Co-Student: Collaborating Strong and Weak Students for Sparsely Annotated Object Detection

Lianjun Wu[1], Jiangxiao Han[1], Zengqiang Zheng[2], and Xinggang Wang[1]✉

[1] School of EIC, Huazhong University of Science and Technology
[2] Wuhan Jingce Electronic Group Co., Ltd.

Abstract. Sparsely Annotated Object Detection (SAOD) tackles the issue of incomplete labeling in object detection. Compared with Fully Annotated Object Detection (FAOD), SAOD is more complicated and challenging. Unlabeled objects tend to provide wrong supervision to the detectors during training, resulting in inferior performance for prevalent object detectors. Shrinking the performance gap between SAOD and FAOD does contribute to reducing the labeling cost. Existing methods tend to exploit pseudo-labeling for unlabeled objects while suffering from two issues: (1) they fail to make full use of unlabeled objects mined from the student detector and (2) the pseudo-labels contain much noise. To tackle those two issues, we introduce *Co-Student*, a novel framework aiming to bridge the gap between SAOD and FAOD via fully exploiting the pseudo-labels from both teacher and student detectors. The proposed *Co-Student* comprises a sophisticated teacher to denoise the pseudo-labels for unlabeled objects and two collaborative students that leverage strong and weak augmentations to excavate pseudo-labels. The students exchange the denoised pseudo-labels and learn from each other with consistency regularization brought by strong-weak augmentations. Without bells and whistles, the proposed *Co-Student* framework with the one-stage detector, *i.e.*, FCOS, can achieve state-of-the-art performance on the COCO dataset with sparse annotations under diverse settings. Compared to previous works, it obtains 1.0%∼3.0% AP improvements under five settings of sparse annotations and achieves 95.1% performance compared to FCOS trained on fully annotated COCO dataset. Code has been made available at `https://github.com/hustvl/CoStudent`.

Keywords: Sparsely Annotated Object Detection · Strong and Weak Students · Pseudo-Labels · Denoising Teacher

1 Introduction

Object detection, as the fundamental task in computer vision, requires detectors to predict bounding boxes and categories of objects in images. Recently, with the rapid development of deep convolutional neural networks (CNN) [8–10, 24, 31],

✉ Corresponding author: Xinggang Wang (`xgwang@hust.edu.cn`)

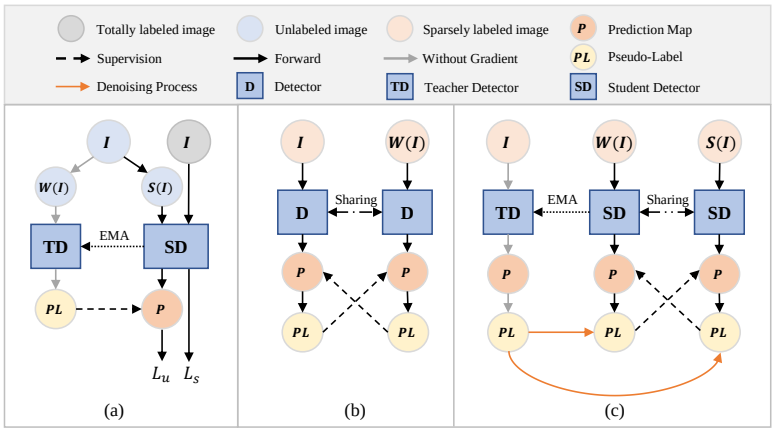


Fig. 1: Comparison of different frameworks for sparsely annotated object detection. (a) The vanilla teacher-student framework in SSOD [4, 17, 38, 43], such as Calibrated Teacher [30]; (b) the Siamese detector with pseudo-labeling, such as Co-mining [33]; (c) the proposed *Co-Student* with denoising teacher.

massive object detection methods [2,13,28,29] have emerged, which heavily rely on fully annotated bounding boxes. However, exhaustively labeling each image with bounding boxes and categories in various real-world scenarios is difficult and costly. Besides, it is also inevitable to miss labels during the human annotation process. The scarcity of annotations will affect the performance and effectiveness of object detectors.

To address the above issues, several works [12,23,30,33,38,42] explore SAOD which aims for high-performance detectors training with partial annotations in images, *i.e.* amounts of foreground objects in SAOD are unlabeled and treated as the background. Different from the semi-supervised object detection (SSOD) task which trains object detectors with both labeled and unlabeled images, the SAOD task attempts to train object detectors on images with partial annotations or incomplete annotations. The unlabeled foreground objects will be regarded as negative samples, which can lead to many ambiguities in optimization and greatly affect the training and performance of the detectors. Recently, several methods [23, 30, 33] have explored pseudo-labeling and consistency regularization on unlabeled objects for SAOD and obtained promising performance. As shown in Fig. 1, we categorize the previous methods into three groups, *i.e.*, (a) the vanilla teacher-student and (b) the Siamese detector with pseudo-labeling. Specifically, the vanilla teacher-student methods [30], inspired by the semi-supervised object detection [4, 17, 18, 38, 43] rely on pseudo-labels by the teacher detector while ignoring the good annotations from the student detector. The Siamese detector methods [1,6,36] are motivated to mine pseudo-labels from two Siamese detectors with weak augmentations while the noisy annotations affect the training. Despite the great success of those methods, the pseudo-labels

for unlabeled objects are still not fully excavated and utilized and the gap between SAOD and FAOD is still large.

