Unbiased Teacher v2: Semi-supervised Object Detection for Anchor-free and Anchor-based Detectors
Supplementary Material

Table 1. Results of the **anchor-free model (FCOS)** on *VOC*.

	Models	Labeled	Unlabeled	$AP_{50:95}$
Supervised	FCOS-R50	VOC07	None	41.53
STAC [4]	FCOS-R50			44.47 (+2.94)
Unbiased Teacher [3]	FCOS-R50	VOC07	VOC12	49.50 (+7.97)
Ours	FCOS-R50			**56.71** (+15.13)
STAC [4]	FCOS-R50		VOC12	44.89 (+3.36)
Unbiased Teacher [3]	FCOS-R50	VOC07	+	51.15 (+9.62)
Ours	FCOS-R50		*COCO20cls*	**57.91** (+16.38)

1. Additional Results on Anchor-free Detector

COCO-additional. To examine whether the fully-supervised **anchor-free detector (FCOS)** can be improved by using the additional unlabeled images, we also consider *COCO-additional* and *VOC*. We present *COCO-additional* in Table 2, and our method can also perform 7.19 mAP improvement compared to the supervised baseline, 6.42 mAP absolute improvement compared with STAC [4], and 4.64 mAP absolute improvement compared with Unbiased Teacher [3].

VOC. A similar trend is also found in *VOC* as presented in Table 1, where our model achieves 56.71 mAP, which shows 15.13 mAP improvement over the supervised only baseline. By leveraging unlabeled *COCOcls20* in the training, our model can further improve and obtain 57.91 mAP, which is 16.38 mAP higher than the supervised baseline. These results confirm the effectiveness of our method in improving the fully-supervised object detector with additional unlabeled images.

2. Effect of margin between localization uncertainties of Teacher and Student

We presented the *Listen2Student* mechanism, which selects pseudo-labels based on whether the boundary predictions satisfy $\delta_t + \sigma \leq \delta_s$, where δ_t is the estimated boundary uncertainty and δ_s is the estimated boundary uncertainty of student.

We ablate the margin and examine its sensitivity in Table 3. We observe that when the margin is set as $\sigma = 0.0$ (*i.e.,* select all instances where the teacher has lower uncertainty than the student), the model can achieve 22.15 mAP. By further increasing the margin to 0.2, the result can be improved to 22.73 mAP, since the ratio of instances correctly reflecting teacher's prediction is better than the student's prediction is higher when the selection condition becomes stricter. However, when we further increase the margin to 0.4, the performance drops to 21.91 mAP. This is because the number of pseudo-labels becomes fewer when the extremely strict conditions are enforced. We thus use the margin $\sigma = 0.2$ for *COCO-standard* experiments.

3. Definition of Misleading and Beneficial Instances

To better approach the pseudo-labeling method in semi-supervised object detection, we explicitly differentiate the unlabeled instances with pseudo-labels into two categories, *beneficial* and *misleading* instances. To learn object detectors in a semi-supervised manner, the pseudo-labeling method updates the model θ^n with both a supervised loss $\mathcal{L}_s$ and an unsupervised loss $\mathcal{L}_u$,

$$\theta^{n+1} = \theta^n + \gamma \frac{\partial}{\partial \theta^n}(\mathcal{L}_s + \boldsymbol{\lambda}_u \mathcal{L}_u) \tag{1}$$

$$= \theta^n + \gamma\Big(\frac{\partial}{\partial \theta^n}\mathcal{L}_s + \boldsymbol{\lambda}_u \underbrace{\Big(\frac{\partial}{\partial \theta^n} \sum_{\hat{y}_i^u \in \Psi^+} \mathcal{L}_u(x_i^u, \hat{y}_i^u)}_{\text{Beneficial Gradient}} + \tag{2}$$

$$\underbrace{\frac{\partial}{\partial \theta^n} \sum_{\hat{y}_i^u \in \Psi^-} \mathcal{L}_u(x_i^u, \hat{y}_i^u)}_{\text{Misleading Gradient}}\Big)\Big), \tag{3}$$

where x_u^i is the i-th unlabeled image in a batch, $\hat{y}_i^u$ is the corresponding pseudo-label, Ψ^+ is the beneficial set, Ψ^- is the misleading set, γ is the learning rate, and $\boldsymbol{\lambda}_u$ is the unsupervised loss weight. Both the confidence thresholding and our *Listen2Student* aim to remove the misleading gradient, which is derived from the incorrect pseudo-labels.

Table 2. Experimental results of the anchor-free model (FCOS-ResNet50) on *COCO-additional*. Note that 1x represents 90K training iterations, and Nx represents $N\times$90K training iterations.

