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

Fangyuan Zhang,[1 2] **Tianxiang Pan,** [1 3] **Bin Wang** [1 4]

[1] Software School of Tsinghua University
[2] zhangfy19@mails.tsinghua.edu.cn
[3] ptx9363@gmail.com
[4] wangbins@mails.tsinghua.edu.cn

Abstract

This study delves into semi-supervised object detection (SSOD) to improve detector performance with additional unlabeled data. State-of-the-art SSOD performance has been achieved recently by self-training, in which training supervision consists of ground truths and pseudo-labels. In current studies, we observe that class imbalance in SSOD severely impedes the effectiveness of self-training. To address the class imbalance, we propose adaptive class-rebalancing self-training (ACRST) with a novel memory module called Crop-Bank. ACRST adaptively rebalances the training data with foreground instances extracted from the CropBank, thereby alleviating the class imbalance. Owing to the high complexity of detection tasks, we observe that both self-training and data-rebalancing suffer from noisy pseudo-labels in SSOD. Therefore, we propose a novel two-stage filtering algorithm to generate accurate pseudo-labels. Our method achieves satisfactory improvements on MS-COCO and VOC benchmarks. When using only 1% labeled data in MS-COCO, our method achieves 17.02 mAP improvement over supervised baselines, and 5.32 mAP improvement compared with state-of-the-art methods.

Introduction

Object detection to classify and localize objects in the image is one of the most important research topics in computer vision. In recent years, significant progress has been witnessed in deep-learning-based object detection. The majority of the existing studies follow a fully supervised setting and heavily rely on large datasets with bounding-box annotations. However, creating fully annotated detection datasets costs thousands of hours (Russakovsky, Li, and Fei-Fei 2015; Dollar et al. 2012), thereby hindering the practicability of current studies. Therefore, a surge of attention has been dedicated to semi-supervised object detection (SSOD). Although SSOD has made immense progress, the current SSOD methods and their fully supervised counterparts continue to have a significant performance gap.

State-of-the-art SSOD performance has been achieved recently by the self-training paradigm, in which pseudo-labels of unlabeled data are generated to train detectors. However, the majority of advanced self-training algorithms (Tarvainen and Valpola 2017; Xie et al. 2020b; Laine and Aila 2017) are designed specifically for classification. In the experiments, we observe that using them directly is suboptimal when *class imbalance* in SSOD considerably hinders the use of self-training.

Class imbalance is a longstanding challenge in object detection that mainly includes foreground-background imbalance (Lin et al. 2017; Ren et al. 2015; Chen et al. 2019; Cao et al. 2020) and foreground-foreground imbalance (Peng et al. 2020; Oksuz et al. 2020). As presented in Figure 1(a), background instances are predominant in the training targets (background instances account for 90% of all training instances) when the foreground-background imbalance exists in detection data. This problem is compounded in pseudo-labels (foreground instances only account for 5% of all training instances).

Apart from foreground-background imbalance, we find a severe foreground-foreground imbalance problem in SSOD. As shown in Figure 1(b), there are some neglected classes in pseudo-labels in 1% COCO-standard. The cause of this balance is two folds. First, pseudo-labels in SSOD are inaccurate due to the high complexity of the detection task. Moreover, the model trained on foreground-background imbalanced labeled data is prone to generate biased predictions.

The above two-type imbalance yields biased pseudo-labels in self-training-based SSOD. Subsequent training on biased pseudo-labels further intensifies the class imbalance, thereby aggravating the performance of the final model.

To address the preceding issues, an intuitive idea is to utilize data-rebalancing algorithms in classification tasks (Pang et al. 2019; Ouyang et al. 2016; Ren et al. 2015). However, this idea is impeded by entanglements of foreground-background and foreground-foreground instances in detection data. To decouple these entanglements, we introduce a novel memory module called *CropBank* to store ground truths/pseudo labels of foreground instances in labeled/unlabeled data. With the CropBank, we propose two detection-specific data-rebalancing algorithms: foreground-background rebalancing (FBR) and adaptive foreground-foreground rebalancing (AFFR). We extend the original self-training paradigm to adaptive class-rebalancing self-training (ACRST) based on FBR and AFFR.

We first propose FBR to address the foreground-background imbalance in SSOD. FBR first extracts foreground instances from entire datasets according to ground truths/pseudo labels stored in the CropBank. Thereafter, foreground instances are augmented and pasted to random
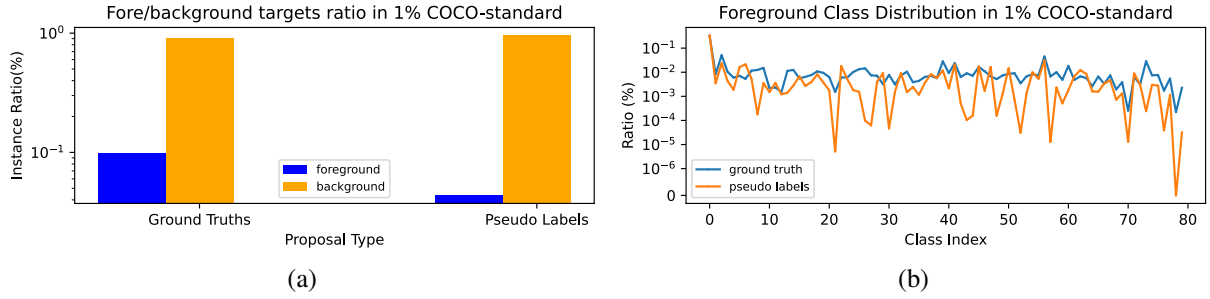
Figure 1: Class imbalance in pseudo-labels in 1% COCO-standard. (a) Foreground-background imbalance. In Faster-RCNN, background instances are predominant in the training targets of pseudo-labels. (b) Foreground-foreground imbalance is compounded in pseudo-labels.

locations in training images. With synthetic data, FBR can increase the proportion of foreground instances in training targets and alleviate the foreground-background imbalance.