In this paper, we are motivated to explore the multi-source pseudo-labels for unlabeled objects, *i.e.*, annotations from both the teacher detector and the student detector, and present *Co-Student*, as shown in Fig. 1(c). The proposed *Co-Student* is composed of a sophisticated teacher and two collaborative students. Inspired by previous SSOD methods [17, 22, 25, 35], we adopt Exponential Moving Average (EMA) to update the teacher detector, which can generate better pseudo-labels compared to the student in the Siamese detectors. In *Co-Student*, the teacher detector is used to denoise the pseudo labels of two collaborative student detectors, which can handle the mutual interference of the student detectors caused by the noise in their pseudo-labels, and also suppress the accumulation of noise in the pseudo-labels earlier during training. Furthermore, since the two collaborative student detectors exchange the denoised pseudo-labels, which contain many high-quality unlabeled objects treated as pseudo annotations, this further narrows the gap between SAOD and FAOD.

Our experiments have shown that our *Co-Student* with the denoising teacher based on FCOS [26] is highly effective for SAOD. We briefly summarize our contributions below:

- We present a novel SAOD framework named *Co-Student* to fully utilize pseudo annotations for unlabeled objects, in which the two students with strong and weak augmentation collaboratively excavate more pseudo annotations and exchange the supervisions.
- We introduce an EMA teacher detector to denoise the pseudo labels excavated by the students for correcting the guidance for unlabeled objects during training.
- The experimental results on the COCO dataset demonstrate the effectiveness and superiority of the proposed *Co-Student*, which remarkably outperforms previous methods and achieves 95.1% performance compared to the fully-supervised baselines.

2 Related Works

Semi-Supervised Object Detection (SSOD): Pseudo-label-based methods [17, 22, 25, 35] inherit the self-training strategy from Semi-Supervised Learning (SSL) [11, 21, 39], which trains a teacher model on fully annotated images. They use the teacher to generate pseudo-labels on unlabeled images and then use pseudo-labels to train the student model. [39] shows the key of this strategy is that the student model should be perturbed and the teacher should be undisturbed. *Unbiased Teacher* [17] demonstrates that class imbalance in ground-truth labels will make the model predict unevenly, which can be handled by an EMA teacher model. *Unbiased Mean Teacher* [4] demonstrates that the EMA teacher model is challenging to provide reliable and robust predictions for guiding the student model's learning across domains. De-biased Teacher [32] is designed to

de-bias the training proposals generated by the pseudo-label-based IoU matching to recall more labels in unlabeled images. Consistent-Teacher [34] found the inconsistency of pseudo-labels in SSOD and proposed an adaptive anchor assignment strategy and 3D feature alignment module to remedy the unreliable supervision signal during training. DenseTeahcer [43] replaced the sparse pseudo-boxes with the Dense Pseudo-Label (DPL) as a straightforward form of pseudo-label, which does not involve any post-processing method and retains richer information.

Two-Stage Detectors for SAOD: Two-stage object detectors [3,7,19] classify and regress class-agnostic proposals to find foreground objects in the first stage, followed by using RoI heads to classify and regress by pooling into a fixed size. Soft Sampling [38] down-weights the gradients for each RoI in the second stage according to a function of its overlap with sparse annotations. However, this method will also reduce the gradient from real backgrounds, which also brings a negative effect on network learning. It has not made improvements at the framework level, leading to its inability to further enhance performance using pseudo-labels. SparseDet [23] adopts a dual-branch input and utilizes the RPN network to differentiate RoIs into foreground, background, or unlabeled regions. As for the unlabeled regions, it enforces the ROI features of the two branches to be consistent with each other in the same region. Although it adopts a dual-branch framework with consistency loss to utilize the unlabeled part information, it fails to explicitly utilize pseudo-labels as the positive supervisions for the model results in the unlabeled objects providing limited assistance in learning for the overall detector. Our method utilizes the Co-Student framework to generate pseudo-labels, with a teacher model to revise these pseudo-labels. The refined pseudo-labels as positive supervision enable the model to gain more knowledge from the unlabeled objects.

One-Stage Detectors for SAOD: One-stage object detectors [5,13,15,24,26, 29] densely predict classes and offsets without random sampling for label assignment. They usually handle unbalanced positives and negatives via re-weighting loss with FocalLoss [15]. However, samples from missing labeled areas are likely to be incorrectly up-weighted, leading to an increase in corresponding loss [42]. *Background Recalibration Loss* [42] redesigns FocalLoss to reduce the negative effect of incorrect supervision but does not fully leverage the valuable information extracted from unlabeled objects. It also has not made improvements at the framework level, leading to its inability to further enhance performance using pseudo-labels. Co-mining [33] introduces a dual-branch framework and merges the pseudo-labels generated by each branch with sparse annotations to positively supervise the other branch. However, our experimental results indicate that simple data augmentation alone is not sufficient to uncover enough unlabeled objects. Moreover, simply increasing the strength of data augmentation may backfire due to the issue of noise. Calibrated Teacher [30], which transforms the confidence score of the teacher to fit the real precision and leaves the threshold of generating pseudo-label unchanged on different detectors. Calibrated Teacher is based on the vanilla teacher-student framework and ignores

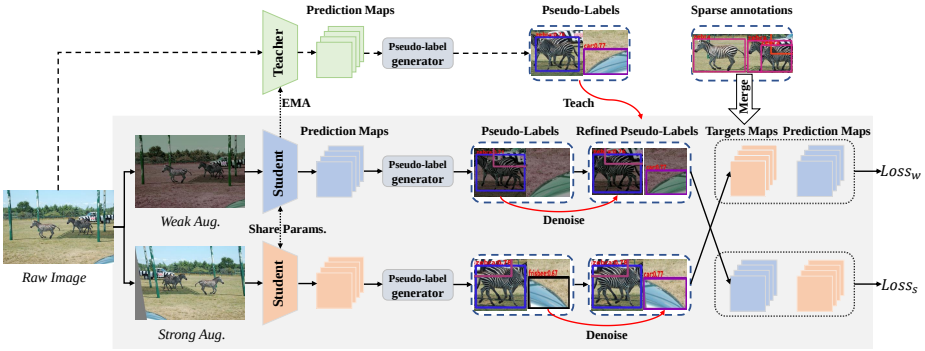


Fig. 2: The overview of our *Co-Student* is composed of a sophisticated teacher and two collaborative students that leverage strong and weak augmentations to excavate pseudo-labels. The students exchange the denoised pseudo-labels and learn from each other with consistency regularization brought by strong-weak augmentations.

the good pseudo-labels from the student. Therefore, to address the aforementioned issues, we propose a more comprehensive approach to leverage the unlabeled objects mined by the student network and mitigate the noise issues in the pseudo-labels resulting from sparsely annotated issues and data augmentation.

