# Label Matching Semi-Supervised Object Detection (Supplementary Materials)

Binbin Chen[2], Weijie Chen[1,2,], Shicai Yang[2], Yunyi Xuan[2], Jie Song[1]
Di Xie[2], Shiliang Pu[2], Mingli Song[1], Yueting Zhuang[1,]
[1]Zhejiang University, [2]Hikvision Research Institute
{chenbinbin8,chenweijie5,yangshicai,xuanyunyi,xiedi,pushiliang.hri}@hikvision.com
{sjie,songml,yzhuang}@zju.edu.cn

## A. Consistent Class Distribution Assumption

LabelMatch is based on the assumption that consistent class distribution exists between the labeled and unlabeled data since they are drawn from the same data distribution. To further verify this hypothesis, we present the comparisons between the labeled and unlabeled data in COCO-standard and VOC using the ground-truth labels. As shown in Fig. 1, the foreground-foreground class distribution and the foreground-background ratio of the unlabeled data are close to those of the labeled data in these SSOD settings.

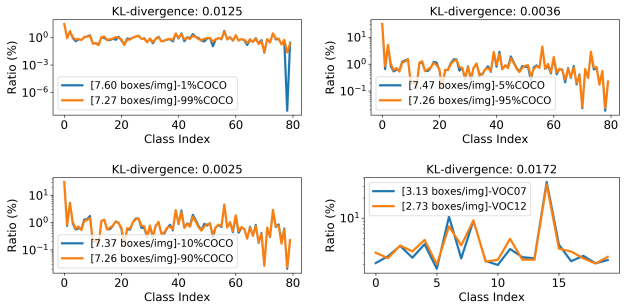



Figure 1. Comparisons on class distribution between the labeled and unlabeled data. The blue and orange lines denote the foreground-foreground class distribution in the labeled and unlabeled data, respectively. "boxes/img" in the legend represents the foreground-background ratio.

## B. More Results on COCO-standard

In this section, we present more experimental results on COCO-standard using the ablation study setting (see the fifth column in Tab. 9). Firstly, we carry out more analysis about ACT in Appendix B.1. Then, we study the effect of hyper-parameter in RPLM in Appendix B.2 and more analysis about proposal self-assignment in Appendix B.3. Finally, more qualitative results are exhibited in Appendix B.4.

### B.1. Analysis of ACT

In this part, we present more analysis about the proposed ACT from flexibility and implementation.

**Flexibility.** To further demonstrate the flexibility of our method, we extend STAC [13] with the proposed ACT, denoted as STAC$^\star$ for short. The original STAC first uses a pretrained model to generate pseudo labels and then uses a threshold of 0.9 to filter out low-quality pseudo labels, which are finally fed back into the network with strong data augmentation for model fine-tuning. Alternatively, STAC$^\star$ replaces the fixed threshold with the proposed ACT for pseudo labeling and updates the thresholds every epoch. Since there is no mean teacher in STAC, the label assignment strategy of STAC$^\star$ simply follows the *ignore assignment*, where uncertain pseudo labels are set as ignore labels.

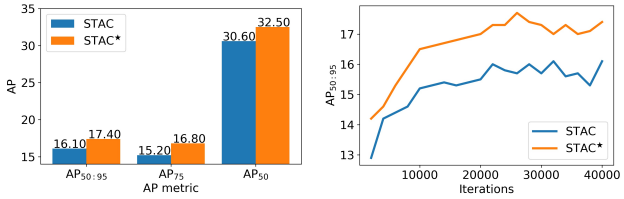
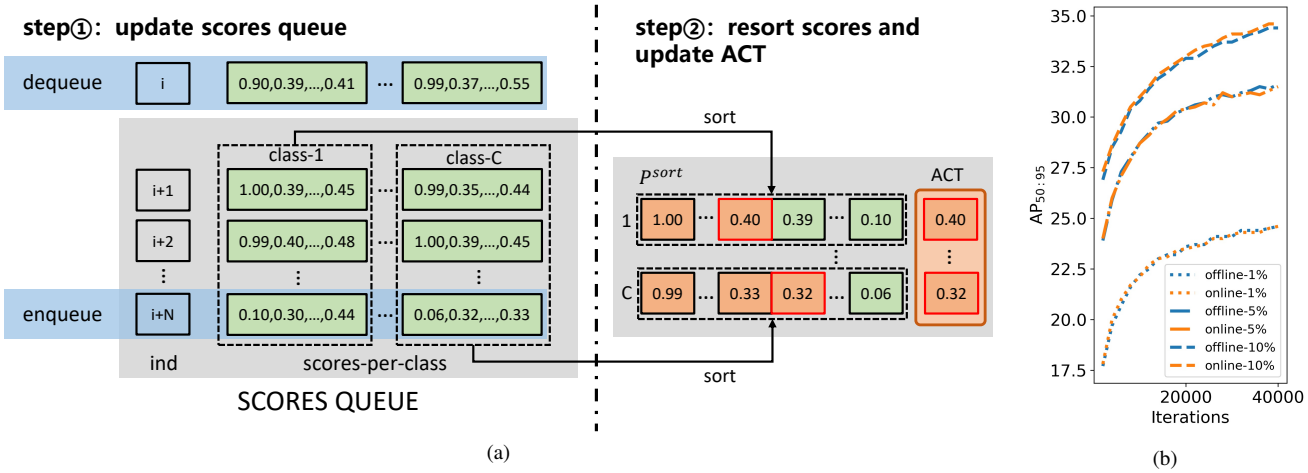



