseudo-labels $v_i = \{l_k\}_{k=1}^K, l_k \in \{0, 1\}$, where K is the number of classes and l_k indicates whether there are instances of class k in the image. In the

training stage, predictions of the teacher model are converted to image-level pseudo-labels, and a focal binary cross entropy loss to optimize the student model. SSMLL is a much easier auxiliary task compared with SSOD and enables reliable references generation for two-stage pseudo-label selection. Note that we also extend ACRST to alleviate the class imbalance in SSMLL, and the total training of SSMLL only takes $\frac{1}{5}$ of SSOD's time due to fewer steps, smaller input size, and a more simplified framework.

Two-stage Pseudo-label Filtering. For predictions from the teacher model, we adopt a two-stage pseudo-label filtering scheme to get accurate pseudo-labels with confidence scores s and image-level pseudo-labels v. In the first stage, predictions with scores $s < \tau_{cls}$ are removed to get pseudo-labels with high objectness. In the second stage, predictions with classes that activate negative in v (i.e., activation values are smaller than τ_{ml}) are removed to get pseudo-labels with correct class labels. Note that we use negative instead of positive multi-label as references because negative learning has much higher accuracy and recall than positive learning.

Selective Supervision

While bounding-box regression losses in previous SSOD researches (Liu et al. 2021) are removed due to inaccurate regression, they are beneficial for our framework. We attribute the success to the CropBank, which alleviates noise from partially detected instances that take a large proportion (81.2% in 1% COCO-standard) in biased predictions. Learning blindly with these noisy pseudo-labels will heavily aggravate the model performance. However, in our work, when the partially detected instances from the CropBank are cropped and pasted to other images, they become independent and complete in new backgrounds, thereby providing additional accurate supervision for regression learning.

With selective supervision, loss function $\mathcal{L}_{unsup}$ in Equation 5 can be represented as follows:

$$\begin{aligned}\mathcal{L}_{unsup} = \Sigma_i \mathcal{L}_{cls}^{rpn}(x_i^u, \widetilde{y}_i^u) + \mathcal{L}_{reg}^{rpn}(x_i^u, \widetilde{y}_i^{ss}) \\ + \mathcal{L}_{cls}^{roi}(x_i^u, \widetilde{y}_i^u) + \mathcal{L}_{reg}^{roi}(x_i^u, \widetilde{y}_i^{ss}),\end{aligned} \tag{10}$$

where $\widetilde{y}_i^{ss}$ are the instances from CropBank.

Experiments

Datasets

We evaluate our method on three SSOD benchmarks from MS-COCO (Lin et al. 2014) and PASCAL VOC (Everingham et al. 2010). (1).*COCO-standard*: We sample $0.5/1/2/5/10\%$ of the COCO2017-train as the labeled dataset and take the remaining data as the unlabeled dataset. (2).*COCO-additional*: We use the COCO2017-train as the labeled dataset and the additional COCO2017-unlabeled as the unlabeled dataset. (3).*VOC*: We use the VOC07-trainval as the labeled dataset and the VOC12-trainval as the unlabeled dataset. We evaluate the model on the COCO2017-val for (1)(2) and VOC07-test for (3).

Implementation Details

For fair comparisons, we follow previous methods (Sohn et al. 2020; Liu et al. 2021) to use Faster-RCNN with FPN and ResNet50 and build our framework upon the Detectron2 (Wu et al. 2019). Following (Liu et al. 2021), the batch-sizes of labeled and unlabeled images are both 32. We use the SGD optimizer with learning rate=0.01 and momentum rate=0.9. We set $\lambda_{ema} = 0.9996$, $\tau_{cls} = 0.7$, $\lambda_{unsup} = 4$. For specific parameters in our work, we set $\beta = 0.6$, and $\tau_{ml} = 0.2$. The pre-training takes 3000/5000/5000/5000/10000 steps and the total training takes 180000 steps for $0.5/1/2/5/10\%$ *COCO-standard*. For *VOC*, the pre-training takes 5000 steps and the total training takes 72000 steps. We apply color jittering, Gaussian blur and CutOut for strong augmentations, and we apply randomly resize and flip, crop for weak augmentations. The widely used mAP ($AP_{50:95}$) serves as metric for comparisons. For SSMLL, the batch-sizes of labeled and unlabeled images are both 64. The pre-training takes 2k/2k/6k steps and the total training takes 18k/36k/96k steps for VOC/COCO-standard/COCO-additional, where we use Adam optimizer with lr=1e-5. Data augmentations are the same with SSOD but images are resized into 576*576.

