Dual-Curriculum Teacher for Domain-Inconsistent Object Detection in Autonomous Driving

Longhui Yu[1]
yulonghui@stu.pku.edu.cn
Yifan Zhang[2]
yifan.zhang@u.nus.edu
Lanqing Hong[3†]
honglanqing@huawei.com
Fei Chen[3]
chen.f@huawei.com
Zhenguo Li[3]
li.zhenguo@huawei.com

[1] Peking University, Beijing, China
[2] National University of Singapore, Singapore, Singapore
[3] Huawei Noah's Ark Lab, Hong Kong, China

Appendix

We organize the Appendix as follows:

- In Appendix 1, we add some additional explanation for the Domain-Inconsistent Object Detection.
- In Appendix 2, we introduce the mutual learning framework used in our DucTeacher.
- In Appendix 3, we demonstrate the implementation details about DucTeacher.
- In Appendix 4, we analyze the influence of the hyper-parameter, thresholds τ and scale factors μ, introduced in the DucTeacher.
- In Appendix 5, we show more results about SODA10M.
- In Appendix 6, we explain how DucTeacher can show a good cross-domain generalization ability.

1 Domain-Inconsistent Object Detection Setting

As exhibited in the main paper, the targeted setting, Domain-Inconsistent Object Detection is different from the Classical Semi-Supervised Object Detection. Figure 1 shows the difference for detail. In the Classical Semi-Supervised Object Detection, Labeled data and

† Lanqing Hong is the corresponding author.

Unlabeled data are from the same data distribution. However, in the Domain-Inconsistent Semi-Supervised Object Detection, Labeled Data and Unlabeled data can from different different distribution. This distribution shift would cause two detailed challenges hindering the learning of unlabeled data and this influence analysis has elaborated in the main paper.

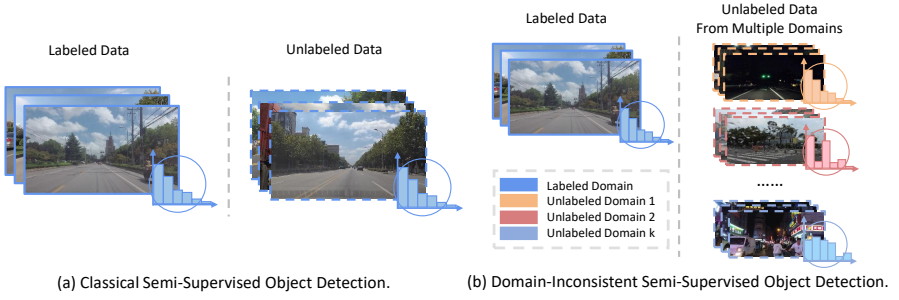

Figure 1: Illustration of domain-inconsistent semi-supervised object detection (SSOD). (a) Existing SSOD often considers labeled data and unlabeled data from the same data distribution. (b) Domain-inconsistent SSOD aims to tackle the problem with both data distribution shifts and class distribution shifts between labeled and unlabeled data.

2 Mutual Learning Framework

Existing state-of-the-art SSOD methods [7, 9, 12] usually adopt a Teacher-Student Mutual Learning framework. Similar to the knowledge distillation [5, 11], a student model is partially supervised by the teacher model and is trained with combined loss function $\mathcal{L} = \mathcal{L}_s + \mathcal{L}_u$ with the supervised loss $\mathcal{L}_s$ and the unsupervised loss $\mathcal{L}_u$,

$$\mathcal{L}_s = \frac{1}{N_l}\sum_i \mathcal{L}_{cls}(x_i^l, y_i^l) + \mathcal{L}_{reg}(x_i^l, y_i^l), \tag{1}$$

$$\mathcal{L}_u = \frac{1}{N_u}\sum_i \mathcal{L}_{cls}(x_i^u, y_i^u) + \mathcal{L}_{reg}(x_i^u, y_i^u), \tag{2}$$

where $\mathcal{L}_{cls}$ and $\mathcal{L}_{reg}$ represent the classification loss and regression loss, respectively, y_i^l is the annotation of the labeled image x_i^l, and y_i^u is the pseudo-labels generated by teacher model on unlabeled data x_i^u. Before the Teacher-Student Mutual Learning stage, the teacher model is pre-trained on the labeled set $\mathcal{D}_L$, while in the mutual learning stage, the teacher model is updated by exponential moving average (EMA) mechanism.