Figure 2. Performance comparisons between STAC and STAC$^\star$ on COCO-standard with 1% labeled data.

As shown in Fig. 2, there is an apparent performance gain after equipping STAC with ACT, demonstrating the universality of the proposed ACT.

**Online vs. Offline.** As discussed in the paper, ACT are updated to the evolved teacher during the training phase, avoiding a negative bias caused by the outdated predictions. There are two patterns to update ACT, one of which is introduced in the paper, leveraging a subset of unlabeled data to update ACT every $K$ iterations, termed as offline version. Here, we describe another pattern, named as online version,

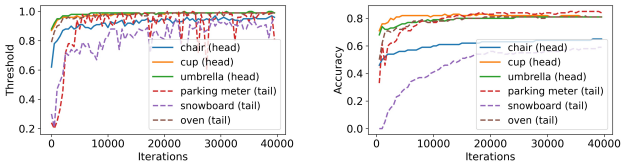



(a) (b)

Figure 3. (a) Implementation of the online version of ACT. Each training iteration consists of two steps: (1) obtain predictions from the teacher model and update the scores queue; (2) re-sort scores and update the ACT in real-time. (b) Performance comparison between the online and offline version of ACT during the training phase on three COCO-standard settings with 1%, 5% and 10% labeled data.

| | ACT | iterations | 1% | 5% | 10% |
|---|---|---|---|---|---|
| LabelMatch | Offline | 40K | 24.6 | 31.6 | 34.6 |
| LabelMatch | Online | 40K | 24.6 | 31.5 | 34.6 |

Table 1. Performance ($AP_{50:95}$) comparisons between the online and offline versions of ACT. We only run 1-fold using the ablation training setting.

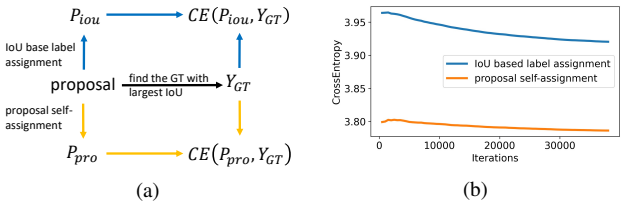



(a) (b)

Figure 4. (a) Thresholds in the training phrase. (b) The quality of reliable pseudo labels.

| $(T_{score}, T_{iou})$ | $AP_{50:95}$ | $(T_{score}, T_{iou})$ | $AP_{50:95}$ | $(T_{score}, T_{iou})$ | $AP_{50:95}$ |
|---|---|---|---|---|---|
| (0.7, 0.7) | 34.4 | (0.8, 0.7) | 34.4 | (0.9, 0.7) | 34.4 |
| (0.7, 0.8) | 34.5 | (0.8, 0.8) | **34.6** | (0.9, 0.8) | 34.5 |
| (0.7, 0.9) | 34.5 | (0.8, 0.9) | **34.6** | (0.9, 0.9) | 34.3 |

Table 2. Effect of hyper-parameters in RPLM.