For foreground-foreground imbalance, we propose adaptive foreground-foreground rebalancing(AFFR) based on FBR. In particular, we design a novel criterion called *pseudo recall* to judge whether a class is neglected or over-focused in SSOD. Thereafter, pseudo-labels of the neglected classes are sampled more frequently because of higher negative confidence. Consequently, the entire dataset is foreground-foreground rebalanced, thereby leading to a minimally biased detector for online pseudo-labeling in the subsequent self-training.

However, as presented in Figure 2, the accuracy of pseudo-labels is undesirable. We observe that FBR and AFFR suffer from noisy pseudo-labels in the mutual-training stage. Therefore, we exploit additional high-level semantics to filter noisy pseudo-labels. In particular, we propose a semi-supervised multi-label classification module to generate image-level pseudo-labels for unlabeled data. Thereafter, we design a two-stage filter mechanism to filter out pseudo-labels that activates negative in classification confidences or image-level pseudo-labels.

Our method outperforms previous state-of-the-art methods 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 state-of-the-arts. When using VOC07 (Everingham et al. 2010) as labeled data, our method outperforms state-of-the-arts by 1.43 mAP improvement.

We summarize our contributions as follows:

- We design a novel memory module called CropBank to disentangle fore/background and fore/foreground instances in detection data. With the CropBank, we further propose adaptive class-rebalancing self-training (ACRST) to address the foreground-background and foreground-foreground imbalance in SSOD.

- We propose a semi-supervised multi-label classification module to mine high-level semantics from unlabeled data. Thereafter, we propose a two-stage pseudo-label filtering mechanism with classification confidence and high-level semantics. This mechanism is effective in pseudo-label

denoising, thereby further facilitating FBR and AFFR.

- The proposed data-rebalancing and pseudo-label filtering algorithms are plug-and-play for any self-training-based SSOD framework. Moreover, the CropBank provides an effective detection-specific data augmentation algorithm.
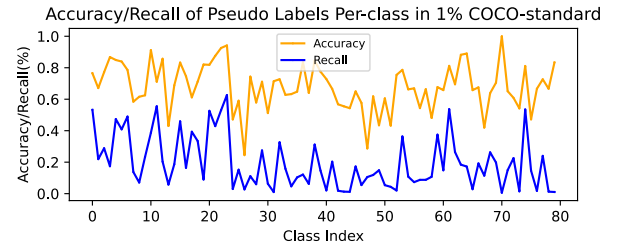

Figure 2: **Accuracy and Recall of pseudo-labels for each class in 1% COCO-standard.**

Related work

Semi-supervised Learning

Semi-supervised learning (SSL) utilizes unlabeled data to facilitate model training when large-scale annotated datasets are unavailable. The majority of the SSL methods typically consist of two types: consistency regularization and pseudo-labeling. Consistency regularization (Berthelot et al. 2019b,a; Xie et al. 2020a; Takeru et al. 2018; Sajjadi, Javanmardi, and Tasdizen 2016) enforces the prediction consistency of different augmented views of the same image. Pseudo-labeling (Tarvainen and Valpola 2017; Bachman, Alsharif, and Precup 2014; Arazo et al. 2019; Iscen et al. 2019) exploits high-quality pseudo-labels of unlabeled data to refine the model pre-trained on few labeled data.SSL has made remarkable success in image classification. However, foreground-background imbalance and foreground-foreground in SSOD heavily impede the effectiveness of the current SSL approaches. To address these limitations, we propose a pseudo-labeling-based method to alleviate the class imbalance.

Semi-supervised Object Detection

Object detection is a fundamental task in computer vision. Existing object detection frameworks include two- and one-stage detectors. 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 thereafter for the final bounding-boxes classification and regression. For one-stage detectors (Redmon et al. 2016; Lin et al. 2017; Law and Deng 2018; Duan et al. 2019), their predictions are performed on dense grids directly. Although existing studies have made remarkable progress in the past years, they have primarily focused on detectors in a fully supervised setting. SSOD algorithms, which train detectors with a combination of labeled and unlabeled data following standard SSL settings, have increasing attention recently. CSD (Jeong et al. 2019) utilizes a consistency-based mechanism, which enforces the model predictions consistency of different flipped versions of the same image for generalized feature learning. Similar to CSD, ISD (Jeong et al. 2020) imposes consistency-regularization on input images and their mixed versions. STAC (Sohn et al. 2020) introduces a pseudo-labeling-based method, which first pre-trains a detector on available labeled data and generates pseudo-labels on unlabeled data thereafter to re-train the detector. Instant Teaching (Zhou et al. 2021) develops an SSOD framework with MixUp (Zhang et al. 2018) and Mosaic (Bochkovskiy, Wang, and Liao 2020). Although these studies have improved the performance against the model in a supervised setting, they lack considerations into serious data imbalance issues in SSOD. Recently, Unbiased-Teacher (Liu et al. 2021) has been proposed recently to utilize focal loss (Lin et al. 2017) to alleviate class imbalance. However, the effectiveness of focal loss on unlabeled data is impeded by noisy pseudo-labels. Moreover, Unbiased-Teacher does not facilitate an increase in data diversity. To address the preceding issues and enhance the performance of SSOD, we propose FBR and AFFR to alleviate foreground-background and foreground-foreground imbalance simultaneously. We also devise a two-stage pseudo-label filtering algorithm with classification confidence and high-level semantics.