3 Method

The overall framework of the proposed Co-Student is shown in Fig. 2, which comprises a denoising teacher detector and two collaborative students. Two students share the same set of parameters. The teacher detector updated by exponential moving average is the smoother version of the student detector. A raw image is fed to the teacher detector. A strongly augmented image and a weakly augmented image are fed to the strong and weak students, respectively. We use the standard FCOS head to predict the classification map, regression offsets map, and center-ness map to generate the original pseudo-label sets. Each pseudo-label set is then revised by the output pseudo-labels of the teacher model and merged with the sparse annotations. This way, the revised ground truth of the student detector is created and provides more precise supervision, which can train a more favorable student model. Then it can, in turn, enhance the denoising capability of the teacher model through EMA. We will describe Co-Student with the denoising teacher model in detail below.

3.1 *Co-Student*

We applied weak and strong data augmentation separately to each student branch to introduce more diversity in the data for *Co-Student* and enhance the possibility of mining unlabeled objects in two collaborative students. In the

top branch, *i.e.* the DTM branch, we do not use any augmentations. In the two bottom branches, *i.e.* the *Co-Student* branches, We use RandFlip and ColorJiter for the weak student. We use RandFlip, ColorJiter, Contrast, Equalize, Solarize, Sharpness, Posterize, RandTranslate, RandShear and RandRrase for the strong student. We divide all those augmentations into four groups: a random horizontal flip, a series of color space adjustments, a series of geometric transformations, and a random erase augmentation. The probabilities of augmentations in each group sum up to 1. Meanwhile, only one augmentation will be applied in each group at each iteration. Then for each branch, we use the predictions of the detector to generate pseudo-labels. The output of the teacher model is used to generate pseudo-labels through the same pseudo-label generator to provide stable and reliable pseudo-labels for *Co-Student*. Pseudo-label generator will be described in detail next. The denoising process will be described in detail in Section 3.2.

Pseudo-label generator: The generator's input is the prediction set of each branch. Let $C^j \in \mathbb{R}^{\hat{H}\cdot\hat{W}\cdot D}$ denote the classification map, where $j \in \{$raw, weak, strong$\}$, D stands for the categories number in the dataset. Then, the predictions with lower classification scores than a threshold θ, θ_s for student and θ_t for teacher, are filtered out, and the highest *top-k* predictions are kept as the candidate outputs. Note that the selections of candidates are calculated independently at each level of predictions in the corresponding level feature map. Here we take one of them as an example. We can obtain the candidate pseudo-labels P^j_{cd} as follows:

$$P^j_{cd} = \{S, B, C\} = \{\sqrt{\sigma(C^j[K] * T^j[K//D])}, BD(R^j[K//D]), K\%D\} \quad (1)$$

where {S, B, C} denotes {Scores, Bboxes, Classes}, $T^j \in \mathbb{R}^{\hat{H}\cdot\hat{W}\cdot 1}$ denotes the center-ness map, $R^j \in \mathbb{R}^{\hat{H}\cdot\hat{W}\cdot 4}$ denotes the regression offsets map, BD denotes the Box Decode, which is the reverse process of getting the regression targets.

Let P^j_{acd} denote all candidate pseudo-labels from all levels. Then, we apply Non-Maximum Suppression (NMS) with an IoU threshold α on P^j_{acd} to get the final output $P^j = \{x_0, y_0, x_1, y_1, score, class\}^N \in \mathbb{R}^{N\cdot 6}$. The teacher model has the same process of generating pseudo-labels with hyper-parameters score threshold θ_t and same IoU threshold α.

3.2 Denoising Student Labels using Teacher Model

Denoising Student Pseudo-label: We have three sets of pseudo-labels generated by the raw, weak, and strong branches, respectively. Each pseudo-label set is specific to its transformation, so direct comparison is not feasible. We transfer the pseudo-labels of the teacher P^r to the weak augmentation field first. Then, we use the transferred teacher's pseudo-labels P to denoise the student's pseudo-labels $P^\star$, as shown in Algorithm 1. $P^\star$ is determined based on the IoU, class, and score similarities with P. If the IoU between two pseudo-labels exceeds the threshold α_1 and the class and score conditions are met, $P^\star$ is revised by the

Algorithm 1 Algorithm of Denoising Student Pseudo-label

Input: Teacher model pseudo-label set $P \in \mathbb{R}^{N\cdot 6}$; Student model pseudo-label set $P^{\star} \in \mathbb{R}^{M\cdot 6}$ which need to be revised; BBoxes B, $B^{\star}$; Scores S, $S^{\star}$; Classes C, $C^{\star}$; IoU threshold α_1, α_2 for revision, and $\alpha_1 < \alpha_2$.
Output: denoised $\bar{P} = \{\bar{B}, \bar{S}, \bar{C}\}$;