(a) (b)

Figure 5. (a) The solution to measure the quality of predictions. (b) The quality comparison of the predictions between two label assignment methods in the training phase (lower is better).

which maintains a scores queue, as shown in Fig. 3a. The teacher's prediction is pushed into the scores queue for refreshing the ACT in each training iteration, which can be seen as a special case of the offline version with $K = 1$. Both versions of ACT can get satisfactory performance, as shown in Fig. 3b and Tab. 1. We use the offline version in all the experiments and will release the online version as well.

**Thresholds evolve alone training.** We select three head classes and three tail classes on offline version for analysis. As shown in Fig. 4, the thresholds (Eq.6 in the paper to filter reliable pseudo labels) increase in both head and tail classes during optimization. Specifically, the thresholds for tail classes are more fluctuated than those for head classes due to the scarce samples. Also, the quality of reliable pseudo labels get increased as training goes on.

## B.2. Analysis of RPLM Hyper-Parameter

There are two hyper-parameters $(T_{score}, T_{iou})$ in the component of reliable pseudo label mining (RPLM). Here we use COCO-standard with 10% labeled data as the experimental setting. As shown in Tab. 2, the best performance appears when $(T_{score}, T_{iou}) = (0.8, 0.8)$. Therefore, we use $(T_{score}, T_{iou}) = (0.8, 0.8)$ by default in all experiments throughout the paper. It is also worth mentioning that our method is not sensitive to these hyper-parameters.

## B.3. Analysis of proposal self-assignment

To further analyze the quality of the teacher's RoI head predictions on the student's proposals (proposals self-


Figure 6. Qualitative comparisons between the single confidence threshold and the proposed LabelMatch. Red rectangles highlight the false negatives, and yellow rectangles highlight the false positives. The score threshold for visualization is 0.6.

| Data Split | Normal→Foggy | Small→Large | Across cameras | Synthetic→Real |
|---|---|---|---|---|
| labeled data | Cityscapes (train) | Cityscapes (train) | KITTI | Sim10K |
| unlabeled data | Cityscapes-foggy (train) | BDD100K (train) | Cityscapes (train) | Cityscapes (train) |
| test data | Cityscapes-foggy (val) | BDD100K (val) | Cityscapes (val) | Cityscapes (val) |

Table 3. Four differnt domain shifts in DA-OD, which are contructed by five different datasets, including Cityscapes [2], Cityscapes-foggy [11], KITTI [11], Sim10k [5] and BDD100K [5].

| Method | truck | car | rider | person | train | motor | bicycle | bus | mean |
|---|---|---|---|---|---|---|---|---|---|
| Source only | 19.2 | 47.9 | 40.8 | 34.8 | 7.8 | 24.2 | 36.0 | 36.4 | 30.9 |
| CVPR2020:GPA [16] | 24.7 | 54.1 | 46.7 | 32.9 | 41.1 | 32.4 | 38.7 | 45.7 | 39.5 |
| CVPR2020:HTCN [1] | 31.6 | 47.9 | 47.5 | 33.2 | 40.9 | 32.3 | 37.1 | 47.4 | 39.8 |
| CVPR2021:MeGA [14] | 25.4 | 52.4 | 49.0 | 37.7 | 46.9 | 34.5 | 39.0 | 49.2 | 41.8 |
| CVPR2021:UMT [3] | 34.1 | 48.6 | 46.7 | 33.0 | 46.8 | 30.4 | 37.3 | 56.5 | 41.7 |
| LabelMatch (Ours) | **42.0** | **62.2** | **55.4** | **45.3** | **55.1** | **43.5** | **51.5** | **64.1** | **52.4** |

Table 4. Results of adaptation from normal to foggy weathers. "Source only" refers to the model trained by labeled source data.

| Method | truck | car | rider | person | train | motor | bicycle | bus | mean |
|---|---|---|---|---|---|---|---|---|---|
| Source only | 18.3 | 50.0 | 33.3 | 35.8 | - | 18.4 | 27.6 | 17.0 | 28.7 |
| CVPR2019:SW-Faster [10] | 15.2 | 45.7 | 29.5 | 30.2 | - | 17.1 | 21.2 | 18.4 | 25.3 |
| CVPR2020:CR-DA [15] | 19.5 | 46.3 | 31.3 | 31.4 | - | 17.3 | 23.8 | 18.9 | 26.9 |
| LabelMatch (Ours) | 39.4 | 54.6 | 37.4 | 42.9 | - | 25.7 | 29.8 | 41.7 | 38.8 |
| LabelMatch† (Ours) | **39.8** | **55.4** | **44.5** | **44.8** | - | **38.6** | **41.5** | **47.1** | **44.5** |

Table 5. Results of adaptation from small to large scale datasets. † is an ideal setting that uses the ground-truth labels of the unlabeled data for class distribution estimation.

assignment *vs.* IoU-based label assignment), we use the ground truth for quantitative measurement. For each proposal, we calculate the cross-entropy between the corresponding prediction and the nearest ground truth (set as background if IoU<0.5). As shown in Fig. 5, the predictions by proposal self-assignment show better quality than IoU-based one.

