Appendix: End-to-End Semi-Supervised Object Detection with Soft Teacher

Mengde Xu[1†*] Zheng Zhang[1,2* ‡] Han Hu[2‡] Jianfeng Wang[2] Lijuan Wang[2] Fangyun Wei[2]
Xiang Bai[1] Zicheng Liu[2]
[1]Huazhong University of Science and Technology
{mdxu,xbai}@hust.edu.cn, macaroniz1990@gmail.com
[2]Microsoft
{zhez, hanhu,jianfw,lijuanw,fawe,zliu}@microsoft.com

1. Implementation Details

We use the Faster R-CNN [4] equipped with FPN [3] (Feature Pyramid Network) as our default detection framework to evaluate the effectiveness of our method, and an ImageNet pre-trained ResNet-50 [2] is adopted as the backbone. Our implementation and hyper-parameters are based on MMDetection [1]. Anchors with 5 scales and 3 aspect ratios are used. 2k and 1k region proposals are generated with a non-maximum suppression threshold of 0.7 for training and inference. In each training step, 512 proposals are sampled from 2k proposals as the box candidates to train RCNN. Since the amount of training data of **Partially Labeled Data** setting and **Full Labeled Data** setting has large differences, the training parameters under the two settings are slightly different.

Partially Labeled Data: The model is trained for 180k iterations on 8 GPUs with 5 image per GPU. With SGD training, the learning rate is initialized to 0.01 and is divided by 10 at 110k iteration and 160k iteration. The weight decay and the momentum are set to 0.0001 and 0.9, respectively. The foreground threshold is set to 0.9 and the data sampling ratio s_r is set to 0.2 and gradually decreases to 0 over the last 10k iterations.

Fully Labeled Data: The model is trained for 720k iterations on 8 GPUs with 8 image per GPU. In SGD training, the learning rate is initialized to 0.01 and is divided by 10 at 480k iteration and 680k iteration. The weight decay and the momentum are set to 0.0001 and 0.9, respectively. The foreground threshold is set to 0.9 and the data sampling ratio s_r is set to 0.5 and gradually decreases to 0 in the last 20k iterations.

For estimating the box localization reliability, we set N_{jitter} as 10, and threshold is set as 0.02 to select the pseudo-labels for box regression. The jittered boxes are randomly sampled by adding the offsets on four coordinates, and the offsets are uniformly sampled from [-6%, 6%] of the height or width of the pseudo box candidates. In addition, we follow STAC and FixMatch to use different data augmentation for pseudo-label generation, labeled image training and unlabeled image training. The details are summarized in Table .1.

Augmentation	Labeled image training	Unlabeled image training	Pseudo-label generation
Scale jitter	short edge $\in (0.5, 1.5)$	short edge $\in (0.5, 1.5)$	short edge $\in (0.5, 1.5)$
Solarize jitter	p=0.25, ratio $\in (0, 1)$	p=0.25, ratio $\in (0, 1)$	-
Brightness jitter	p=0.25, ratio $\in (0, 1)$	p=0.25, ratio $\in (0, 1)$	-
Constrast jitter	p=0.25, ratio $\in (0, 1)$	p=0.25, ratio $\in (0, 1)$	-
Sharpness jitter	p=0.25, ratio $\in (0, 1)$	p=0.25, ratio $\in (0, 1)$	-
Translation	-	p=0.3, translation ratio $\in (0, 0.1)$	-
Rotate	-	p=0.3, angle $\in (0, 30°)$	-
Shift	-	p=0.3, angle $\in (0, 30°)$	-
Cutout	num $\in (1, 5)$, ratio $\in (0.05, 0.2)$	num $\in (1, 5)$, ratio $\in (0.05, 0.2)$	-

Table 1. The summary of the data augmentation used in our approach. We follow the practice of STAC [6] and FixMatch [5] to provide different data augmentation for pseudo-label generation, labeled image training and unlabeled image training. “-” indicates the augmentation is not used.

[*]Equal contribution. †This work is done when Mengde Xu was intern in MSRA. ‡Contact person.

References

[1] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[3] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 1

[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1

[5] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *NIPS*, 2020. 1

[6] 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