for j = 1,2,..., M **do**
  $temp = \{ B_t, S_t, C_t \}$;
  for i = 1,2,..., N **do**
    if IoU(B_i, $B_j^{\star}$) $> \alpha_1$ and $C_i == C_j^{\star}$ and $S_i > S_j^{\star}$ **then**
      $temp$ += $\{B_i, S_i, C_i\}$;
    else if IoU(B_i, $B_j^{\star}$) $> \alpha_1$ and $C_i == C_j^{\star}$ and $S_i <= S_j^{\star}$ **then**
      $temp$ += $\{B_j^{\star}, S_j^{\star}, C_j^{\star}\}$;
    else if IoU(B_i, $B_j^{\star}$) $> \alpha_2$ and $C_i != C_j^{\star}$ and $S_i > S_j^{\star}$ **then**
      $temp$ += $\{B_i, S_i, C_i\}$;
    else
      $temp$ += $\{B_j^{\star}, S_j^{\star}, C_j^{\star}\}$;
  $temp_{max}$ = Max($\{B_t, S_t, C_t\}$), do max operate in S_t;
  $\bar{P}$ += $temp_{max}$
for i = 1,2,..., N **do**
  if IoU(B_i, $B_{j=1,2,...,M}^{\star}$) <= α_1 **then**
    $\bar{P}$ += $\{B_i, S_i, C_i\}$;
return $\bar{P}$

highest-confidence pseudo-label from P. The operation $Max(B_t, S_t, C_t)$ selects the highest-confidence pseudo-label from P that meets the correction criteria for a given student's pseudo-label. During training, the students exchange denoised pseudo-labels, so the final pseudo-label set of the weak student is the denoised pseudo-label set of the strong student.

Merging operation: Each student takes the combination of its sparse ground-truth $GT \in \mathbb{R}^{N\cdot 5}$ and a denoised pseudo-labels $\bar{P} \in \mathbb{R}^{M\cdot 6}$ as supervisions. The final supervisions of the weak student $\mathcal{A}^w = GT^w \ merge \ \bar{P}_i^w$. GT^w will merge all $\bar{P}_i^w$ where i belongs to the union of the following two sets:

$$\begin{aligned} &\{i \mid \forall j \in \{1,\ldots,N\},\ \text{IoU}(B_i^w,\ GB_j^w) \leq \alpha_3\} \\ &\{i \mid \exists j \in \{1,\ldots,N\},\ \text{IoU}(B_i^w,\ GB_j^w) > \alpha_3 \text{ and } C_i^w \neq GC_j^w\} \end{aligned} \tag{2}$$

where B and GB stand for the boxes of $\bar{P}^w$ and GT^w respectively. C and GC stand for the classes of $\bar{P}^w$ and GT^w respectively. α_3 is the IoU threshold. The overall loss of the proposed framework is defined as follows:

$$Loss = Loss_w + Loss_s = \mathcal{L}_{det}(\mathcal{A}^s, \mathcal{P}^w) + \mathcal{L}_{det}(\mathcal{A}^w, \mathcal{P}^s) \tag{3}$$

where $\mathcal{P}$ is the prediction map.

Table 1: The main results in `COCO-miss50p` setting. The methods marked with $\star$ are trained on fully annotated `COCO-train2017` set. The results marked with † are inherited from their publications.

Method	Backbone	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
FCOS* [26]	ResNet-50-FPN	38.9	57.6	42.0	23.0	42.7	50.1
FCOS [26]	ResNet-50-FPN	33.5	51.6	35.6	19.5	35.9	43.8
Co-mining [33]	ResNet-50-FPN	35.7	53.4	38.4	19.6	38.9	47.0
Co-Student (ours)	ResNet-50-FPN	**37.0**	**55.2**	**39.9**	**21.4**	**40.4**	**47.9**
SparseDet [23]	ResNet-101-FPN	35.9†	-	-	-	-	-
FCOS* [26]	ResNeXt-32x8d-101-BiFPN	47.1	66.6	51.4	29.8	50.7	62.4
FCOS [26]	ResNeXt-32x8d-101-BiFPN	39.2	58.2	42.5	22.8	41.3	52.8
Co-mining [33]	ResNeXt-32x8d-101-BiFPN	40.8	59.2	44.2	20.2	44.6	56.5
Co-Student (ours)	ResNeXt-32x8d-101-BiFPN	**43.4**	**62.8**	**47.4**	**25.2**	**47.5**	**57.0**

4 Experiments

4.1 Datasets and Evaluation

We conduct all experiments on the MS COCO dataset [16] (`train2017` and `val2017` split), and five different settings of the sparsely annotated training datasets `Easy` [42], `Hard` [42], `Extreme` [42], `COCO-50missp` [42], `Keep1` [12]. `Easy`: Randomly remove one annotation in each image, and above 20.60% annotations are removed. `Hard`: Randomly remove half of the annotations in each image, and above 39.00% annotations are removed. `Extreme`: Only one annotation in each image randomly remains, and above 64.95% annotations are removed. `COCO-50missp`: Randomly erases 50% of annotations for each category. It does not ensure that every training image has at least one annotation. A total of 16039 images with no annotation are removed from the original training set with 118287 images. `Keep1`: Keep only one annotation for each existing category in each image, and above 60% annotations are removed. We validate our model on 5000 images `COCO val2017` split and use the standard COCO style Average Precision (AP) as our metrics.

4.2 Implementation Details

We adopt FCOS [26] as our basic detector and ResNet-50 [8] with FPN [14] as the backbone. All basic hyper-parameters are the same as the best setting on FCOS (ResNet-50), which achieves 38.9 AP on COCO: the radius 1.5 for center sampling, predicting the center-ness scores on regression branch, using the *GIoU Loss* [20] for the regression branch and using Group Normalization [37] for the convolutional layers in heads. Moreover, we conduct our method on a deeper model ResNeXt-32x8d-101-BiFPN [24] based FCOS with deformable convolutions v2 [45] but without multi-scales training and test-time data augmentation. We adopt Co-mining [33] as our baseline in the setting of sparsely annotated datasets. We resize each input image's shorter edge to 800 during training and

Table 2: The main results in four settings of sparse annotations. We compare our model with previous state-of-the-art methods under the same training settings they used. The results marked with † are inherited from their publications.