The quality of pseudo-labels y_i^u is important for SSOD. Nevertheless, in domain-inconsistent SSOD, the input data distribution shifts and class distribution shifts would cause the teacher model to produce inaccurate and biased pseudo-labels y_i^u. To tackle this problem, we proposed two curricula, DEC and DMC, to provide reliable pseudo-labels y_i^u.

DucTeacher is based on the teacher-student mutual learning framework and the consistency pseudo-labeling strategy, which are also included in existing state-of-the-art SSOD methods. We introduce the technical details as follows.

Teacher-Student Mutual Learning. In the mutual learning stage, the student model is trained with supervision of the ground truths and the pseudo-labels. The student model is

updated by gradient descent, while the teacher model is updated by exponential moving average (EMA) mechanism,

$$\theta_s \leftarrow \theta_s + \frac{\partial \mathcal{L}}{\partial \theta_s}, \tag{3}$$

$$\theta_t \leftarrow \alpha\theta_t + (1-\alpha)\theta_s, \tag{4}$$

where θ_t, θ_s represent the parameter of teacher model and student model.

Consistency Pseudo-Labeling. Consistency Pseudo-Labeling produces pseudo-label based on both the consistency regularization and pseudo-labeling. It produces a pseudo-label on a weakly-augmented unlabeled image and screens out the pseudo-label with the high confidence score, which would be used as a target for the model fed with a strongly-augmented version of the same image. The detailed data augmentations used in this work are shown in Table 1.

Table 1: Used augmentations in DucTeacher.

Weak Augmentation		
Process	Probability	Parameters
Horizontal Flip	0.5	-
Strong Augmentation		
Process	Probability	Parameters
Grayscale	0.2	-
GaussianBlur	0.5	(sigma x, sigma y) = (0.1, 2.0)
CutoutPattern1	0.7	scale=(0.05, 0.2), ratio=(0.3, 3.3)
CutoutPattern2	0.5	scale=(0.02, 0.2), ratio=(0.1, 6)
CutoutPattern3	0.3	scale=(0.02, 0.2), ratio=(0.05, 8)

3 Implementation details

We implement our DucTeacher based on Detectron2 [10]. For a fair comparison, we follow STAC [9] and Unbiased Teacher [7] to use Faster-RCNN with FPN [5] and ResNet-50 backbone [2] as the object detector. We adopt the teacher-student mutual learning framework as Unbiased Teacher [7], which trains the student model with Focal loss [6] and updates the teacher model with EMA. The EMA rate α is 0.9996. The base pre-defined threshold in DucTeacher τ is set as 0.7, and the scale factor μ is set as 0.1. The learning rate is set as 0.01, and the max training iteration is set as 160k. The batch size of training data is set as 32 (16 for both labeled and unlabeled) for both the SODA10M and COCO datasets. The pre-trained model obtained in DucTeacher is trained on the labeled domain $\mathcal{D}_l$ with 2k iterations. We conduct experiments with 8 Nvidia V100 GPU (32GB) cards with Intel Xeon Platinum 8168 CPU (2.70GHZ).

Pre-train Stage. For SODA10M, the Pre-train stage is using the labeled domain D_l to get an initial model. For COCO, following the Unbiased Teacher [27], we use a small amount of labeled data to pre-train a detector with 2000 iterations for fair comparisons.

DEC for SODA10M. For SODA10M, according to the similarity score provided by DEC, we divide the whole unlabeled training set into 4 subsets. The earlier trained subset has a higher similarity score than that of subset trained later.

Adapting to the COCO. Unlike the SODA10M dataset, there is no domain label in COCO [4]. To adapt the DucTeacher for classical SSOD on COCO, we modify the proposed DEC. First, we also pre-train a model only with the labeled data until 2,000 iterations then use

the pre-trained model to evaluate the unlabeled data and obtain the average bboxes score for each image. After that, we sort the unlabeled data in descending order according to the average bboxes score and divide the unlabeled data into different phases. Unlabeled data in each phase can be regarded as a similar difficulty degree. Also, we use the class distribution of labeled data as the target for distribution matching.