Results and Comparisons

COCO-standard & COCO-additional. We first evaluate our method on COCO-standard. As shown in Table 1, when using only 1% to 10% labeled data, our model consistently performs better against all previous approaches. When trained on the 1% COCO-standard, our method achieves 5.32 mAP improvement compared with Unbiased-Teacher, and 3.61 mAP improvement than CSD trained on 10% COCO-standard. When using 10% COCO-standard, our method achieves 11.06 mAP improvement compared with supervised baselines. In Table 2, our model has a 0.72 mAP gains on COCO-additional and 3.08 mAP gains on 0.5% COCO-standard compared with previous methods. This result indicates that our method achieves satisfying gains even on extremely small/large-scale labeled datasets. We attribute the success of model performance to the class rebalanced data and accurate pseudo-labels.

VOC. We evaluate models on a balanced dataset VOC to demonstrate the generalization of our method. Table 3 provides the mAP results of CSD, STAC, Unbiased Teacher, Humble Teacher, and ours. Our method achieves 7.99 mAP improvement compared with the supervised baseline and 1.26 mAP improvement against Humble Teacher, even though Humble Teacher has witnessed performance saturation in VOC. We owe the success to the generalization ability of ACRST. Albeit training data is already foreground-foreground balanced in VOC, FBR alleviates the inevitable foreground-background imbalance in SSOD. Besides, the two-stage pseudo-label filtering scheme and selective supervision further improve the model performance.

Ablation Studies

Foreground–Background Rebalancing. We first verify the effect of FBR. Table 4 shows that applying FBR improves mAP in 1% labeled COCO from 21.05 to 23.32. To

	COCO-standard ($AP_{50:95}$)			
	1%	2%	5%	10%
Supervised	9.05 ± 0.16	12.70 ± 0.15	18.47 ± 0.22	23.86 ± 0.81
CSD (Jeong et al. 2019)	$10.51 \pm 0.06(+1.46)$	$13.93 \pm 0.12(+1.23)$	$18.63 \pm 0.07(+0.16)$	$22.46 \pm 0.08(-1.4)$
STAC (Sohn et al. 2020)	$13.97 \pm 0.35(+4.92)$	$18.25 \pm 0.25(+5.55)$	$24.38 \pm 0.12(+5.91)$	$28.64 \pm 0.21(+4.78)$
Instant Teaching (Zhou et al. 2021)	$18.05 \pm 0.15(+9.00)$	$22.45 \pm 0.15(+9.75)$	$26.75 \pm 0.05(+8.28)$	$30.40 \pm 0.05(+6.54)$
Unbiased Teacher (Liu et al. 2021)	$20.75 \pm 0.12(+11.70)$	$24.30 \pm 0.07(+11.60)$	$28.27 \pm 0.11(+9.80)$	$31.5 \pm 0.10(+7.64)$
Humble Teacher (Tang et al. 2021)	$16.96 \pm 0.38(+7.91)$	$21.72 \pm 0.24(+9.02)$	$27.70 \pm 0.15(+9.23)$	$31.61 \pm 0.28(+7.75)$
Soft Teacher (Xu et al. 2021)	$20.46 \pm 0.39(+11.41)$	-	$30.74 \pm 0.08(+12.27)$	$34.04 \pm 0.14(+10.18)$
Ours	**26.07** $\pm$ **0.26**$(+17.02)$	**28.69** $\pm$ **0.17**$(+15.99)$	**31.63** $\pm$ **0.13**$(+13.16)$	**34.92** $\pm$ **0.22**$(+11.06)$

Table 1: Comparison with the state-of-the-arts on 1% to 10% COCO-standard.