Method	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
FCOS	37.3	56.2	40.1	22.0	40.1	47.4
Co-mining	37.6†	-	-	-	-	-
Co-mining	37.5	55.9	40.5	22.5	41.1	47.5
SparseDet	36.8†	-	-	-	-	-
Co-Student	**38.6**	**57.3**	**41.8**	**23.1**	**42.2**	**48.6**

(a) Results in `easy` setting.

Method	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
FCOS	33.5	51.5	35.8	18.8	36.4	43.3
Co-mining	34.4†	-	-	-	-	-
Co-mining	34.2	51.9	36.8	19.4	36.8	44.7
SparseDet	34.0†	-	-	-	-	-
Co-Student	**35.9**	**54.3**	**38.7**	**21.1**	**39.2**	**45.8**

(b) Results in `hard` setting.

Method	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
FCOS	22.6	35.5	23.7	9.9	23.9	34.0
Co-mining	23.1†	-	-	-	-	-
Co-mining	23.1	35.9	24.3	10.4	25.3	33.9
SparseDet	23.4†	-	-	-	-	-
Co-Student	**26.1**	**40.2**	**27.9**	**13.2**	**28.3**	**37.2**

(c) Results in `extreme` setting.

Method	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
FCOS	33.4	49.4	35.9	15.8	36.1	48.4
DMiner	33.9†	-	-	-	-	-
SparseDet	32.9†	-	-	-	-	-
Co-Student	**35.1**	**51.3**	**38.0**	**18.6**	**38.6**	**49.6**

(d) Results in `keep1` setting.

testing. We use 2 NVIDIA GeForce 3090 GPUs with a batch size of 20 for training. The networks are trained for 72000 iterations (about 12 epochs). The initial learning rate is set to 0.0125. A warm-up policy is used in the first 1000 iterations. At 48000 and 64000 iterations, the learning rate will be reduced to 10% of the original. As for hyper-parameters, *topk*, α, α_1, α_2, α_3, θ_s, θ_t are set to 1000, 0.6, 0.5, 0.9, 0.4, 0.5, 0.6, respectively.

4.3 Main Results

We trained our method on the COCO dataset with varied settings of sparse annotations. The results are summarized in Tab. 1 and Tab. 2. Tab. 1 contains the results of four methods training on sparsely annotated `COCO-miss50p`. Tab. 2a $\sim$ Tab. 2d contain the performance of previous state-of-the-art works and our work under `easy`, `hard`, `extreme`, and `keep1` settings.

Comparison with Co-mining [33]: Note that there is a 0.3% AP gap in FCOS [26] between *IoU Loss* and *GIoU Loss*. Their publication gives out the results of FCOS with *IoU Loss* [41] trained in three settings, `easy`, `hard`, `extreme`. Our method with *GIoU Loss* achieves 1.0%, 1.5%, and 3.0% AP improvements respectively in these settings. It is inspiring that our method achieves better results even though more annotations are removed. Our method outperforms the *GIoU Loss* version of Co-mining by 1.1%, 1.7%, and 3.0% AP improvements respectively. Furthermore, their paper did not provide the results of FCOS on `COCO-miss50p`, so we reconstructed the codebase of Co-mining with *GIoU Loss* and trained the model on `COCO-miss50p`. We obtain 35.7% AP by Co-mining and 37.0% by our method. The results demonstrate the superiority of our method.

Comparison with DMiner [12]: DMiner has done thorough research on CenterNet [44] and obtained 29.8% AP under the `keep1` setting by CenterNet-Res101, which is 2.0% AP higher than the baseline, i.e., directly training the CenterNet-Res101 with the `keep1`. However, they obtain 33.9% AP on `keep1` by FCOS-Res50-FPN, which is only 0.4% AP higher than the baseline. Our method achieves 35.1 % AP under the `keep1` setting by FCOS-Res50-FPN, which is 1.7% above the baseline. The results prove that our method is more effective on FCOS.

Comparison with SparseDet [23]: SparseDet has undergone extensive research and effective improvements on the two-stage detector Faster-RCNN, enabling its effective utilization in SAOD tasks. However, under the `COCO-miss50p` setting, its performance based on ResNet-101 is not satisfactory, which is 1.1% AP lower than our method based on FCOS with ResNet-50. And its results based on ResNet-50, which are 1.8% AP, 1.9% AP, 2.7% AP lower than our method across the Easy, Hard, and Extreme settings, respectively.

Using deeper model: To demonstrate our method *Co-Student* can still work on the deeper model, we train *Co-Student* and Co-mining with ResNeXt-32x8d-101-BiFPN [24] on the sparsely annotated datasets `COCO-miss50p`. The results in Tab. 2 shows that Co-mining slightly improves the performance when the model goes deeper, while *Co-Student* still boosts 4.2% AP higher than the Base. This phenomenon shows that Co-mining suffers from *label noise overfitting*, which could be worse when the model goes deeper. In contrast, *Co-Student* handles the noisy pseudo-labels by the denoising teacher model, which has a sample but very effective strategy to make better use of pseudo-labels in SAOD.

Comparison with DenseTeacher [43]: We compared our method with SSOD task method DenseTeacher, which also employs weak and strong data augmentations and dual-student branches but utilizes a different teacher module designed for semi-supervised tasks. When maintaining data augmentation, model size, and sparsely annotated settings as our model, DenseTeacher obtained 32.9% AP. Although its structure is very similar to ours, the results demonstrate that a vanilla student-teacher detector, without specific designs for the SAOD task, is not suitable for sparsely annotated scenarios.