### B.4. Qualitative Results

We perform the qualitative comparisons between the proposed method and the mean teacher frameworks with a fixed and single confidence threshold (varying from 0.7 to 0.9). As shown in Fig. 6, there are many false positives with a low confidence threshold (yellow rectangles in the second column), while many false negatives appear when using a high confidence threshold (red rectangles in the fourth column). Although the manual search threshold (0.8) via trial-and-error can achieve satisfactory results, our method shows even better qualitative results.

## C. Domain Adaptive Object Detection

LabelMatch is based on the consistent class distribution assumption between the labeled and unlabeled data. To ex-

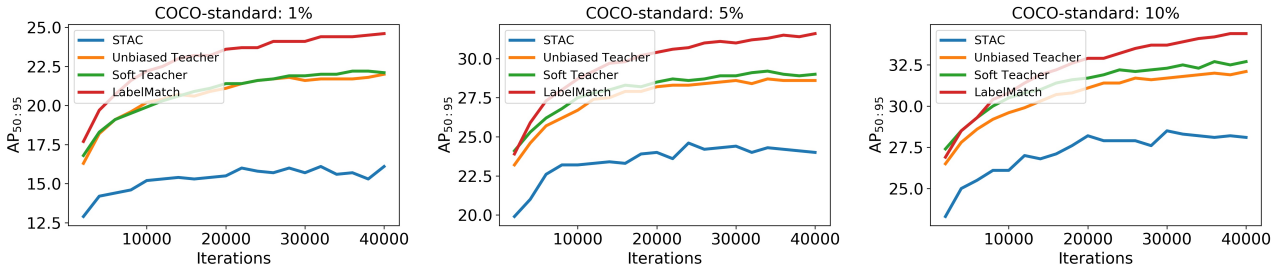



Figure 7. Performance ($AP_{50:95}$) comparisons among different state-of-the-art SSOD methods with exactly the same training settings.

| Method | $AP_{50}$ | network |
|---|---|---|
| Source only | 42.2 | FR+VGG |
| CVPR2019:SW-Faster [10] | 37.9 | FR+VGG |
| CVPR2020:GPA [16] | 47.9 | FR+R50 |
| CVPR2021:MeGA [14] | 43.0 | FR+VGG |
| ICCV2021:SimROD [9] | 47.5 | YOLOv5 |
| LabelMatch (Ours) | 51.0 | FR+VGG |
| LabelMatch† (Ours) | **52.2** | FR+VGG |

Table 6. Results of adaptation across cameras. FR: Faster-RCNN. † is an ideal setting that uses the ground-truth labels of the unlabeled data for class distribution estimation.

| Method | $AP_{50}$ | network |
|---|---|---|
| Source only | 36.5 | FR+VGG |
| CVPR2019:SW-Faster [10] | 40.7 | FR+VGG |
| CVPR2020:GPA [16] | 47.6 | FR+R50 |
| CVPR2021:MeGA [14] | 44.8 | FR+VGG |
| ICCV2021:SimROD [9] | 52.1 | YOLOv5 |
| LabelMatch (Ours) | 52.7 | FR+VGG |
| LabelMatch† (Ours) | **53.8** | FR+VGG |

Table 7. Results of adaptation from synthetic to real. VGG: VGG-16. † is an ideal setting that uses the ground-truth labels of the unlabeled data for class distribution estimation.

| Method | Loss | Threshold | 1% | 5% | 10% |
|---|---|---|---|---|---|
| STAC [13] | Cross-Entropy | 0.9 | 16.1 | 24.0 | 28.1 |
| Unbiased Teacher [8] | Focal-Loss | 0.7 | 22.0 | 28.6 | 32.1 |
| Soft Teacher* [17] | Cross-Entropy | 0.9 | 22.1 | 29.0 | 32.7 |
| LabelMatch (Ours) | Cross-Entropy | ACT | **24.6** | **31.5** | **34.6** |

Table 8. Benchmark results on COCO-standard: our re-implementations with exactly the same training details and data augmentation strategies. * denotes the re-implementation without box-jitter trick. We only run 1-fold using the ablation training setting due to the limitation of computation resources.

plore the robustness of LabelMatch to the prior dependence on this assumption, we extend it to the scenario of domain adaptive object detection (DA-OD) [6, 10, 14, 16] where the labeled source data and the unlabeled target are drawn from two different data distributions.