Method

This section describes our solutions in detail. Section defines the problem, while Section introduces the baseline framework Mean Teacher. Section defines the CropBank, and Section introduces the proposed adaptive class self-training rebalancing (ACRST) algorithm. Section illustrates the two-stage pseudo-label filtering method. Lastly, Section introduces the selective supervision mechanism. The overview of the entire training process is shown in Figure 3.

Problem Definition

Semi-supervised object detection aims to train detectors in a semi-supervised setting, where a small labeled dataset $D_s = \{x_i^s, y_i^s\}_{i=1}^{N_s}$ and a large unlabeled dataset $D_u = \{x_i^u\}_{i=1}^{N_u}$ are available. N_s/N_u presents the number of labeled/unlabeled data. y_i^s contains bounding-box annotations including object locations, sizes, and categories in ith labeled image x_i^s.

Mean Teacher for Semi-supervised Object Detection

This study takes the Mean Teacher (Tarvainen and Valpola 2017) as the SSOD baseline. Mean Teacher consists of a teacher and a student model, in which the entire framework is optimized via a mutual learning mechanism. The training pipeline of Mean Teacher consists of two stages.

Pre-training. The student model is pre-trained with a small amount of labeled data D_s via gradient back-propagation in a supervised manner. Thereafter, we initialize the teacher model with pre-trained model weights of the student model, which produce noisy-less pseudo-labels, thereby facilitating the subsequent training.

Teacher-Student Mutual Learning. In the mutual learning stage, we train the student model with supervision signals consisting of ground truths and pseudo-labels. Once the student model is updated via the gradient back-propagation, the learned knowledge is feedback to the teacher model in an exponential moving average (EMA) mechanism,

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

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

where θ_s/θ_t represents the model parameters of the student/teacher model, and $\mathcal{L}$ represents the total SSOD losses.

By using Faster-RCNN(Ren et al. 2015) as the detection module, the loss function of SSOD can be summarized as a combination of losses on labeled data $\mathcal{L}_{sup}$ and unlabeled data $\mathcal{L}_{unsup}$.

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

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

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

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

To succeed in SSOD, the teacher model must generate accurate pseudo-labels and maintain a reliable performance margin over the student model throughout the training. However, we observe that class imbalance in SSOD significantly hinders the performance of the teacher model.

CropBank

Data-rebalancing algorithms have been proved to be the most simple and effective data-rebalancing method in classification tasks. However, their effectiveness is heavily
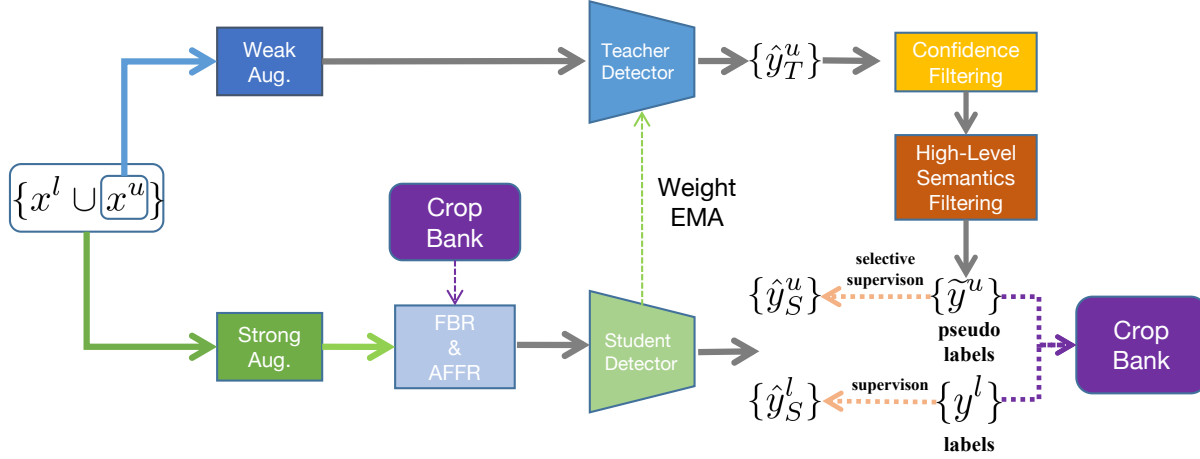
Figure 3: **Overview of our methods**. We take Mean Teacher as our SSOD baseline. In Mean Teacher, the teacher model generates pseudo labels from weakly augmented unlabeled data, and the student model is trained with a combination of ground-truths and pseudo labels. To alleviate the class imbalance in SSOD, we first design a memory module called CropBank which absorbs instance-level annotations including ground truths and pseudo labels. Thereafter, we utilize the CropBank to perform foreground-background rebalancing (FBR) and adaptive foreground-foreground rebalancing (AFFR) on strong augmented unlabeled data for adaptive class-rebalancing self-training (ACRST). To further improve ACRST, we introduce a two-stage pseudo-label filtering algorithm with classification confidence and high-level semantics.

impeded by strong interconnections on both foreground-background and foreground-foreground instances. To separate the entanglement, we propose a novel memory module called CropBank, which stores abundant instance-level annotations. The CropBank consists of two sub-banks, namely, Labeled CropBank $\Phi_L = \{y_i^l\}_{i=1}^{N_L}$ and Pseudo CropBank $\Phi_U = \{\widetilde{y}_i^u\}_{i=1}^{N_U}$, where N_L/N_U represent the size of Labeled/Pseudo CropBank, $y_i^l/\widetilde{y}_i^u$ represents ground truths/pseudo-labels of ith labeled/unlabeled image.