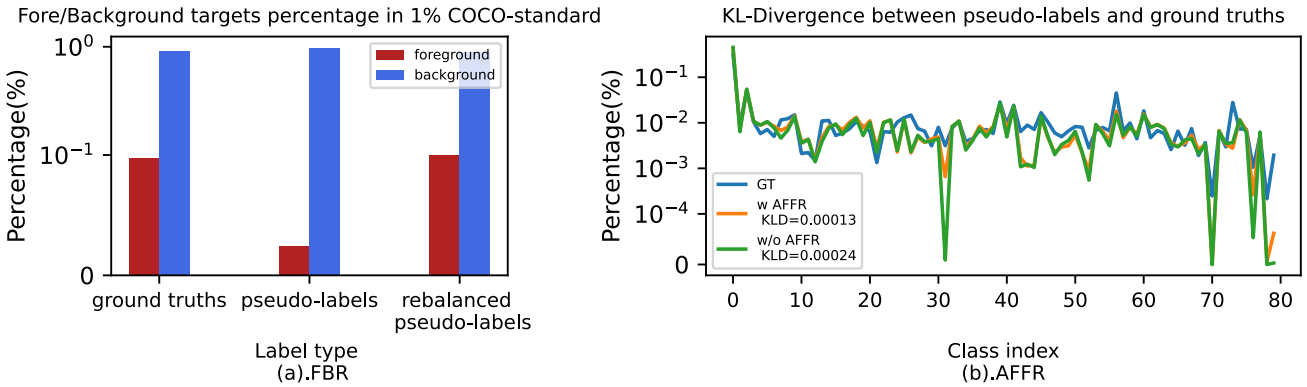

Figure 4: Ablation study on (a) FBR and (b) AFFR.

	$AP_{50:95}$	
	COCO-addtional	0.5% COCO-standard
Supervised	40.20	6.83
CSD	$38.82(-1.38)$	$7.41(+0.58)$
STAC	$39.21(-0.99)$	$9.78(+2.95)$
Unbiased Teacher	$41.30(+1.10)$	$16.94(+10.11)$
Humble Teacher	$42.17(+1.97)$	–
Ours	**42.89**$(+2.69)$	**20.02**$(+13.19)$

Table 2: Comparison with the state-of-the-arts on COCO-additional and 0.5% COCO-standard.

	AP_{50}	$AP_{50:95}$
Supervised	72.63	42.13
CSD	$74.70(+2.07)$	-
STAC	$77.45(+4.82)$	$44.64(+2.51)$
Unbiased Teacher	$77.37(+4.74)$	$48.69(+6.56)$
Humble Teacher	$80.94(+8.31)$	$53.04(+10.91)$
Ours	**81.11**$(+8.48)$	**54.30**$(+12.17)$

Table 3: Comparison with the state-of-the-arts on VOC.

FBR	AFFR	Two-Stage	SS	$AP_{50:95}$
				21.05
		✓		$23.48(+2.43)$
✓				$23.32(+2.27)$
✓	✓			$24.36(+3.31)$
✓	✓	✓		$25.56(+4.51)$
✓	✓	✓	✓	**26.12**$(+5.07)$

Table 4: Ablation study on 1% COCO-standard.

analyze the divergent results, we visualize the foreground-background distribution of the rebalanced pseudo-labels. As shown in Fig.4 (a), the distribution of the foreground instances is rebalanced after FBR. The ratio of foreground instances in rebalanced pseudo-labels is even higher than that of ground truths. Hence, training detectors with rebalanced training data alleviates data bias and produces high mAP. We also perform ablation studies on the type of CropBank and sample range $[N_{min}, N_{max}]$. As shown in Table 5, sampling instances from both Labeled and Pseudo CropBank with a large random sample range achieves the highest mAP.

CropBank	N_{min}	N_{max}	$AP_{50:95}$
Labeled	0	10	25.42
Pseudo	0	10	25.96
Labeled + Pseudo	0	5	26.04
Labeled + Pseudo	0	10	**26.12**
Labeled + Pseudo	10	10	25.74

Table 5: Ablation study on CropBank and sample ranges.