	Anchor-free detectors on COCO-additional						
	Supervised (1x)	Supervised (3x)	CSD (3x)	STAC (3x)	Unbiased Teacher (3x)	Ours (3x)	Ours (4x)
$AP^{50:95}$	37.10	37.29	36.70 (-0.59)	38.06 (+0.77)	39.84 (+2.55)	**44.02**(+6.73)	**44.48** (+7.19)

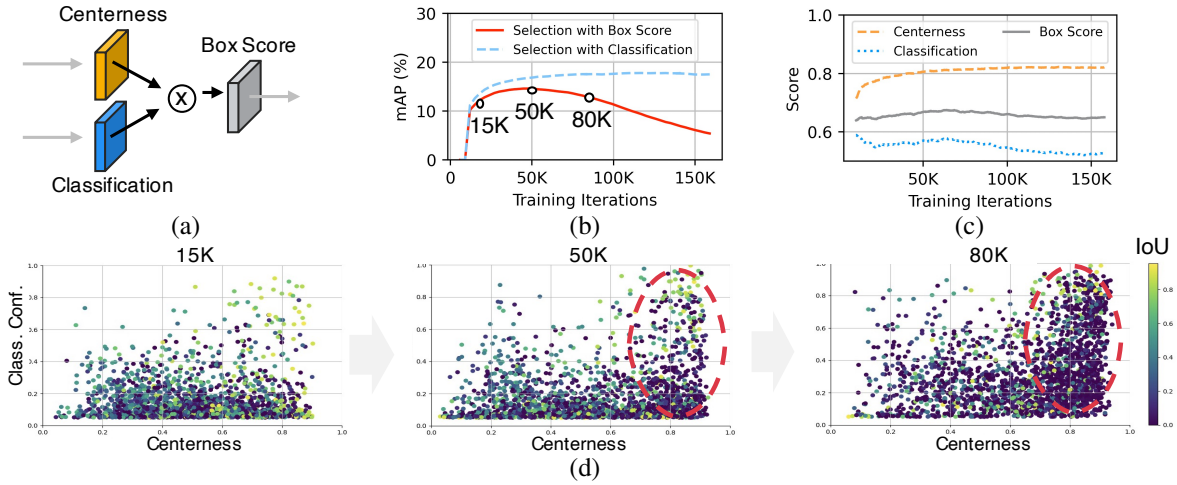

Figure 1. **Centerness bias issue in semi-supervised anchor-free detectors**. We analyze the detector in terms of its (b) mean average precision (mAP), (c) averaged scores of pseudo-boxes, and (d) correlation between centerness and classification score (y-axis) and IoU (colorization) for each predicted boxes. **(a)** Box score of some anchor-free detectors [7, 10] is derived by multiplying the centerness and the classification score. **(b)** By using the box score to select pseudo-labels, the performance of the detector starts to degrade after 50K training iterations. **(c)** We observe the box score of pseudo-boxes is dominated by the centerness score, which cannot precisely reflect whether the predicted boxes are foreground classes. As the predictions of the Teacher model is used as pseudo-labels for the Student model (*i.e., pseudo-labeling*), the high-centerness background instances (points colored in dark blue in **(d)**) appear more frequently when the detector is trained longer.

Table 3. Ablation study of varying margin between the localization uncertainties of Teacher and Student σ in the case of *COCO-standard* 1%. Faster-RCNN is used as the model for all results below.

σ	0.0	0.1	0.2	0.3	0.4
mAP	22.15	22.27	**22.73**	22.57	21.91

Table 4. Pseudo-box selection based on classification scores is more effective than pseudo-box selection based on box score (which contains centerness score) in COCO-additional. All results are based on FCOS [7].

	Class. score	Box score	Δ
COCO-additional	44.48	42.50	**-1.98**

Table 5. Using center-sampling leads to a worse accuracy in COCO-additional. All results are based on FCOS [7].

	w/o Center-Sampling	w/ Center-Sampling	Δ
COCO-additional	43.23	42.50	**-0.53**

Table 6. Ablation study on Listen2student and scale jitter. Note that we experiment on COCO-standard with 10% supervision. Faster-RCNN is used as the model for all results below.

Listen2student	Scale jitter	mAP
-	-	30.91
✓	-	**32.18**
-	✓	31.83
✓	✓	**33.55**

3.1. More on Adaption to Anchor-free Detectors

Further Discussion on Centerness bias issue As presented in Figure 1b, we notice that selecting the pseudo-boxes based on box scores performs worse than solely relying on classification scores in the semi-supervised setting, while FCOS [7] shows using box scores leads to better re-

Table 7. Meanings and values of the hyper-parameters used in **FCOS** in experiments.