In the implementation, the CropBank size is unlimited owing to the negligible memory consumption of instance-level annotations. In the training stage, Φ_L is fixed once generated, while Φ_U is updated periodically with improved pseudo-labels in mutual training. We use CropBank as the basis to decouple instances and design adaptive class-rebalancing self-training (ACRST) to address the class imbalance.
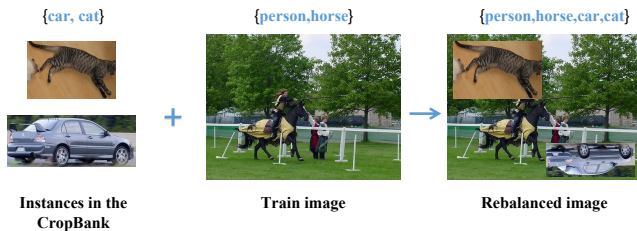

Figure 4: **Data-rebalancing with instances in the Crop-Bank.**

Adaptive Class-Rebalancing Self-Training

The self-training paradigm is an ideal solution to alleviate the lack of labeled data. However, its effectiveness is impeded by the inherent class imbalance in object detection tasks. Therefore, we propose ACRST to alleviate the class imbalance in SSOD. ACRST consists of foreground-background rebalancing (FBR) and adaptive foreground–foreground rebalancing (AFFR).

Foreground-Background Rebalancing

Models trained on foreground-background imbalanced data tend to overfit excessive background instances. Foreground-background imbalance in object detection has been widely explored. Various solutions have been proposed to alleviate such an imbalance, including loss-reweighting (Lin et al. 2017) and region refinement (Ren et al. 2015). Unfortunately, these methods rely on ground truths unavailable in SSOD to guide the rebalancing procedure. Hence, we utilize abundant instance-level annotations including ground truths and pseudo-labels in the CropBank to perform foreground-background rebalancing.

Given a training mini-batch $B = \{x_i, y_i\}_{i=1}^{N_T}$, we fetch a set of foreground instances $F = \{c_j, y_j\}_{j=1}^{N_C}$ from the Crop-Bank Φ_L and Φ_U for each image x_i following a sampling distribution P, where c_j is a foreground instance cropped from original images with annotation y_j. Thereafter, a new training mini-batch $B_{mix} = \{x_i^{mix}, y_i^{mix}\}_{i=1}^{N_T}$ is generated as follows:

$$x_i^{mix} = \alpha x_i + (1 - \alpha)c_j, \qquad (6)$$
$$y_i^{mix} = merge(y_i, y_j). \qquad (7)$$

Where α denotes a binary mask of pasted objects and y_i^{mix} denotes mixed annotations, in which fully occluded instances are removed from mixed image x_i^{mix}. In detail, c_j

is a rectangular region cropped from the image based on instance-level annotations in the CropBank. During training, c_j is augmented and pasted thereafter to random locations of x_i. This combined procedure increases the ratio of foreground instances in the training data for foreground-background rebalancing and also explores essential context semantics from a holistic perspective.

Once mixed images are ready, we take them to train the detector as the pipeline of Mean Teacher. The rebalancing process is shown in Figure 4. As discussed in Section , such a crop-and-paste operation enables higher model performance with selective supervision.

Adaptive Foreground-Foreground Rebalancing

FBR adequately alleviates the foreground-background imbalance with considerable attention on foreground instances. However, a random sampling distribution P, such as uniform distribution, fails to correct the foreground-foreground imbalance. Hence, we propose an adaptive sampling probability distribution P for foreground-foreground rebalancing. In particular, samples in neglected classes are selected more frequently during the training.

To measure the neglected degree of a class, we propose a novel criterion *pseudo recall* (PR), which quantities the proportion of pseudo-labels to ground truths. In detail, we estimate the class distribution of unlabeled data from labeled data on account of distribution similarity between labeled and unlabeled data. Suppose there are K classes $\{1, 2, ..K\}$ in datasets. We calculate *pseudo recall* for class k as

$$PR_k = \frac{N_k^u}{rN_k^l}, \tag{8}$$

where N_k^u and N_k^l denote the number of pseudo-labels and ground truths of class k, and r is the ratio of the unlabeled to labeled data.

Pseudo recall defines how neglected one class is under the SSOD setting. High PR_k indicates that the detector is certain even overconfident on class k. Consequently, lower sampling probabilities should be allocated to samples in class k for overfitting alleviation. By contrast, low PR_k implies that the detector lacks confidence for detecting instances of class k. Therefore, we should select these instances frequently. As a solution, we sort the classes in descending order according to *pseudo recall* and design the following adaptively sampling strategy:

$$\mu_k = \left(\frac{PR_{K-k+1}}{\Sigma_{i=1}^K PR_i} \right)^\beta, \tag{9}$$

where μ_k is the probability of selecting instances of class k, and β is used to tune the sampling probability. This mechanism adaptively allocates higher/lower sampling rates to neglected/over-focused instances. Note that AFFR performs FBR simultaneously. There are numerous ways to rebalance class distribution, and we introduce an effective example. A potential problem with this mechanism is that the noise of pseudo-labels in the neglected classes is amplified. Therefore, we propose a two-stage pseudo-label filtering mechanism in Section .

Two-stage Pseudo-label Filtering

The 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 undesirably amplified in self-training. Consequently, we should filter noisy pseudo-labels from the teacher model predictions and store noisy-less pseudo-labels in the CropBank. In SSOD, the general filtering algorithm sets a threshold τ_{cls} to filter predictions with low classification confidence out. However, such single-stage filtering without additional semantics constraints is prone to produce noisy pseudo-labels. As a solution, we propose a semi-supervised multi-label classification module to learn high-level semantics (i.e., image-level pseudo-labels). Thereafter, we design a two-stage filtering algorithm with classification confidences and high-level semantics to generate accurate pseudo-labels.

Semi-supervised Multi-label Classification.