Adaptive Foreground–Foreground Rebalancing. As shown in Table 4, AFFR improves 3.31 mAP compared to supervised baseline. We further verify the effectiveness of AFFR by analyzing the KL-divergence between the distribution of ground truths and pseudo-labels. Fig.4 (b) indicates that when using AFFR, the KL-divergence is reduced from 0.00024 to 0.00013. This result further confirms the effectiveness of AFFR in handling foreground–foreground imbalance in pseudo-labels and generating unbiased data distributions. And, we explore the selection of hyper-parameter β.

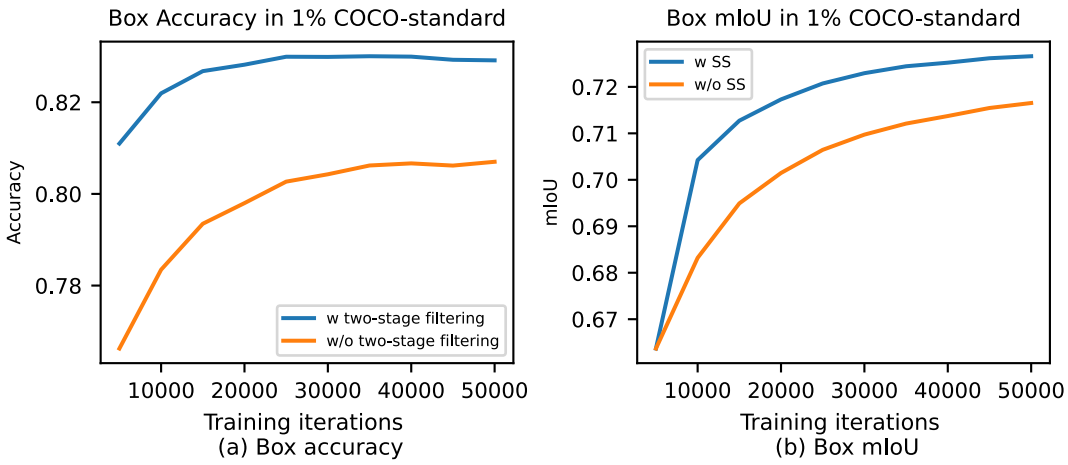

Figure 5: Pseudo-labels improvements in Box Accuracy and Box mIoU in 1% COCO-standard.

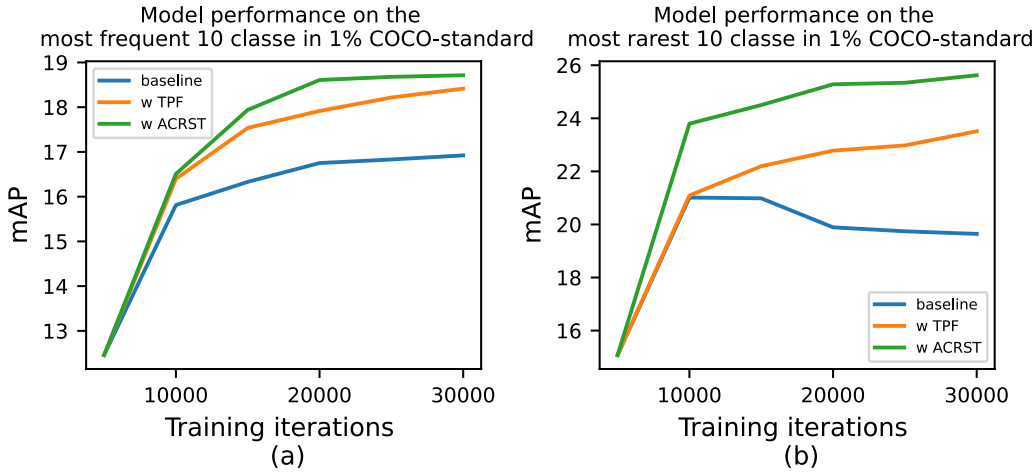

Figure 6: Effect of TPF and ACRST on neglected/over-focused classes on 1% COCO-standard.