Validating on other domain datasets: To demonstrate *Co-Student* can still work well on other domain datasets, we train *Co-Student* and Co-mining with FCOS-ResNet-50 on BDD100k [40] and TUPAC16 [27] datasets. BDD100k is a large-scale driving video dataset designed for autonomous driving, which has three parts, a training set with 70K images, a validation set with 10K images, and a test set with 20K images. Since the label of the test set is not public, we evaluate all methods on the validation set. TUPAC16 is a collection of histopathological images aimed at facilitating research in cancer diagnosis and treatment. TUPAC16 dataset consists of images from 73 breast cancer cases from three pathology centers, lab_1, lab_2, and lab_3. Due to the test set of TUPAC16 also not publicly available, we evaluate all methods on the lab_2 and lab_3 parts. We created a sparse version of BDD100k training set named `BDD100k-miss50p` and a sparse version of TUPAC16 training set named `TUPAC16-`lab_1`-miss50p` with the same generation process as `COCO-miss50p`. We used AP50 and F1-score to

Table 3: The results on other domain datasets. FAOD refers to fully annotated object detection.

Methods	Datasets	AP50	F_1-score
Baseline	BDD100k	51.4	-
Co-mining	BDD100k	52.2	-
Co-Student	BDD100k	**53.2**	-
FAOD	BDD100k	56.9	-
Baseline	TUPAC16	-	62.3
Co-mining	TUPAC16	-	64.9
Co-Student	TUPAC16	-	**66.0**
FAOD	TUPAC16	-	72.9

Table 4: The ablation results on augmentation methods. Rand-CA refers to the color-space augmentations. Rand-T refers to RandTranslate. Rand-S refers to RandShear. Rand-E refers to RandErase.

Rand-CA	Rand-T	Rand-S	Rand-E	AP
				36.0
✓				36.3
	✓			36.8
		✓		36.6
			✓	36.2
✓	✓	✓	✓	**37.0**

Table 5: The ablation results of detector FCOS-Res50-FPN with different components on `COCO-miss50p`. DTB denotes Dual Training Branches with pseudo-labels. SA denotes Strong Augmentations. DTM denotes Denoising Teacher Model. STB denotes Single Training Branch with pseudo-labels, *i.e.* the common Teacher-Student framework shown in Fig. 1.

Method	DTB	SA	DTM	AP	AP_{50}	AP_{75}	ΔAP
Baseline				33.5	51.6	35.6	0
Baseline		✓		28.4	45.2	29.7	-5.1
STB				34.9	53.0	37.5	+1.4
STB		✓		34.4	53.0	37.1	+0.9
Co-Student	✓			35.7	53.4	38.4	+2.2
Co-Student	✓		✓	36.1	53.7	38.8	+2.6
Co-Student	✓	✓		36.4	54.5	39.3	+2.9
Co-Student	✓	✓	✓	**37.0**	**55.2**	**39.9**	**+3.5**
FAOD				38.9	57.6	42.0	+5.4

evaluate the effectiveness of detection results for BDD100k and TUPAC16, respectively. Additionally, We keep all hyper-parameters unchanged. The results are shown in Tab. 3.

4.4 Ablation Studies

All ablation experiments are conducted on `COCO-miss50p` using FCOS with ResNet-50-FPN as a basic detector. We first conducted a series of ablation experiments on data augmentation of the strong student to observe which augmentations are effective. The augmentations of the weak student are unchanged. Results are shown in Tab. 4. The results show that all the data augmentations we used in the strong student play a positive role in our method for the SAOD task.

Then, we conducted a series of ablation experiments on model designs. The results are shown in Tab. 5. The Baseline method FCOS with ResNet-50-FPN achieves a 33.5% AP. Unfortunately, if we apply strong augmentations mentioned in Sec. 3.1 to the Baseline method directly, there is a significant performance degradation, about a 5.1% AP decrease. We believe that the direct introduction

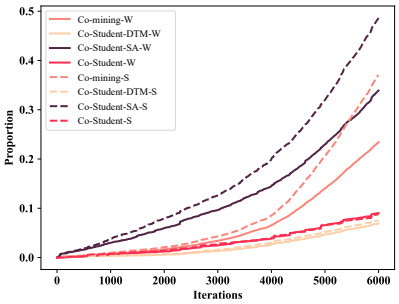


Fig. 3: The proportion of false pseudo-labels generated in each student. $\alpha_f = 0.4$.

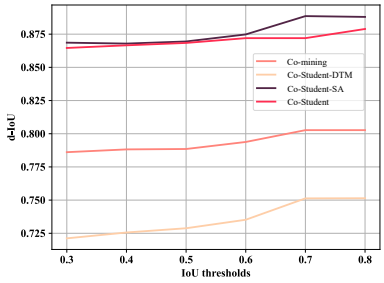
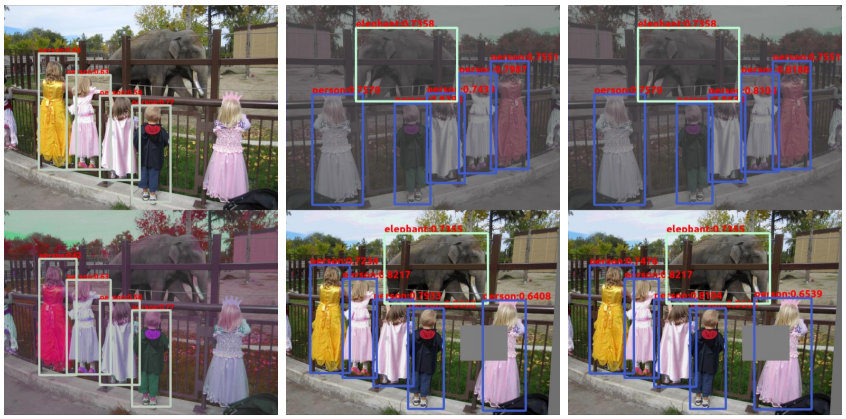


Fig. 4: The *d-IoU*s of all methods under different IoU thresholds α_d.