The proposed semi-supervised multi-label classification module is devised based on Mean Teacher for the classification task. For each image x_i, we particularly aim to predict its image-level pseudo-labels $v_i = \{l_k\}_{k=1}^K, l_k \in \{0, 1\}$, where K is the total category number and l_k determines 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 which supervise the student model. We utilize a focal–binary–cross-entropy loss to optimize the student model.

Two-stage Pseudo-label Filtering.

For bounding-box prediction i with classification score s_i of image x_j with image-level pseudo-label y_j, we perform a two-stage filtering to get bounding-box pseudo-labels with low-noise. In the first stage, we filter predictions out with scores $s < \tau_{cls}$ to remove predictions with low objectness or wrong class labels. In the second stage, predictions whose classes activate negative in y_j (i.e., activation values are smaller than τ_{ml}) are removed. The second filtering stage utilizes high-level semantics to filter noisy predictions inconsistent with image-level pseudo-labels. With the two-stage filtering, the consistency of low- and high-level semantics are achieved.

Selective Supervision

In previous SSOD research, bounding-box regression losses are removed during training to alleviate noise. By contrast, utilizing regression losses in our framework is beneficial to achieve high SSOD performance, which is attributed to the CropBank module.

The contribution is two-fold. First, the CropBank alleviates noise from partially detected instances, which take a large proportion in bias predictions. Learning blindly with these noisy pseudo-labels will heavily aggravate the model performance. However, when the partially detected bounding-boxes from the CropBank are cropped and pasted to training batches, they become independent and complete in the new background, thereby providing additional clean training supervisions. Second, the CropBank provides

a detection-specific data augmentation method. The additional augmented data continuously improves the regression accuracy of pseudo-labels.

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

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

where $\widetilde{y}_i^{ss}$ denotes instances from the CropBank in x_i^u.

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 set as the labeled dataset and take the remaining data as the unlabeled dataset.

2. *COCO-additional*: We use the COCO2017-train set as the labeled dataset and the additional COCO2017-unlabeled set as the unlabeled dataset.

3. *VOC07&12*: We use the VOC07-trainval set as the labeled dataset and the VOC12-trainval set as the unlabeled dataset.

We evaluate the model performance on the COCO2017-val set for (1)(2) and VOC07-test set for (3).

Implementation Details

We use FPN-Faster-RCNN with ResNet-50 backbone as the detection module. ResNet-50 is initialized with ImageNet pre-trained weights. We set the hyper-parameters $\lambda_{unsup} = 2$, $\beta = 2$. For two-stage pseudo-label filtering, we use classification confidence threshold $\tau_{cls} = 0.7$ and multi-label confidence threshold $\tau_{ml} = 0.2$. We use $AP_{50:95}$, i.e, mAP as the evaluation metric. We construct each training batch with 32 labeled and 32 unlabeled images for all the training settings. For the *COCO-standard*, the pre-training stage takes 2500/5000/10000/20000/40000 steps for 0.5/1/2/5/10% *COCO-standard* and 180000 steps for the whole training stage of *COCO-standard*. For *COCO-additional*, the pre-training stage takes 90000 steps in total 270000 training steps. For *VOC07&12*, the pre-training stage takes 12000 steps and 36000 steps for the entire training stage. Strong augmentations in our research consist of random jittering, Gaussian noise, and random crop. Weak augmentations in our study consist of random resize and flip. Moreover, we set the above hyper-parameters without aggressively searching. Consequently, high model performance may be achieved with improved choices.

Results and Comparisons

COCO-standard

We first evaluate the efficacy of our method on COCO-standard. As shown in Table 1, when only 0.5% to 10% of

the entire dataset are labeled, our model consistently performs better against all previous studies in CSD, STAC, Instant Teaching, and Unbiased Teacher. When trained on the 1% COCO-standard, our method achieves 5.32 mAP improvement compared Unbiased-Teacher. The mAP is even higher than CSD trained on 10% COCO-standard. When trained on 10% COCO-standard, our method achieves 10.42 mAP improvement compared with supervised baselines. We attribute the success of model performance to two factors.

Class rebalanced data. Our method alleviates the class imbalance in SSOD with two rebalancing algorithms (i.e., FBR and AFFR). Foreground-background rebalanced data prevents the model from overfitting on background instances and helps mine beneficial information from enormous unlabeled data. Foreground-foreground rebalanced data benefits the model predictions with information from neglected classes and avoids biased predictions on over-focused classes.

Noise-less pseudo-labels. When using the self-training paradigm for SSOD, accurate and reliable pseudo-labels from pre-trained models should be generated. We propose a teacher–student mutual learning mechanism for progressive pseudo-labels refinement. In addition, we introduce a two-stage pseudo-label filtering algorithm to remove noisy predictions with classification confidence and high-level semantics. With accurate pseudo-labels, the student model is well optimized and gives beneficial feedback to the teacher model. We present an ablation study on two-stage pseudo-label filtering in Section .

COCO-additional

In this section, we verify whether our method can further improve the model trained on a large-scale labeled dataset with additional unlabeled data. Table 2 shows that our model has a 0.41 mAP improvement compared with those of previous methods, and 1.51 mAP improvement compared with the supervised baseline. This result indicates that our method achieves satisfying improvement even on the well-trained model.

VOC07&12

We evaluate models on a less imbalanced dataset VOC07&12 to demonstrate the generalization of our method. Table 3 provides the mAP results of CSD, STAC, Unbiased Teacher, and our method. Our method achieves 7.99 mAP improvement compared with the supervised setting, and 1.43 mAP improvement against previous state-of-the-art methods, even though Unbiased Teacher has witnessed performance saturation in VOC07&12. We owe the success of our approach to the generalization ability of ACRST. Even if training data is foreground-foreground balanced, FBR can substantially alleviate the inevitable foreground-background imbalance in SSOD. Furthermore, the two-stage pseudo-label filtering mechanism benefits the model trained on VOC07&12.