Hyper-parameter	Description	*COCO-standard*	*VOC*	*COCO-additional*
τ	Class. confidence threshold	0.5	0.5	0.5
λ_u^{cls}	Unsupervised loss weight for classification	3	3	2
λ_u^{reg}	Unsupervised loss weight for regression	0.2	0.2	0.2
α	EMA rate	0.9999	0.9999	0.9999
b_l	Batch size for labeled data	8	16	32
b_u	Batch size for unlabeled data	8	16	32
γ	Learning rate	0.01	0.01	0.01

Table 8. Meanings and values of the hyper-parameters used in **Faster-RCNN** in experiments.

Hyper-parameter	Description	*COCO-standard*	*VOC*	*COCO-additional*
δ	Confidence threshold	0.7	0.7	0.7
λ_u^{cls}	Unsupervised loss weight for classification	4	2	2
λ_u^{reg}	Unsupervised loss weight for regression	1.0	1.0	1.0
α	EMA rate	0.9996	0.9996	0.9996
b_l	Batch size for labeled data	8/32	8	32
b_u	Batch size for unlabeled data	8/32/40	8	32
γ	Learning rate	0.01	0.01	0.01

sults in the fully-supervised setting. We observed that this is because the box score of some anchor-free detectors [7, 10] is defined as the multiplication of classification score and centerness score (see Figure 1a), and the pseudo-boxes selected based on the box score have relatively high centerness score but low classification scores (see Figure 1c). This reveals that the box score is dominated by the centerness score in the pseudo-labeling mechanism. However, with the limited amount of labels used in the training, the centerness score is not reliable for reflecting whether a prediction is a foreground instance since there is no supervision to suppress the centerness score for background instances in the centerness branch. As a result, these selected high centerness pseudo-box are likely to be the background instances (see Figure 1d), and adding these false-positive pseudo-boxes in the semi-supervised training degrades the effectiveness of the pseudo-labeling and also aggravates the centerness bias issue.

Analysis on COCO-additional. We also analyze whether an object detector also suffers from the centerness bias issue and unreliable label assignment. We thus follow the experimental setup in COCO-additional and show the analysis in Table 4 and 5. The trend is consistent with what we observed in the COCO-standard (*i.e,* randomly sample $0.5-10\%$ as the labeled set).

4. Complete Implementation Details

Network and framework. We build our method upon the Detectron2 framework [8] (with Apache License). For the anchor-free experiments, we use the FCOS, with implementation from AdelaiDet [6]. For the anchor-based experiments, we use the official implementation in Detectron2. For a fair comparison, both FCOS and Faster-RCNN use the ResNet50 as the feature backbone pretrained on the ImageNet, which is a common procedure in the prior works [2–5, 11]

Training Details and Hyper-parameters. Our training procedure follows the Unbiased Teacher [3], which contains the burn-in stage (*i.e.,* train an object detector with the available labeled data) and the mutual learning stage (*i.e.,* additionally use the pseudo-labels for unlabeled data with teacher-student mechanism). For the *COCO-standard* with FCOS, we train 180k iterations, which includes $10/15/20/25/30$k iterations for $0.5\%/1\%/2\%/5\%/10\%$ in the Burn-In stage and the remaining iterations in the Teacher-Student Mutual Learning stage. For the *COCO-additional*, we train 360k iterations, which includes 90k iterations in the Burn-Up stage and the remaining 270k iterations in the Teacher-Student Mutual Learning stage. For the Faster-RCNN, we follow the setups in Unbiased Teacher [3]. We summarize the hyper-parameters for FCOS in Table 7 and Faster-RCNN in Table 8.

Data augmentation. We follow the data augmentation used in Unbiased Teacher [3], which uses a random horizontal flip for weak augmentation and randomly adding color jittering, grayscale, Gaussian blur, and cutout patches [1] for the strong augmentation. We additionally consider scale jitter used in SoftTeacher [9] to further improve the performance, and we present the effect of scale jitter in Table 6. Note that Image-level or box-level geometric augmentations,

such as rotation, translation, and Mosaic [11], are not used in our method.

References

[1] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 3

[2] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3

[3] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 1, 3

[4] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 1, 3

[5] Peng Tang, Chetan Ramaiah, Ran Xu, and Caiming Xiong. Proposal learning for semi-supervised object detection. *arXiv preprint arXiv:2001.05086*, 2020. 3

[6] Zhi Tian, Hao Chen, Xinlong Wang, Yuliang Liu, and Chunhua Shen. AdelaiDet: A toolbox for instance-level recognition tasks. https://git.io/adelaidet, 2019. 3

[7] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 3

[8] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 3

[9] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. *arXiv preprint arXiv:2106.09018*, 2021. 3

[10] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3

[11] Qiang Zhou, Chaohui Yu, Zhibin Wang, Qi Qian, and Hao Li. Instant-teaching: An end-to-end semi-supervised object detection framework. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 4