As shown in Table 6, equipped with $\beta = 0$, AFFR is equivalent to uniform sample and degrades to FBR. With larger $\beta = 0.6$, AFFR delivers 1.04 mAP performance gains. Note that AFFR with $\beta = 0.4$ or $\beta = 0.8$ obtains similar gains, these results prove that AFFR is insensitive to the only hyper-parameter β.

β	0	0.2	0.4	0.6	0.8
$AP_{50:95}$	23.32	23.90	24.22	**24.36**	24.28

Table 6: Results for different values of β in AFFR.

Two-stage Pseudo-label Filtering. We also verify the effectiveness of the two-stage pseudo-label filtering with detection confidences and image-level pseudo-labels. As presented in Table 4, the model that filters pseudo-labels with additional multi-label information favorably achieves 2.43 performance gains compared to single-stage filtering. Fig.5 (a) shows a continuous improvement in the accuracy of pseudo-labels with the two-stage filtering scheme effective in removing noisy predictions in SSOD. Besides, the two-stage filtering scheme is necessary to build an accurate Pseudo CropBank and improve the performance of ACRST. Table 4 indicates that applying the two-stage filtering scheme to ACRST improves the mAP from 24.36 to 25.56. All the results confirm that the two-stage filtering scheme is effective in handling the noisy pseudo-labels.

Selective Supervision. In this section, we examine the effectiveness of selective supervision in SSOD. As presented in Table 4, the selective supervision improves the mAP from 25.56 to 26.12 in 1% COCO-standard. We owe the improvement to the crop-paste operation in ACRST, in which incomplete instances are pasted to new backgrounds. Accordingly, transferring these incomplete predictions to complete objects in a new background alleviates regression noise in the pseudo-labels and improves the model performance. We further analyze the accuracy of regression in pseudo-labels. As shown in Fig.5 (b), selective supervision continuously improves the mIoU of pseudo-labels. While selective supervision is an effective method to exploit partially detected pseudo-labels in SSOD, there is still room for improvement. For instance, the current strategy fails to handle noise when

objects are overlapped with each other in pseudo-labels.

Ablation Study on other SSOD frameworks. To prove that our method can be seamlessly incorporated into other SSOD frameworks, we re-implement a representative work STAC (Sohn et al. 2020), equipped with proposed ACRST, two-stage pseudo-label filtering (Two-stage), and selective supervision (SS). As shown in Table 7, while pseudo-labels in STAC are not updated online, our proposed methods achieve significant gains on 1% COCO-standard and show strong generalization ability.

Method	$AP_{50:95}$
STAC	13.97
STAC+ACRST	15.52(+1.55)
STAC+ACRST+Two-stage	16.64(+2.67)
STAC+ACRST+Two-stage+SS	**16.92**(+2.95)

Table 7: Ablation study for STAC on 1% COCO-standard.

Ablation Study on the Most Frequent and Rarest Classes. We perform another ablation study on the effect of proposed modules on the over-focused(most frequent)/neglected(rarest) classes in Fig. 6. The results in both Fig. 6 (a) and (b) indicate that both two-stage pseudo-labels filtering (TPF) and ACRST perform well on over-focused/neglected classes. As shown in (b), ACRST achieves significant improvements on the neglected classes with AFFR, while baseline has witnessed a performance drop in the rarest classes.

Additional Results and Analysis

CropBank:A Strong Data Augmentation for Detection. Appropriate strong augmentations play a vital role in semi-supervised learning (SSL). While image-level data augmentations (e.g.color jittering, CutOut (Devries and Taylor 2017)) are effective in boosting SSL on classification, they are not powerful enough for SSOD (Zhou et al. 2021). Recently, (Zhou et al. 2021) combines MixUp (Zhang et al. 2018) and Mosaic (Bochkovskiy, Wang, and Liao 2020) as a strong augmentation to change the image semantics and improves the model performance. However, MixUp and Mosaic are designed specifically for the classification and degrade in SSOD. While CropBank is designed for ACRST, it is a strong detection-specific augmentation for SSOD. The strength of CropBank is two-folds. First, CropBank decouples foreground instances and background in detection data and creates complicated training data with decoupled elements. Second, the CropBank alleviates noise in pseudo-labels with selective supervision.

