MixTeacher: Mining Promising Labels with Mixed Scale Teacher for Semi-Supervised Object Detection

Supplementary Material

Liang Liu[1], Boshen Zhang[1], Jiangning Zhang[1], Wuhao Zhang[1], Zhenye Gan[1]
Guanzhong Tian[3], Wenbing Zhu[4], Yabiao Wang[1†], Chengjie Wang[1,2†],
[1]Youtu Lab, Tencent [2]Shanghai Jiao Tong University
[3]Ningbo Research Institute, Zhejiang University, [4]Rongcheer Co., Ltd
{leoneliu, boshenzhang, vtzhang, wuhaozhang, wingzygan}@tencent.com;
gztian@zju.edu.cn; louis.zhu@rongcheer.com; {caseywang, jasoncjwang}@tencent.com

1. Experimental Details

Different SSOD methods may implement with different data augmentation strategies and training hyper-parameters which have a great impact on the performance. As the choice of the majority, our implementation and hyper-parameters are based on MMDetection, with the base model of `FasterRCNN-R50-FPN`. We implement MixTeacher without any modification on the model design and loss formulation, except for the necessary module and losses introduced by the mixed scale teacher in training (which are dropped in testing). The training hyper-parameters are summarized in Table 1.

Training Setting	COCO-Partial	COCO-Additional	VOC
Batch size for labeled data	8	32	16
Batch size for unlabeled data	32	32	16
Learning rate	0.01	0.01	0.01
Learning rate step	(120k, 160k)	(480k, 640k)	-
Iterations	180k	720k	40k
Unsupervised loss weight λ	4.0	2.0	2.0
EMA rate	0.999	0.999	0.999
Temperature T	3	3	3
Mine score thresh τ_l	0.7	0.7	0.7
Mine diff thresh δ	0.1	0.1	0.1
Test score threshold	0.001	0.001	0.001

Table 1. The summary of training settings for different settings.

Note that, as we illustrated in the main manuscript, we adopt the confidence score thresholds $\tau_h = 0.9$ and $\tau_l = 0.7$ to select and mine pseudo labels for the classification loss of RCNN and the classification and regression losses of RPN. Moreover, we follow the practice in Soft Teacher [11] which adopts a different strategy to filter out pseudo labels for the regression loss of RCNN. Concretely, the pseudo labels with a confidence score higher than 0.5 are selected as the candidates, and the candidates with uncertainty lower than 0.02 are selected for RCNN regression. The estimation of uncertainty for the box localization reliability is implemented by jittering each predicted box 10 times as a group of proposals, and computing the standard deviation of the corresponding location predictions for the group of proposals. The offsets of jittering are uniformly sampled from [-6%, 6%] of the height or width of the pseudo box candidates. If necessary, please refer to [11] for the details of this part.

In addition, strong-weak augmentation is commonly used in semi-supersvied learning, we follow previous works to use different augmentations for labeled data, unlabeled images, and pseudo labels generation during training. The details of data augmentation are summarized in Table 2.

2. Training Efficiency

Since our method brings extra computations for additional scales, the major concern might be the training efficiency of our method. Although we have reported the training speed in Table 5 of the main manuscript, we further investigate the convergence speed for different models. We plot the evaluation results during training for different models in Figure 1. Compared with our baseline, *i.e.* the original version of soft teacher, our method obtains a similar result of 33.9 mAP with only 1/3 of total iterations, and finally reaches a significant improvement to 36.7 mAP. Comparing with the most recent method PseCo [4] which also uses an additional down-sampled view but still generates pseudo labels from the regular scale, our method also behaves superiority, for which obtains a comparable result with only 40% iteration. Furthermore, we conduct a experiment to compare the proposed MixTeacher with a version named MixTeacher-RD

† Corresponding Authors.