**Dataset.** As described in Tab. 3, following the existing DA-OD works, there are four common types of domain shifts in DA-OD. We evaluate our method on these settings and compare it with the state-of-the-arts.

**Network Architecture.** For a fair comparison with the existing DA-OD arts, we switch the backbone from ResNet-50 [4] to VGG-16 [12] and remove the FPN [7] neck.

**Implementation Details.** The implementation is nearly the same as SSOD, and more training hyper-parameters can be found in Appendix E. Following previous works, we use $AP_{50}$ as our evaluation metric.

**Results.** To examine the prior dependence on the consistent class distribution assumption, we evaluate LabelMatch in two class distribution estimation manners: 1) The first one is the same as described in the main body of the paper, which estimates the class distribution of the unlabeled target data by the annotations of the labeled source data; 2) The second one is an ideal setting, which determines the class distribution of the unlabeled target data by the ground-truth labels of the unlabeled data.

- Normal→Foggy: This scenario is different from the following DA-OD settings. In this scenario, the labeled source data and the unlabeled target data meet exactly the same class distribution since the target foggy data is rendered from the normal source data via a foggy translation model. As shown in Tab. 4, benefited from the given class distribution, we achieve a +21.5 mAP improvement over the "source only" baseline, exceeding previous state-of-the-arts by a large margin.

- Small→Large: Although there exists bias between the labeled and unlabeled data on foreground-foreground class distribution ($KL = 0.36$) and foreground-background ratio (18.5 boxes/img vs. 13.9 boxes/img), our method can still achieve 38.8 mAP, surpassing all the previous arts as far as we know. With access to the accurate class distribution (the ideal setting), our method can be further improved to 44.5 mAP.

- Across cameras & Synthetic→Real: In these settings, there is only one foreground class and exits foreground-background ratio bias (4.3 boxes/img vs. 9.6 boxes/img and 5.8 boxes/img vs. 9.6 boxes/img). Even using a biased class distribution, our method can still achieve satisfactory results. And our method can get further improvement equipped with the accurate class distribution (aka the ideal setting).

These DA-OD experiments demonstrate the robustness of the proposed LabelMatch framework, since the introduction of proposal self-assignment and RPLM weaken the prior dependence on the consistent class distribution assumption. From another perspective, these experiments also indicate that an accurate class distribution estimation can

| training setting | COCO-standard | COCO-additional | VOC | Ablation | DA-OD |
|---|---|---|---|---|---|
| batch size for labeled data | 16 | 32 | 4 | 32 | 16 |
| batch size for unlabeled data | 16 | 32 | 4 | 32 | 16 |
| learning rate | 0.01 | 0.02 | 1.25e-3 | 0.02 | 0.016 |
| learning rate step | - | (360K, 480K) | - | - | - |
| iterations | 160K | 540K | 160K | 40K | 20K |
| unsupervised loss weight $\lambda$ | 2.0 | 2.0 | 2.0 | 2.0 | 2.0 |
| EMA rate | 0.996 | 0.996 | 0.996 | 0.996 | 0.9996 |
| reliable ratio $\alpha$ | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 |
| mean score thresh $T_{score}$ | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 |
| mean iou thresh $T_{iou}$ | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 |
| multi-scale (strong augmentation) | (0.2, 1.8) | (0.2, 1.8) | (0.2, 1.8) | (0.5, 1.5) | (0.5, 1.5) |
| test score thresh | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |

Table 9. Training settings for different datasets and different tasks. "Ablation" means the training setting of the ablation studies in the main body of the paper, which is also used in all SSOD experiments in the Appendix.