To verify the effectiveness of CropBank, we provide the results of Instant Teaching (Zhou et al. 2021) with different data augmentations in Table 8. The CropBank improves the mAP from 16.00 to 16.85 compared to MixUp and Mosaic.

Semi-supervised Multi-Label Learning. Here, we provide the results from the semi-supervised multi-label learning (SSMLL) with different τ_{ml} and corresponding SSOD performance. As shown in Table 9, SSMLL generates accurate image-level pseudo-labels and the SSOD performance

Augmentations	$AP_{50:95}$
MixUp and Mosaic (Zhou et al. 2021)	16.00
CropBank	**16.85**

Table 8: Instant Teaching performance under different data augmentations on 1% COCO-standard.

is insensitive to τ_{ml}. Note that positive image-level pseudo-labels are less accurate, the accuracy is 0.740 and the recall is 0.325 when using a 0.7 threshold.

τ_{ml}	Accuracy	Recall	$AP_{50:95}$
0.05	0.994	0.968	23.27
0.1	0.992	0.984	23.28
0.2	0.990	0.991	**23.32**

Table 9: Accuracy and Recall of image-level negative pseudo-labels on 1% COCO-standard.

Then we clarify the reasons for using negative instead of positive pseudo-labels as reference. We provide the results in three settings: (1) **Single-stage:** Predictions with low detection confidence are filtered. (2) **Two-stage filtering:** Predictions with low detection confidence or activating negative in image-level pseudo-labels are filtered. (3) **Two-stage Mining:** Predictions with high detection confidence or activating positive in image-level pseudo-labels are reserved.

Setting	Accuracy	Recall	$AP_{50:95}$
Single-stage	0.788	0.377	21.05
Two-stage Filtering	**0.815**	0.367	**23.48**
Two-stage Mining	0.712	**0.448**	21.79

Table 10: Model Performance, Accuracy and Recall of pseudo-labels on 1% COCO-standard.

As shown in Table 10, while the two-stage mining achieves higher recall gains compared with the two-stage filtering, the latter achieves 1.69 mAP gains. This result indicates that the improvement in accuracy of pseudo-labels is relatively important in SSOD.

Conclusion

This study proposes a simple but effective ACRST to address the class imbalance in SSOD. With CropBank, ACRST considerably alleviates foreground-background and foreground-foreground imbalances with FBR and AFFR. To further improve FBR and AFFR, we design a two-stage pseudo-label filtering algorithm with detection confidences and high-level semantics. Over iterations on rebalanced training data, SSOD detectors become unbiased and ameliorate the model performance progressively. Extensive experiments demonstrate the effectiveness of our method.

Acknowledgments

This work was supported by the NSFC under Grant 62072271.

References

Arazo, E.; Ortego, D.; Albert, P.; O'Connor, N. E.; and McGuinness, K. 2019. Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning. *CoRR*, abs/1908.02983.

Bachman, P.; Alsharif, O.; and Precup, D. 2014. Learning with Pseudo-Ensembles. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 3365–3373.

Berthelot, D.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Sohn, K.; Zhang, H.; and Raffel, C. 2019a. ReMixMatch: Semi-Supervised Learning with Distribution Alignment and Augmentation Anchoring. *CoRR*, abs/1911.09785.

Berthelot, D.; Carlini, N.; Goodfellow, I. J.; Papernot, N.; Oliver, A.; and Raffel, C. 2019b. MixMatch: A Holistic Approach to Semi-Supervised Learning. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 5050–5060.

Bochkovskiy, A.; Wang, C.; and Liao, H. M. 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. *CoRR*, abs/2004.10934.

Devries, T.; and Taylor, G. W. 2017. Improved Regularization of Convolutional Neural Networks with Cutout. *CoRR*, abs/1708.04552.

Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; and Tian, Q. 2019. CenterNet: Keypoint Triplets for Object Detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 6568–6577. IEEE.

Everingham, M.; Gool, L. V.; Williams, C.; Winn, J.; and Zisserman, A. 2010. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2): 303–338.