Ablation Study

Foreground–Background Rebalancing (FBR)

We first verify the effect of FBR. Table 4 shows that applying FBR improves mAP in 1% labeled COCO from

Table 1: Experimental results on COCO-standard comparing with CSD, STAC and Unbiased Teacher.

	COCO-standard ($AP_{50:95}$)				
	0.5%	1%	2%	5%	10%
Supervised	6.83 ± 0.15	9.05 ± 0.16	12.70 ± 0.15	18.47 ± 0.22	23.86 ± 0.81
CSD(Jeong et al. 2019)	7.41 ± 0.21	10.51 ± 0.06	13.93 ± 0.12	18.63 ± 0.07	22.46 ± 0.08
STAC(Sohn et al. 2020)	9.78 ± 0.53	13.97 ± 0.35	18.25 ± 0.25	24.38 ± 0.12	28.64 ± 0.21
Instant Teaching(Zhou et al. 2021)	-	18.05 ± 0.15	22.45 ± 0.15	26.75 ± 0.05	30.40 ± 0.05
Unbiased Teacher(Liu et al. 2021)	16.94 ± 0.23	20.75 ± 0.12	24.30 ± 0.07	28.27 ± 0.11	31.5 ± 0.10
Ours	$\mathbf{19.62 \pm 0.37}$	$\mathbf{26.07 \pm 0.46}$	$\mathbf{28.69 \pm 0.17}$	$\mathbf{31.35 \pm 0.13}$	$\mathbf{34.92 \pm 0.22}$

Table 2: Experimental results on COCO-additional with CSD, STAC, and Unbiased Teacher. Note that N×represents N×90K training iterations.

	COCO-additional ($AP_{50:95}$)				
	Supervised(3×)	CSD(3×)	STAC(6×)	Unbiased Teacher(3×)	Ours(3×)
$AP_{50:95}$	40.20	38.82	39.21	41.30	**42.79**

Table 3: Experimental results on VOC07&12 comparing with CSD, STAC and Unbiased Teacher.

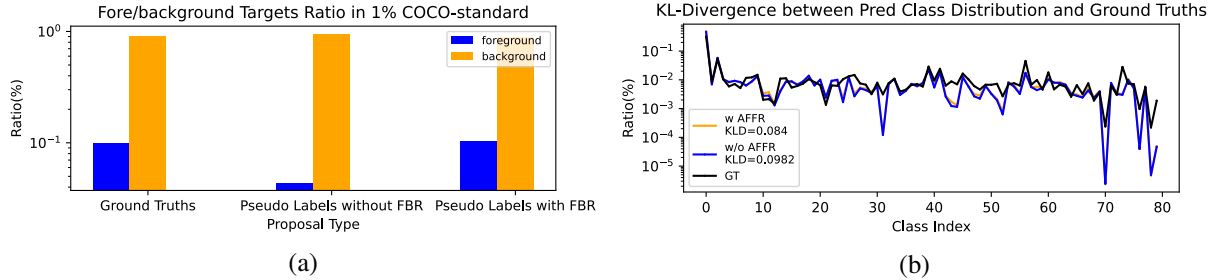|	Labeled	Unlabeled	AP_{50}	$AP_{50:95}$
Supervised	VOC07	None	72.63	42.13
CSD(Jeong et al. 2019)	VOC07	VOC12	74.70	-
STAC(Sohn et al. 2020)	VOC07	VOC12	77.45	44.64
Unbiased Teacher(Liu et al. 2021)	VOC07	VOC12	77.37	48.69
Ours	VOC07	VOC12	**78.16**	**50.12**

(a)　　　　　　　　　　　　　　　(b)

Figure 5: Ablation study on the FBR (a) and AFFR (b). (a) FBR alleviates foreground-background imbalance in pseudo-labels in 1% COCO-standard. (b) AFFR reduces the KL-Divergence(KLD) between the pseudo-labels distribution and the ground truths distribution from 0.0982 to 0.084 in 1% COCO-standard.
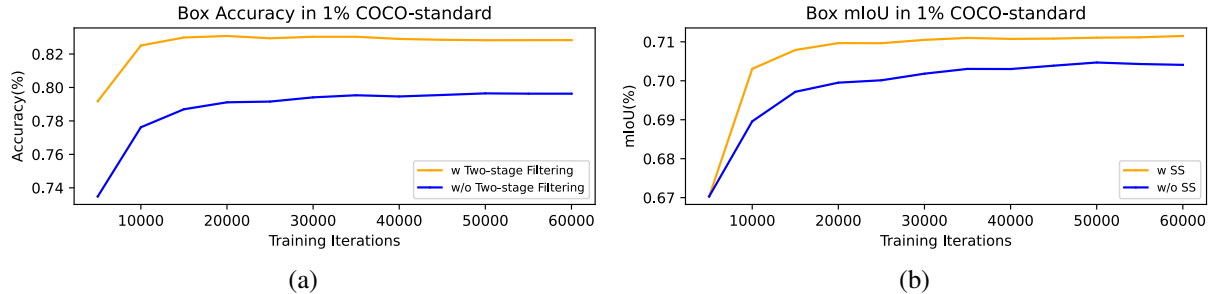

(a)　　　　　　　　　　　　　　　(b)

Figure 6: Pseudo-labels improvement on Box Accuracy and Box mIoU in 1% COCO-standard. (a) Box accuracy of pseudo-labels with/without two-stage pseudo-label filtering. (b) Box mIoU between pseudo-labels and ground truths with/without selective supervision (SS).