of strong augmentations will make the model more confused for the sparse annotations, which is counterproductive. We use the pseudo-labels generated by a simple EMA-based teacher model, following the common practice in Teacher-Student methods, to train the Single Training Branch (STB) with pseudo-labels. We achieve an improvement of 1.4% AP over the baseline. Similarly, strong augmentations do not work in STB. If we only use the Dual Training Branch (DTB) with pseudo-labels in our method *Co-Student*, it only achieves an improvement of 2.2% AP over the baseline. Afterwards, we combined the Denoising Teacher model on top of DTB, called Co-Student-DTM, and achieved an improvement of 2.6% AP. However it is not enough to demonstrate that the DTM model indeed improves the performance with its denoising ability. We will supplement relevant experimental results to prove this in the following section. We combined Strong Augmentations (SA) on top of DTB, called Co-Student-SA, an improvement of 2.9% AP is achieved. Note, we use strong augmentation only in one student. Similarly, it is not enough to prove that SA can mine more unlabeled objects. In the following sections, we will supplement relevant experimental results to prove this too. When combining those modules, Co-Student has an improvement of 3.5% AP compared to the baseline.

4.5 Quantitative Results

In order to elucidate the effectiveness of our method, we conducted an experiment to quantitatively analyze the denoising effect of DTM on false pseudo-labels and the mining ability of *Co-Student*. For ease of experimentation, we created a sparse version of `COCO-val2017` with the same generation process as `COCO-miss50p`, named `COCO-50p-val`. Then, we trained four different designs, Co-mining, Co-Student-DTM, Co-Student-SA, and Co-Student as mentioned in Tab. 4.4 on `COCO-50p-val` using a batch size of 12 and recorded all pseudo-labels and annotations after each iteration. All implementation details were identical to



(a) **(b)** **(c)**

Fig. 5: Qualitative results showing the qualities of the pseudo-labels generated by different methods. (a) The pseudo-labels generated by Co-mining; (b) The pseudo-labels generated by *Co-Student*; (c) The refined pseudo-labels denoised by the teacher model. It demonstrates the higher recall of unlabeled objects in *Co-Student* compared to Co-mining, along with the effective denoising ability of the teacher model.

those used in training on `COCO-miss50p`, except that the training was conducted for a total of 6k iterations.

To quantify how many false pseudo-labels sprout, we first define false pseudo-labels as follows: 1. IoU with all annotations lower than threshold α_f; 2. IoU with the maximum overlap annotations higher than threshold α_f and class mismatch with those annotations. The annotations here are fully annotated annotations. The proportions of false pseudo-labels in all pseudo-labels are shown in Fig. 3. Empirical evidence shows that the claim is valid. Specifically, during the entire training schedule, methods equipped with DTM, *i.e.* Co-Student-DTM and Co-Student, have very low proportions of false labels. In contrast, false pseudo-labels of the model without DTM become more and more prominent as the training progresses. Moreover, DTM can help reduce the negative impact on model performance caused by the presence of noise generated by SA. We mentioned earlier that the DTM can prevent the accumulation of incorrect information in the pseudo-labels generated by the student network, while the positive gain feeds back to the teacher network to further strengthen its denoising performance.

To quantify the extent of diverse supervisions brought by SA, we define a differential IoU ($d\text{-}IoU$) as follows: $d\text{-}IoU = 1 - \frac{Set_W \cap Set_S}{Set_W \cup Set_S}$, where Set_W and Set_S are the recalled unlabeled annotations sets of weak student W and strong student S, respectively. We define a set of recalled pseudo-labels for each annotation box as the collection of all pseudo-labels that have the same category as that of the box, and whose IoU with the box exceeds threshold α_d. We define the pseudo-label among the set of recalled pseudo-labels that has the highest IoU with the box as recalled unlabeled annotation. The results are shown in Fig. 4.

It can be seen that under different IoU thresholds α_d, all methods with SA, *i.e.* Co-Student-SA and Co-Student, have higher *d-IoU*s, which means strong augmentations with dual training branches can introduce appreciable differentiation to the pseudo-labels and increase the probability of obtaining more valuable information, broadening the supervision source for model learning and enable the model to extract more valuable insights from the sparse annotations.

Based on the results shown in Tab. 5, we find that SA contributes to the performance improvement by nearly 60%. This includes not only the positive gains brought by data augmentation itself but also the broader supervision sources it introduces. We believe that these two factors are the main reasons why SA can maintain certain performance despite having a larger proportion of noisy pseudo-labels. Additionally, we can further boost performance by denoising the pseudo-label set using DTM.

4.6 Qualitative Results

We visualized the pseudo-labels generated by both the Co-Student and Co-mining methods after training for half of the total iterations. The qualitative results are shown in Fig. 5. Fig. 5 (a) depicts the pseudo-labels generated by Co-mining, where it can be observed that both branches have missed the elephant and the little girl on the far right. Fig. 5 (b) depicts the pseudo-labels generated by our Co-Student method, demonstrating a higher recall of unlabeled instances. However, in the pseudo-labels corresponding to the strong student, as shown in the lower part of Fig. 5 (b), the localization of the little girl on the far left and the far right appears inaccurate, introducing noise. Finally, Fig. 5 (c) displays the pseudo-labels after denoising by the teacher, showing improved accuracy and reliability.

5 Conclusion

In this paper, we proposed a novel *Co-Student* framework to fully utilize the pseudo annotations for unlabeled objects, in which the two students with strong and weak augmentation collaboratively excavate more pseudo annotations and exchange the supervisions. To address the noise issue in pseudo-labels, we utilized an EMA-based teacher model to denoise the pseudo-labels for the students, effectively mitigating the performance degradation caused by noise in pseudo-labels. Our approach achieved state-of-the-art results on various sparse settings of the MS COCO dataset and outperformed the previous method Co-mining in the datasets from two different domains.

