SIMULATION-AND-MINING: TOWARDS ACCURATE SOURCE-FREE UNSUPERVISED DOMAIN ADAPTIVE OBJECT DETECTION

Peng Yuan[2,⋆], *Weijie Chen*[1,2,⋆], *Shicai Yang*[1,2,⋆], *Yunyi Xuan*[2], *Di Xie*[2], *Yueting Zhuang*[1,✉], *Shiliang Pu*[2]

[1]Zhejiang University, [2]Hikvision Research Institute

ABSTRACT

Vanilla *unsupervised domain adaptive* (UDA) object detection typically requires the labeled source data for joint-training with the unlabeled target data, which is usually unavailable in real-world scenarios due to data privacy, leading to source data-free UDA object detection. Herein, we first analyze the phenomenon of cross-domain detection degradation varying from easy to hard samples (*e.g.* the objects with different scales or occlusion degrees), termed as domain generalization differentiation. In detail, the ability to detect easy samples is well transferred while the one to detect hard samples is dramatically degraded. To this end, we then revisit the existing self-training method, which is of great challenge to deal with the abundant false negatives (hard samples). Assumed that true positives (easy samples) labeled by the source model can be exploited as supervision cues. UDA is finally modeled into an unsupervised false negatives mining problem. Thus, we propose a Simulation-and-Mining (S&M) framework, which simulates false negatives by augmenting true positives and mines back false negatives alternatively and iteratively. Experimental results show the effectiveness.

Index Terms— Domain Adaptation, Self-Training, Object Detection, Domain Generalization Differentiation

1. INTRODUCTION

Deep convolutional neural networks have significantly improved the performance of object detection [1, 2, 3], but remain reliant on massive annotated data. When generalized to new environments with different data distributions, the annotated target data is unavailable for networks fine-tuning due to the high cost of annotations. To address this problem, UDA object detection [4] is proposed to mitigate the domain shift by transferring the knowledge from the semantic related source domain to the target domain. It soon becomes an attractive topic and many representative works have been proposed [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15], most of which focus on generating domain invariant representation by reducing cross-domain discrepancy. Moreover, Cai *et al.* [16] modifies the mean teacher paradigm [17] to reduce the domain gap via consistency regularization. Hsu *et al.* [18] and A.L. *et al.* [19] perform pixel-adaptation by a style transfer for target-like synthesis. Kim *et al.* [20] and Yang *et al.* [21] apply self-training framework to solve UDA problem.

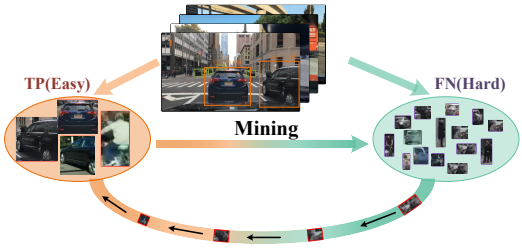


Fig. 1. True positives (TP) are leveraged as supervision cues to drive unsupervised false negatives (FN) mining during self-training for source data-free UDA object detection.

However, the aforementioned algorithms come to a practical limitation that there is usually no access to the labeled source data in many scenarios due to data privacy. Li *et al.* [22] first formulates the source data-free UDA object detection problem and proposes SFOD to search for a relatively good confidence threshold to drive self-training via pseudo-labeling. Despite the promising results, it still suffers from the quality of pseudo labels. We claim that the quality of pseudo labels is strongly related to the domain generalization performance in object detection tasks. In this work, we divide the objects with different visual appearances into two types: easy samples (e.g. large-scale) and hard samples (e.g. small-scale and occluded). Extensive quantitative evaluations on domain benchmarks show a dramatic degeneration in the detection of hard samples. We refer to this phenomenon as domain generalization differentiation, which deteriorates the performance of UDA object detection tasks by producing numerous false negatives during pseudo-labeling. As shown in Fig.3, more than 50% objects with scale less than 100 pixels are degenerated into background class when adapted to a new domain, which cannot be solved by an elaborate confidence threshold for pseudo labeling. This diagnosis motivates us to incorporate the false negatives suppression into pseudo labeling, instead of performed during inference in conventional supervised object detection tasks.

Based on the above analysis, we propose a simple yet effective Simulation-and-Mining (S&M) framework (Fig.2) which models the source data-free UDA object detection task

⋆ Equal Contributions. ✉ Corresponding Author.