4 Hyper-parameters

In this appendix, we study the hyper-parameters introduced by the DucTeacher, including the threshold τ and the scale factor μ. We report the results in Table 2, which gives the following observations. When the threshold τ and scale factor μ are set as 0.7 and 0.1, respectively, DucTeacher achieves the best performance (48.4mAP). For the threshold τ, too low or too high threshold τ would both damage the performance, as a low threshold would cause noisy pseudo-labels and a high threshold would discard much useful information. For the scale factor μ, too small scale factor μ would cause the dynamic thresholding strategy in DucTeacher hard to exert its effectiveness since too small scale factor μ would cause the thresholds for different categories in different domains almost the same, which cannot dynamically filter biased pseudo-labels. Too large scale factor μ would also cause an unstable dynamic threshold then affect performance. Since DucTeacher accumulates the pseudo-label distribution according to each training iteration, too large scale factor μ would cause DucTeacher dependent on the sampling of each batch too much then affect the performance.

Table 2: Ablation study on the effects of different pre-defined thresholds τ and different pre-defined scale factors μ.

τ	mAP	AP_{50}	AP_{75}	μ	mAP	AP_{50}	AP_{75}
0.6	45.9	70.6	50.0	0.05	47.1	71.8	51.2
0.7	48.4	73.5	52.4	0.10	48.4	73.5	52.4
0.8	44.7	69.3	48.5	0.15	46.9	71.4	50.9

Table 3: The cross-domain generalization ability of weak-stong augmentation and EMA mechanism. The table shows the ablation studies about mAP performance on weak-stong augmentation (DucTeacher w/o aug) and EMA mechanism (DucTeacher w/o EMA).

Method	Overall	Daytime	Night
Supervised-only	37.9	43.1	21.1
UMT	44.7	45.1	35.9
MT-MTDA	45.2	47.1	37.1
DucTeacher	48.4	49.6	40.7
DucTeacher w/o aug	39.1	44.9	22.4
DucTeacher w/o EMA	35.7	42.1	16.3

5 More experimental results on SODA10M

Table 4 further shows that the proposed DucTeacher can improve the mAP performance for almost all the domains. Moreover, compared with the state-of-the-art cross-domain object

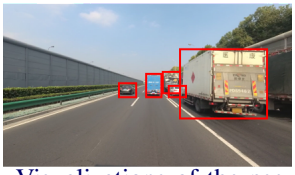


Figure 2: Visualizations of the pseudo-labels produced on similar and dissimilar domain. In the similar domain, there are fewer False Negatives errors (i.e., five ground truth objects versus five predicted bboxes). While in the dissimilar domain, objects are hard to detect and cause more False Negatives errors (five ground truth objects versus only two predicted bboxes).

Table 4: Comparison of mAP for different semi-supervised methods on SODA10M detailed in each domain. '-' means no validation image in this domain.

Model	Overall mAP	City street (Car)			Highway (Car)			Country road (Car)	
		Clear	Overcast	Rainy	Clear	Overcast	Rainy	Clear	Overcast
Daytime									
Supervised	43.1	70.0	64.9	56.6	68.3	65.9	65.9	69.4	63.5
STAC [0]	45.3$^{+2.2}$	74.2	69.6	58.0	71.7	70.3	70.7	75.2	69.8
UMT [0]	45.1$^{+2.0}$	73.4	67.5	56.9	68.5	68.7	68.2	70.2	64.7
MT-MTDA [8]	47.1$^{+4.0}$	71.8	66.0	52.9	68.3	67.8	69.8	74.5	67.5
Unbiased Teacher [0]	47.7$^{+4.6}$	73.0	68.1	55.3	69.1	62.0	71.3	72.6	70.0
DucTeacher (ours)	**49.6**$^{+6.5}$	76.7	68.5	55.6	69.5	70.0	71.6	73.5	69.1
Night									
Supervised	21.1	36.3	37.7	-	37.5	37.3	79.5	38.9	72.8
STAC [0]	28.2$^{+7.1}$	45.5	46.8	-	46.2	45.6	83.7	47.2	75.4
UMT [0]	35.9$^{+14.8}$	58.4	59.7	-	58.7	60.2	81.1	60.4	72.2
MT-MTDA [8]	37.1$^{+16.0}$	60.4	61.2	-	60.7	62.2	80.6	62.4	73.6
Unbiased Teacher [0]	39.7$^{+18.6}$	65.3	66.2	-	66.2	67.2	83.6	67.5	75.2
DucTeacher (ours)	**40.7**$^{+19.6}$	65.3	67.0	-	66.8	67.4	84.3	67.7	76.5