| Weak Augmentation | | | |
|---|---|---|---|
| Process | Prob | Parameters | Descriptions |
| Horizontal Flip | 0.5 | None | None |
| Multi-Scale | 1.0 | scale=(500, 800) | The short edge of image is random resized from 500 to 800. |
| **Strong Augmentation** | | | |
| Process | Prob | Parameters | Descriptions |
| Horizontal Flip | 0.5 | None | None |
| Multi-Scale | 1.0 | ratio=(0.2, 1.8) | The short edge of image is random resized from $0.5l_{short}$ to $1.5l_{short}$. |
| Color Jittering | 0.8 | (brightness, contrast, saturation, hue) = (0.4, 0.4, 0.4, 0.1) | Brightness factor is chosen uniformly form [0.6, 1.4], contrast factor is chosen uniformly from [0.6, 1.4], saturation factor is chosen uniformly from [0.6, 1.4], and hue value is chosen uniformly from [-0.1, 0.1]. |
| Grayscale | 0.2 | None | None |
| GaussianBlur | 0.5 | (sigma_x, simga_y)=(0.1, 2.0) | Gaussian filter with $\sigma_x = 0.1$ and $\sigma_y = 2.0$ is applied |
| CutoutPattern1 | 0.7 | scale=(0.05, 0.2), ratio=(0.3, 3.3) | Randomly selects a rectangle region in an image and erases its pixels. |
| CutoutPattern2 | 0.7 | scale=(0.02, 0.2), ratio=(0.1, 6.0) | Randomly selects a rectangle region in an image and erases its pixels. |
| CutoutPattern3 | 0.7 | scale=(0.02, 0.2), ratio=(0.05, 8.0) | Randomly selects a rectangle region in an image and erases its pixels. |

Table 10. Details of data augmentations. In our ablation study, we use multi-scale with ratio=(0.5, 1.5) in order to use large batch size.

further promote the performance of DA-OD, emphasizing the importance of class distribution estimation. How to estimate an accurate class distribution when the labeled data and the unlabeled data are drawn from two different data distributions is an interesting future work.

## D. MMDetection-based SSOD Codebase

Since different SSOD algorithms use different data augmentation strategies which have great impact on the performance, we build a unified MMDetection-based SSOD codebase for a fair comparison, named MMDet-SSOD for short, containing STAC [13], Unbiased-Teacher [8], Soft-Teacher [17] and LabelMatch.

We comprehensively run all algorithms in our MMDet-SSOD on COCO-standard dataset using the ablation training setting, and report the performance in Tab. 8 and Fig. 7. It is worth mentioning that the data augmentation, training iterations, batch size, and other training settings are all kept the same among these algorithms for a fair comparison. The entire source code will be released soon to support the development of SSOD in the community.

## E. Implementation and Training Details

**Training.** We utilize different training settings for different datasets in our implementation. We use the SGD optimizer with a momentum rate 0.9 and weight decay 0.0001 in all experiments. The different training settings are summarized in Tab. 9.

**Data augmentation.** Our data augmentation strategies are modified from Unbiased Teacher [8], and the details are shown in Tab. 10. The weak augmentation is applied to the unlabeled data for pseudo labeling, and the strong augmentation is applied to both labeled and unlabeled data for model training. In our implementation, no cutout augmentation is applied to the labeled data when using strong data augmentation. In order to save computation resources, we use multi-scale with ratio=(0.5, 1.5) in the ablation studies.

## References


[1] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8869–8878, 2020. 3

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016. 3

[3] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4091–4101, 2021. 3

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[5] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 746–753, 2017. 3

[6] Xianfeng Li, Weijie Chen, Di Xie, Shicai Yang, Peng Yuan, Shiliang Pu, and Yueting Zhuang. A free lunch for unsupervised domain adaptive object detection without source data. In *AAAI*, 2021. 4

[7] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017. 4

[8] 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 *ICLR 2021: The Ninth International Conference on Learning Representations*, 2021. 4, 5, 6

[9] Rindra Ramamonjison, Amin Banitalebi-Dehkordi, Xinyu Kang, Xiaolong Bai, and Yong Zhang. Simrod: A simple adaptation method for robust object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3570–3579, 2021. 4

[10] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6956–6965, 2019. 3, 4

[11] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018. 3

[12] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *Computer Science*, 2014. 4

[13] 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, 4, 5

[14] Vibashan Vs, Vikram Gupta, Poojan Oza, Vishwanath A. Sindagi, and Vishal M. Patel. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4516–4526, 2021. 3, 4

[15] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11724–11733, 2020. 3

[16] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12355–12364, 2020. 3, 4

[17] 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. 4, 5