Table 4: Ablation study in 1% *COCO-standard*.

FBR	AFFR	Two-Stage	SS	$AP_{50:95}$
				20.75
		✓		23.48
✓				23.32
✓	✓			24.17
✓	✓	✓		25.56
✓	✓	✓	✓	**26.12**

20.75 to 23.32. To analyze the divergent results, we visualize the foreground-background distribution of the rebalanced pseudo-labels. Figure 5(a) shows that after rebalancing, the distribution of the foreground and background instances is rebalanced. The ratio of foreground instances in rebalanced pseudo-labels is even higher than that of ground truths. Therefore, training detectors with rebalanced training data alleviates data bias and produces high mAP.

Adaptive Foreground–Foreground Rebalancing (AFFR)
Table 4 shows that AFFR improves mAP from 23.32 to 24.17 based on FBR. We verify the effectiveness of AFFR by analyzing the foreground class distribution. Figure 5(b) presents that AFFR alleviates foreground–foreground imbalance and reduces KL-divergence from 0.0982 to 0.084. This result confirms the effectiveness of AFFR in handling foreground–foreground imbalance issues in pseudo-labels. Using AFFR can generate an unbiased training data distribution and results in a higher mAP.

Two-stage Pseudo-label Filtering
We also verify the effectiveness of the two-stage pseudo-label filtering with classification confidences and high-level semantics. As presented in Table 4, the model that filters pseudo-labels with high-level semantics can favorably improve model performance against the model with only classification confidence. The two-stage filtering mechanism utilizes an uncertainty mechanism where only predictions with high objectness and follow image-level constraints are regarded as accurate pseudo-labels. Figure 6(a) shows that the two-stage filtering mechanism has a continuous improvement on the accuracy of pseudo-labels. This confirms the effectiveness of the two-stage filtering mechanism in removing noisy predictions in SSOD. Moreover, a two-stage filtering mechanism is necessary to build a clean Pseudo CropBank and further improve the performances of ACRST. Table 4 indicates that applying the two-stage filtering mechanism improves the mAP from 20.75 to 23.48. Moreover, applying the mechanism further improves the mAP of the model trained with ACRST from 24.17 to 25.56. This result confirms that the two-stage filtering mechanism is effective in handling the noisy pseudo labels.

Selective Supervision (SS)
Lastly, 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-and-paste operation in ACRST, in which incomplete instances are pasted to a new background in the training data. As a result, utilizing these instances as targets of regression optimization produces less noise compared to regress them directly in originating images. We further analyze the distribution of regression accuracy of pseudo-labels. Figure 6(b) shows that selective supervision improves the mIoU of pseudo-labels. Accordingly, transferring these incomplete predictions to complete objects in a new background alleviates noise in the regression targets and improves the model performance. Selective supervision is still under exploration. For example, the current strategy fails to handle noise in which objects are overlapped with each other.

Discussions

CropBank
Data augmentations that generate different views of the same image are necessary to pseudo-labeling-based and consistency-regularization-based semi-supervised learning. Conventional data augmentations consist of color jittering (Berthelot et al. 2019a), rotation (Jeong et al. 2019), and Gaussian Noise (Chen et al. 2020). Although these methods are effective in vision tasks, they fail to change the image semantics. MixUp (Zhang et al. 2018) and Mosaic (Bochkovskiy, Wang, and Liao 2020) are proposed to change the image semantics by mixing images. However, they fail to decouple instances semantics in detection data.

The CropBank provides a detection-specific data augmentation that effectively decouples entangled semantics in images. The idea of the CropBank is inspired by the Cut-Mix (Yun et al. 2019).

The difference between the CropBank and the CutMix is two-fold. First, the CropBank decouples instances in detection data and creates training new detection datasets with complex disentangled semantics, while the CutMix is classification-specific and unable to decouple semantics. Second, the CropBank adaptively injects semantics from the entire dataset to training images, but the CutMix only exchanges image-to-image semantics. Table 5 provides the model performance with different data augmentations in SSOD. The CropBank improves $AP_{50:95}$ from 16.00 to 16.85 compared to MixUp and Mosaic in (Zhou et al. 2021).

Moreover, the CropBank can be easily embedded in semantic and instance segmentation with few modifications. We will release related researches in future studies.

Table 5: STAC(Sohn et al. 2020) performance in 1% COCO-standard under different data augmentations.

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

Image-level Pseudo-labels
There is an unexplored question in our study: Why we utilize high-level semantics (i.e., image-level pseudo-labels) to fil-

Table 6: Model Performance, Accuracy and Recall of pseudo-labels in 1% COCO-standard in three settings.

Setting	Accuracy	Recall	$AP_{50:95}$
One-stage	0.735	0.299	20.75
Two-stage Filtering	**0.792**	0.292	**23.48**
Two-stage Mining	0.722	**0.376**	21.56

ter noisy pseudo-labels instead of mining neglected predictions? To answer this question, we first evaluate the model trained in three different settings. (1) **One-stage:** Predictions with low classification confidence are filtered. (2) **Two-stage filtering:** Predictions with low classification confidence or low activation in image-level pseudo-labels are filtered. (3) **Two-stage Mining:** Predictions with high classification confidence or high activation in image-level pseudo-labels are reserved. Thereafter, we calculate the $AP_{50:95}$, accuracy, and recall of pseudo-labels.

Table 6 indicates that the two-stage mining mechanism improves recall from 0.299 to 0.376 and reduces accuracy from 0.735 to 0.722, compared to the one-stage strategy. The two-stage filtering mechanism improves the accuracy from 0.735 to 0.792 and reduces the recall from 0.299 to 0.292 compared to the one-stage filtering. Although the recall improvement of the two-stage mining is higher than the accuracy improvement of the two-stage filtering, the $AP_{50:95}$ improvement of the latter is 2.73, which is higher compared to the former.