detection method UMT [0] and our implemented multi-target domain adaptation method MT-MTDA [8], the proposed method DucTeacher shows prominent superiority, which outperforms the UMT and MT-MTDA about 3.7mAP and 3.2mAP respectively, as shown in the main paper. Also, as shown in Table 4, it is interesting that our DucTeacher performs better than UMT and MT-MTDA on both the daytime domains (source) and the night domains (target). The key to the great performance of our DucTeacher on the multiple unlabeled domains is the "weak-strong data augmentation" and "EMA mechanism", where the ablation experiments are shown in appendix 6.

How dose domain shifts affect semi-supervised object detection ? In our main paper, the first thing is that models trained in a similar (easy) domain would perform poor in the dissimilar (hard) domain, which would cause the the produced pseudo-label contain lots of noise. Motivated by this, we design our DucTeacher which includes two curricula to handle the noise pseudo-label in two levels, the training order of data from different domains and the adjustment of pseudo-label thresholds for different domains. Except for these two perspectives, there is also an interesting distinctive problem in semi-supervised object detection, where models produce high-confidence pseudo-box in different levels for easy domains and hard domains. As shown in Figure 2, the high-confidence pseudo-box predicted in the similar (easy) domain is more than that in the dissimilar (hard) domain. The less predicted boxes would cause the False Negatives error in object detection, which causes that the model suppress the activation and tends to predict all the object as background.

6 Why dose DucTeacher have a good cross-domain generalization ability?

In the main paper, we have shown that DucTeacher has a good cross-domain generalization ability. In this appendix, we further analyze the reason behind it. In Table 3, we analyze the effectiveness of weak-strong augmentation and EMA mechanism for the cross-domain generalization ability of our DucTeacher. Compared with the supervised method, the main improvement of DucTeacher is in the Night domain, which is 19.6mAP improvement compared with the 6.5mAP improvement in the Daytime domain, as shown in Table 3. Moreover, DucTeacher also shows a higher cross-domain generalization performance compared with the existing state-of-the-art cross-domain object detection method UMT [1] and multi-target domain adaptation method MT-MTDA [8]. We find the reason for the high cross-domain generalization ability of our DucTeacher is the combination of weak-strong augmentation and exponential moving average (EMA) mechanism. As shown in Table 3, without weak-strong augmentation or EMA mechanism, the performance drops greatly, especially in the Night domain, where the decreasing performance is 18.3mAP and 24.4mAP. With the combination of weak-strong augmentation and EMA mechanism, DucTeacher produces the pseudo-labels on the weakly-augmented images and the pseudo-labels would be used for the strongly-augmented images. This weak-strong consistency regularization guarantees DucTeacher's high cross-domain generalization ability. Besides, the EMA mechanism gradually updates the DucTeacher model to construct a temporal ensembles model of student model in different steps, which can produce more reliable pseudo-labels for self-training.

References

[1] 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.

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

[3] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[5] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[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.

[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.

[8] Le Thanh Nguyen-Meidine, Atif Belal, Madhu Kiran, Jose Dolz, Louis-Antoine Blais-Morin, and Eric Granger. Unsupervised multi-target domain adaptation through knowledge distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1339–1347, 2021.

[9] 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.

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

[11] Longhui Yu, Zhenyu Weng, Yuqing Wang, and Yuesheng Zhu. Multi-teacher knowledge distillation for incremental implicitly-refined classification. *arXiv preprint arXiv:2202.11384*, 2022.

[12] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 4081–4090, 2021.