Acknowledgements: This work was supported by the National Science and Technology Major Project under Grant No. 2023YFF0905400. We are grateful to Tianheng Cheng and Cheng Wang for their help with the preparation of figures and writing in this paper.

References

1. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the eleventh annual conference on Computational learning theory. pp. 92–100 (1998)
2. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6154–6162 (2018)
3. Dai, J., Li, Y., He, K., Sun, J.: R-fcn: Object detection via region-based fully convolutional networks. Advances in neural information processing systems **29** (2016)
4. Deng, J., Li, W., Chen, Y., Duan, L.: Unbiased mean teacher for cross-domain object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4091–4101 (2021)
5. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: Centernet: Keypoint triplets for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6569–6578 (2019)
6. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. Advances in neural information processing systems **31** (2018)
7. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR **abs/1704.04861** (2017), `http://arxiv.org/abs/1704.04861`
10. Huang, G., Liu, Z., Weinberger, K.Q.: Densely connected convolutional networks. CoRR **abs/1608.06993** (2016), `http://arxiv.org/abs/1608.06993`
11. Ke, Z., Wang, D., Yan, Q., Ren, J., Lau, R.W.H.: Dual student: Breaking the limits of the teacher in semi-supervised learning (2019)
12. Li, H., Pan, X., Yan, K., Tang, F., Zheng, W.S.: Siod: Single instance annotated per category per image for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14197–14206 (2022)
13. Li, X., Wang, W., Hu, X., Li, J., Tang, J., Yang, J.: Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11632–11641 (2021)
14. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
15. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
16. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
17. Liu, Y.C., Ma, C.Y., He, Z., Kuo, C.W., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P.: Unbiased teacher for semi-supervised object detection. arXiv:2102.09480 (2021)

18. Liu, Y.C., Ma, C.Y., Kira, Z.: Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9819–9828 (2022)
19. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
20. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 658–666 (2019)
21. Scudder, H.: Probability of error of some adaptive pattern-recognition machines. IEEE Transactions on Information Theory **11**(3), 363–371 (1965)
22. Sohn, K., Zhang, Z., Li, C.L., Zhang, H., Lee, C.Y., Pfister, T.: A simple semi-supervised learning framework for object detection. arXiv preprint arXiv:2005.04757 (2020)
23. Suri, S., Rambhatla, S., Chellappa, R., Shrivastava, A.: Sparsedet: Improving sparsely annotated object detection with pseudo-positive mining. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6770–6781 (2023)
24. Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10781–10790 (2020)
25. Tang, Y., Chen, W., Luo, Y., Zhang, Y.: Humble teachers teach better students for semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3132–3141 (2021)
26. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9627–9636 (2019)
27. Veta, M., Heng, Y.J., Stathonikos, N., Bejnordi, B.E., Beca, F., Wollmann, T., Rohr, K., Shah, M.A., Wang, D., Rousson, M., et al.: Predicting breast tumor proliferation from whole-slide images: the tupac16 challenge. Medical image analysis **54**, 111–121 (2019)
28. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv preprint arXiv:2207.02696 (2022)
29. Wang, C.Y., Yeh, I.H., Liao, H.Y.M.: You only learn one representation: Unified network for multiple tasks. arXiv preprint arXiv:2105.04206 (2021)
30. Wang, H., Liu, L., Zhang, B., Zhang, J., Zhang, W., Gan, Z., Wang, Y., Wang, C., Wang, H.: Calibrated teacher for sparsely annotated object detection (2023)
31. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B.: Deep high-resolution representation learning for visual recognition. CoRR **abs/1908.07919** (2019), `http://arxiv.org/abs/1908.07919`
32. Wang, K., Zhuang, J., Li, G., Fang, C., Cheng, L., Lin, L., Zhou, F.: De-biased teacher: Rethinking iou matching for semi-supervised object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2573–2580 (2023)
33. Wang, T., Yang, T., Cao, J., Zhang, X.: Co-mining: Self-supervised learning for sparsely annotated object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 2800–2808 (2021)

34. Wang, X., Yang, X., Zhang, S., Li, Y., Feng, L., Fang, S., Lyu, C., Chen, K., Zhang, W.: Consistent-teacher: Towards reducing inconsistent pseudo-targets in semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3240–3249 (2023)
35. Wang, Z., Li, Y., Guo, Y., Fang, L., Wang, S.: Data-uncertainty guided multi-phase learning for semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4568–4577 (2021)
36. Wei, H., Feng, L., Chen, X., An, B.: Combating noisy labels by agreement: A joint training method with co-regularization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13726–13735 (2020)
37. Wu, Y., He, K.: Group normalization. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
38. Wu, Z., Bodla, N., Singh, B., Najibi, M., Chellappa, R., Davis, L.S.: Soft sampling for robust object detection. arXiv preprint arXiv:1806.06986 (2018)
39. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10687–10698 (2020)
40. Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V., Darrell, T., et al.: Bdd100k: A diverse driving video database with scalable annotation tooling. arXiv preprint arXiv:1805.04687 **2**(5), 6 (2018)
41. Yu, J., Jiang, Y., Wang, Z., Cao, Z., Huang, T.: Unitbox: An advanced object detection network. In: Proceedings of the 24th ACM international conference on Multimedia. pp. 516–520 (2016)
42. Zhang, H., Chen, F., Shen, Z., Hao, Q., Zhu, C., Savvides, M.: Solving missing-annotation object detection with background recalibration loss. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1888–1892. IEEE (2020)
43. Zhou, H., Ge, Z., Liu, S., Mao, W., Li, Z., Yu, H., Sun, J.: Dense teacher: Dense pseudo-labels for semi-supervised object detection (2022)
44. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)
45. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: More deformable, better results. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9308–9316 (2019)