This result indicates that a higher accuracy of pseudo-labels is considerably more important than a higher recall. Hence, future studies should carefully consider the balance between accuracy and recall. We will further analyze this problem in future studies.

Conclusion

This study proposes ACRST based on a novel memory module called CropBank to address the class imbalance in SSOD. ACRST considerably alleviates foreground-background and foreground-foreground imbalance with proposed FBR and AFFR. To further improve FBR and AFFR, we design a two-stage pseudo-label filtering algorithm with classification confidence and high-level semantics. Over iterations on rebalanced training data, SSOD detectors become unbiased and ameliorate the model performance progressively. Extensive experiments on benchmarks demonstrate the effectiveness of our method.

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. URL http://arxiv.org/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. URL https://proceedings.neurips.cc/paper/2014/hash/66be31e4c40d676991f2405aaecc6934-Abstract.html.

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. URL http://arxiv.org/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. URL https://proceedings.neurips.cc/paper/2019/hash/1cd138d0499a68f4bb72bee04bbec2d7-Abstract.html.

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

Cao, Y.; Chen, K.; Loy, C. C.; and Lin, D. 2020. Prime Sample Attention in Object Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 11580–11588. IEEE. doi:10.1109/CVPR42600.2020.01160. URL https://doi.org/10.1109/CVPR42600.2020.01160.

Chen, K.; Li, J.; Lin, W.; See, J.; Wang, J.; Duan, L.; Chen, Z.; He, C.; and Zou, J. 2019. Towards Accurate One-Stage Object Detection With AP-Loss. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 5119–5127. Computer Vision Foundation / IEEE. doi:10.1109/CVPR.2019.00526. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Chen_Towards_Accurate_One-Stage_Object_Detection_With_AP-Loss_CVPR_2019_paper.html.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. E. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, 1597–1607. PMLR. URL http://proceedings.mlr.press/v119/chen20j.html.

Dollar, P.; Wojek, C.; Schiele, B.; and Perona, P. 2012. Pedestrian Detection: An Evaluation of the State of the Art. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(4): 743–761. doi:10.1109/TPAMI.2011.155.

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. doi:10.1109/ICCV.2019.00667. URL https://doi.org/10.1109/ICCV.2019.00667.

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. doi:10.1109/ICCV.2015.169. URL https://doi.org/10.1109/ICCV.2015.169.

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. doi:10.1109/CVPR.2014.81. URL https://doi.org/10.1109/CVPR.2014.81.

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. doi:10.1109/ICCV.2017.322. URL https://doi.org/10.1109/ICCV.2017.322.

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. doi:10.1109/CVPR.2019.00521. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Iscen_Label_Propagation_for_Deep_Semi-Supervised_Learning_CVPR_2019_paper.html.

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. URL https://proceedings.neurips.cc/paper/2019/hash/d0f4dae80c3d0277922f8371d5827292-Abstract.html.

Jeong, J.; Verma, V.; Hyun, M.; Kannala, J.; and Kwak, N. 2020. Interpolation-based semi-supervised learning for object detection. *ArXiv* abs/2006.02158.

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. URL https://openreview.net/forum?id=BJ6oOfqge.

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. doi:10.1109/ICCV.2017.324. URL https://doi.org/10.1109/ICCV.2017.324.

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.-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. *ArXiv* abs/2102.09480.

Oksuz, K.; Cam, B. C.; Kalkan, S.; and Akbas, E. 2020. Imbalance Problems in Object Detection: A Review. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 1–1.

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. doi:10.1109/CVPR.2016.100. URL https://doi.org/10.1109/CVPR.2016.100.

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. doi:10.1109/CVPR.2019.00091. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Pang_Libra_R-CNN_Towards_Balanced_Learning_for_Object_Detection_CVPR_2019_paper.html.

Peng, J.; Bu, X.; Sun, M.; Zhang, Z.; Tan, T.; and Yan, J. 2020. Large-Scale Object Detection in the Wild From Imbalanced Multi-Labels. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 9706–9715. IEEE. doi:10.1109/CVPR42600.2020.00973. URL https://doi.org/10.1109/CVPR42600.2020.00973.

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. doi:10.1109/CVPR.2016.91. URL https://doi.org/10.1109/CVPR.2016.91.

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. URL https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html.

Russakovsky, O.; Li, L.; and Fei-Fei, L. 2015. Best of both worlds: Human-machine collaboration for object an-

notation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2121–2131. doi:10.1109/CVPR.2015.7298824.

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. URL https://proceedings.neurips.cc/paper/2016/hash/30ef30b64204a3088a26bc2e6ecf7602-Abstract.html.

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.

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. URL https://proceedings.neurips.cc/paper/2017/hash/68053af2923e00204c3ca7c6a3150cf7-Abstract.html.

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*. URL https://proceedings.neurips.cc/paper/2020/hash/44feb0096faa8326192570788b38c1d1-Abstract.html.

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. doi:10.1109/CVPR42600.2020.01070. URL https://doi.org/10.1109/CVPR42600.2020.01070.

Yun, S.; Han, D.; Chun, S.; Oh, S. J.; Yoo, Y.; and Choe, J. 2019. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 6022–6031. IEEE. doi:10.1109/ICCV.2019.00612. URL https://doi.org/10.1109/ICCV.2019.00612.

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. URL https://openreview.net/forum?id=r1Ddp1-Rb.

Zhou, Q.; Yu, C.; Wang, Z.; Qian, Q.; and Li, H. 2021. Instant-Teaching: An End-to-End Semi-Supervised Object Detection Framework. *CoRR abs/2103.11402*. URL https://arxiv.org/abs/2103.11402.