Girshick, R. B. 2015. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 1440–1448. IEEE Computer Society.

Girshick, R. B.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, 580–587. IEEE Computer Society.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. B. 2017. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2980–2988. IEEE Computer Society.

Iscen, A.; Tolias, G.; Avrithis, Y.; and Chum, O. 2019. Label Propagation for Deep Semi-Supervised Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 5070–5079. Computer Vision Foundation / IEEE.

Jeong, J.; Lee, S.; Kim, J.; and Kwak, N. 2019. Consistency-based Semi-supervised Learning for Object detection. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 10758–10767.

Jeong, J.; Verma, V.; Hyun, M.; Kannala, J.; and Kwak, N. 2021. Interpolation-Based Semi-Supervised Learning for Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 11602–11611. Computer Vision Foundation / IEEE.

Laine, S.; and Aila, T. 2017. Temporal Ensembling for Semi-Supervised Learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Lanchantin, J.; Wang, T.; Ordonez, V.; and Qi, Y. 2021. General Multi-Label Image Classification With Transformers. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 16478–16488. Computer Vision Foundation / IEEE.

Law, H.; and Deng, J. 2018. CornerNet: Detecting Objects as Paired Keypoints. *ArXiv*, abs/1808.01244.

Lin, T.; Goyal, P.; Girshick, R. B.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2999–3007. IEEE Computer Society.

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 Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision – ECCV 2014*, 740–755. Cham: Springer International Publishing. ISBN 978-3-319-10602-1.

Liu, Y.; Ma, C.; He, Z.; Kuo, C.; Chen, K.; Zhang, P.; Wu, B.; Kira, Z.; and Vajda, P. 2021. Unbiased Teacher for Semi-Supervised Object Detection. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Ouyang, W.; Wang, X.; Zhang, C.; and Yang, X. 2016. Factors in Finetuning Deep Model for Object Detection with Long-Tail Distribution. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 864–873. IEEE Computer Society.

Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; and Lin, D. 2019. Libra R-CNN: Towards Balanced Learning for Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 821–830. Computer Vision Foundation / IEEE.

Redmon, J.; Divvala, S. K.; Girshick, R. B.; and Farhadi, A. 2016. You Only Look Once: Unified, Real-Time Object

Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 779–788. IEEE Computer Society.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Cortes, C.; Lawrence, N. D.; Lee, D. D.; Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, 91–99.

Sajjadi, M.; Javanmardi, M.; and Tasdizen, T. 2016. Regularization With Stochastic Transformations and Perturbations for Deep Semi-Supervised Learning. In Lee, D. D.; Sugiyama, M.; von Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 1163–1171.

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

Takeru, M.; Shin-Ichi, M.; Shin, I.; and Masanori, K. 2018. Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1.

Tang, Y.; Chen, W.; Luo, Y.; and Zhang, Y. 2021. Humble Teachers Teach Better Students for Semi-Supervised Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 3132–3141. Computer Vision Foundation / IEEE.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 1195–1204.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. FCOS: Fully Convolutional One-Stage Object Detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 9626–9635. IEEE.

Wei, C.; Sohn, K.; Mellina, C.; Yuille, A. L.; and Yang, F. 2021. CReST: A Class-Rebalancing Self-Training Framework for Imbalanced Semi-Supervised Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 10857–10866. Computer Vision Foundation / IEEE.

Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.-Y.; and Girshick, R. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Xie, Q.; Dai, Z.; Hovy, E. H.; Luong, T.; and Le, Q. 2020a. Unsupervised Data Augmentation for Consistency Training. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Xie, Q.; Luong, M.; Hovy, E. H.; and Le, Q. V. 2020b. Self-Training With Noisy Student Improves ImageNet Classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 10684–10695. IEEE.

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. *CoRR*, abs/2106.09018.

Zhang, H.; Cissé, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Zhou, Q.; Yu, C.; Wang, Z.; Qian, Q.; and Li, H. 2021. Instant-Teaching: An End-to-End Semi-Supervised Object Detection Framework. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 4081–4090. Computer Vision Foundation / IEEE.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2021. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.