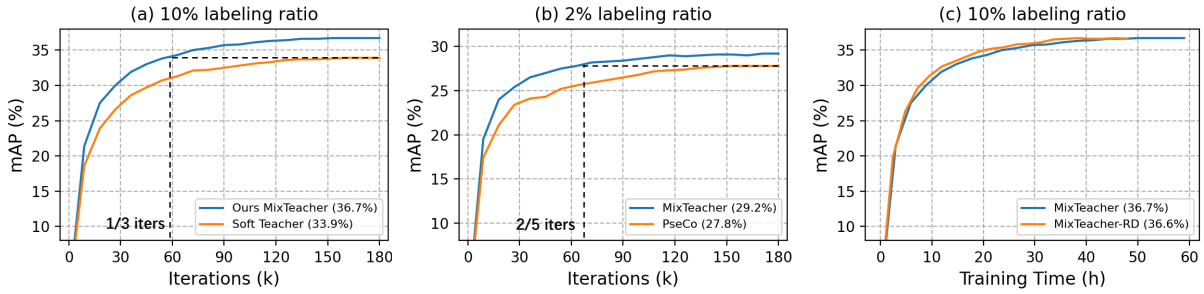

Figure 1. Comparison of model convergence speed in COCO partial labeled setting. (a) Compare MixTeacher against Soft Teacher [11] under 10% labeling ratio. (b) Compare MixTeacher against PseCo [4] under 1% labeling ratio. (c) Compare MixTeacher against MixTeacher-RD under 10% labelling ratio, which randomly drops a path from the regular scale and the mixed scale for unlabeled images in every iteration. In legend, the numbers in brackets refer to the final mAP. Performance is evaluated on the teacher model.

that ramdonly drops a $1\times$ scale path for the student model in each iteration of unlabeled data. As shown in Figure 1 (c), randomly dropping a path can reduce the time consumption of each training iteration, but still reaches a comparable results in the end. More specifically, we report the performance of MixTeacher-RD under all four labelling ratio of COCO partial label settings in Table 3. The results demonstrate that when using a single $1\times$ scale view and a $0.5\times$ scale view as previous multiple views SSOD methods [3,4], our method still improves the performance significantly.

3. Bells and whistles in SSOD

In order to avoid the confusion about what makes results better, we follow a quite simple baseline, in which some tricks that known to improve results are not used. For instance, PseCo [4] uses Focal Loss [6] to replace the cross entropy loss in the original Faster-RCNN implementation, which has been proven can bring $+0.6$ mAP improvement in their work. Unbiased Teacher [7] adopts a larger batch size than ours, and we tried it on our baseline with getting $+0.3$ mAP gains but increasing the training time.

Besides, inspired from the progress in fully supervised object detection, such as GIoU loss [8], dynamically hard label assignment [2, 13], and soft label assignment [5, 14], recent SSOD methods resort to more advanced label assignment strategies [1, 4], or more efficient localization loss [12]. We keep the ordinary implementation to demonstrate the effective of proposed method. We believe it is unnecessary to spend time on trying these components, although they are highly likely to bring better results for MixTeacher, .

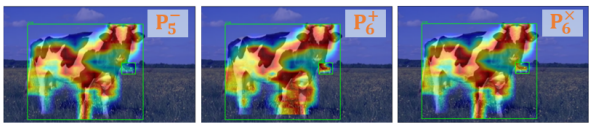

Figure 2. EigenCAM visualization for layers in different feature pyramids. $\gamma = 0.18$ leads the $P_6^\times$ more similar to P_5^-.

4. Feature Visulization

γ is derived from channel-wise attention on the regular scale and down-sampled scale features, serving as a weight in the linear combination of these two features. The weighted sum formulation acts as a gate mechanism to select more appropriate feature for each level. We show activation maps in Fig.2 for an image with two cows in different sizes. In this case, the 5th level of the down-sampled pyramid shows more accurate for the large cow than the 6th level of the regular

Augmentation	Labeled Data Aug.	Unlabeled Strong Aug.	Unlabeled Weak Aug.
Scale jitter	short edge ∈ (0.5,1.5)	short edge ∈ (0.5,1.5)	short edge ∈ (0.5,1.5)
Horizontal flip	p=0.5	p=0.5	p=0.5
Solarize jitter	p=0.25, ratio ∈ (0,1)	p=0.25, ratio ∈ (0,1)	–
Brightness jitter	p=0.25, ratio ∈ (0,1)	p=0.25, ratio ∈ (0,1)	–
Constrast jitter	p=0.25, ratio ∈ (0,1)	p=0.25, ratio ∈ (0,1)	–
Sharpness jitter	p=0.25, ratio ∈ (0,1)	p=0.25, ratio ∈ (0,1)	–
Translation	–	p=0.3, translation ratio ∈ (0,0.1)	–
Rotate	–	p=0.3, angle ∈ (0,30°)	–
Shift	–	p=0.3, angle ∈ (0,30°)	–
Cutout	num ∈ (1,5), ratio ∈ (0.05,0.2)	num ∈ (1,5), ratio ∈ (0.05,0.2)	–

Table 2. The summary of training settings for different datasets and different settings. We follow the practive of Soft Teacher [11], STAC [10], and FixMatch [9] to adopt different hyper-parameters for labeled data augmentation, and unlabeled strong-weak augmentation.

	Unlabeled Data Views Used	COCO Partially Labeled 1%	2%	5%	10%
Supervised Baseline	None	12.15±0.27	16.65±0.18	21.45±0.16	27.10±0.07
Soft Teacher [11]	$\{1\times\}$	20.46±0.39	-	30.74±0.08	34.04±0.14
SED [3]	$\{1\times, 0.5\times\}$	-	-	29.01	34.02
PseCo [4]	$\{1\times, 0.5\times\}$	22.43±0.36	27.77±0.18	32.50±0.08	36.06±0.24
MixTeacher-RD (Ours)	$\{1\times, 0.5\times\}$	23.61±0.38	28.45±0.16	33.64±0.12	36.57±0.20
MixTeacher (Ours)	$\{1\times, 1\times, 0.5\times\}$	**25.16±0.26**	**29.11±0.21**	**34.06±0.13**	**36.72±0.16**

Table 3. Comparison with state-of-the-art methods on COCO benchmark. $AP_{50:95}$ on `val2017` set are reported. Under the Partially Labeled setting, results are the average of all five folds and numbers behind ± indicate the standard deviation. Under the Additional setting, numbers in front of the arrow indicate the supervised baseline. The views of unlabeled image used in each iteration are reported as well.

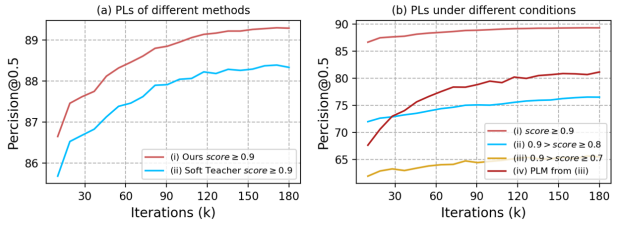

Figure 3. Comparison of the quality of pseudo labels during training. (a) Compare MixTeacher against Soft Teacher [11] under 10% labeling ratio. (b) Compare the pseudo labels of MixTeacher under different conditions. The pseudo labels with IoU overlapping the ground truth greater than 0.5 are regarded as true positives

scale pyramid. On the other hand, the smaller cow shows a higher response in the regular scale, but it is not appropriate to detect in this level due to its size. Thus, a lower γ that tends to use the down-sampled scale is appropriate.

5. Quality of Pseudo Labels

We further investigate the quality of pseudo labels during training. We evaluate the pseudo labels over 5,000 unlabeled images every 10k iterations for all methods. Figure 3 (a) shows the precision of pseudo labels for the proposed MixTeacher and baseline with the same score threshold of 0.9. As the results show in 3 (a), our method obviously produces more accurate pseudo-labels, and thus achieves more accurate results in the end. Figure 3 (b) shows the precision of the pseudo labels in different range of confidence score and the pseudo labels mined by our PLM module. Compared with all the pseudo labels with thresholds in $[0.7, 0.9)$, the pseudo labels mined by our method also have higher accuracy.

References

[1] Binbin Chen, Weijie Chen, Shicai Yang, Yunyi Xuan, Jie Song, Di Xie, Shiliang Pu, Mingli Song, and Yueting Zhuang. Label matching semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14381–14390, 2022. 2

[2] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 303–312, 2021. 2

[3] Qiushan Guo, Yao Mu, Jianyu Chen, Tianqi Wang, Yizhou Yu, and Ping Luo. Scale-equivalent distillation for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 3

[4] Gang Li, Xiang Li, Yujie Wang, Yichao Wu, Ding Liang, and Shanshan Zhang. PseCo: Pseudo Labeling and Consistency Training for Semi-Supervised Object Detection. In *European Conference on Computer Vision*, 2022. 1, 2, 3

[5] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*, 33:21002–21012, 2020. 2

[6] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[7] 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. *arXiv preprint arXiv:2102.09480*, 2021. 2

[8] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 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*, pages 658–666, 2019. 2

[9] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020. 2

[10] 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. 2

[11] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3060–3069, 2021. 1, 2, 3

[12] Lei Zhang, Yuxuan Sun, and Wei Wei. Mind the gap: Polishing pseudo labels for accurate semi-supervised object detection. *arXiv preprint arXiv:2207.08185*, 2022. 2

[13] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 9759–9768, 2020. 2

[14] Xiaosong Zhang, Fang Wan, Chang Liu, Rongrong Ji, and Qixiang Ye. Freeanchor: Learning to match anchors for visual object detection. *Advances in neural information processing systems*, 32, 2019. 2