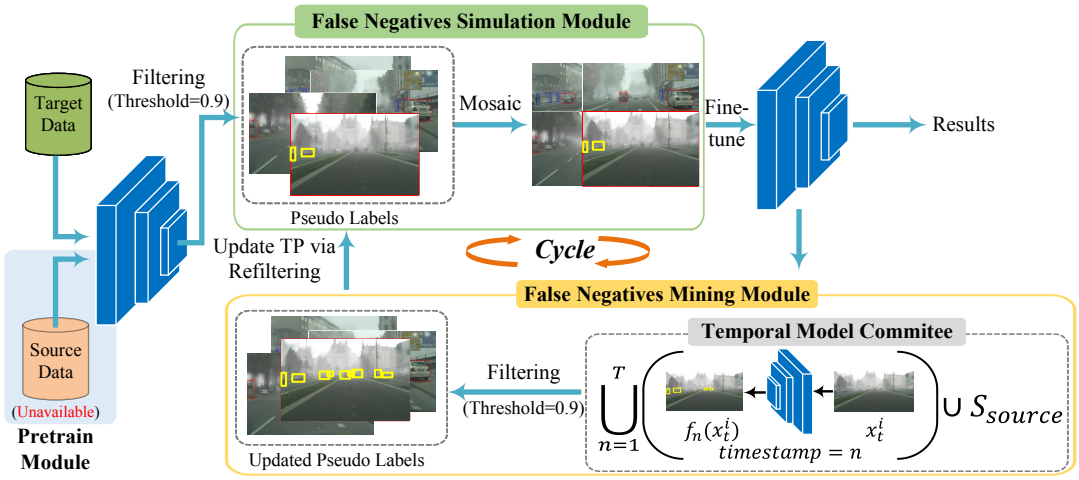


Fig. 2. An overview of S&M self-training framework equipped with False Negatives Simulation and False Negatives Mining.

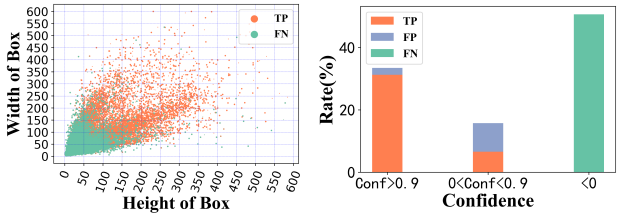


Fig. 3. Domain generalization differentiation in "Sim10k to Cityscapes". **Left**: The object scale distribution. **Right**: The ratio of true positives, false positives and false negatives.

into an unsupervised false negatives mining problem. Supposed that the pseudo boxes with very high confidence provided by the source domain pre-trained model can be treated as true positives. Considering the semantic similarity between true positives and false negatives, these pseudo boxes can be exploited as supervision cues, in turn, mining false negatives. We reason the failure of false negatives annotation for the appearance shift. For instance, some false negatives act as the occluded zoom-out transformation of the true positives. It inspires us to perform False Negatives Simulation (FNS) to minimize the visual appearance discrepancy by enforcing a strong data augmentation on true positives. This simple solution drives the self-training to attract false negatives while rejecting the true negatives. To push the training continuously digging out the missing knowledge, False Negatives Mining (FNM) is further introduced by updating true positives after FNS. Specially, we named this updating procedure as Temporal Model Committee (TMC), which combines the model experiences from different timestamps to generate pseudo boxes via re-filtering true positives. The processes of FNS and FNM are implemented alternatively in network optimization to drive an accurate self-training.

To summarize, our main contributions are listed as follows: **1)** We study the domain generalization differentiation phenomenon in object detection tasks, which inspires us to revisit and improve the existing self-training methods. **2)** We formulate source data-free UDA object detection task as an unsupervised false negatives mining problem and propose a simple yet effective S&M self-training framework. **3)** Extensive experiments show the superiority over the state-of-the-art under various scenarios by a considerable margin, indicating the effectiveness of the proposed framework.

2. DOMAIN GENERALIZATION ANALYSIS

Domain generalization has a great impact on pseudo labeling for cross-domain self-training. However, few works offer insights into domain generalization in object detection. This paper carries out substantial quantitative evaluations on multiple domain adaptation benchmarks to analyze the domain generalization in object detection. A phenomenon is found that object detection ability is differently transferable between various visual appearances, termed as domain generalization differentiation. Especially, the ability to detect hard samples is dramatically degenerated while the ability to detect easy samples is well transferred. Here we take Sim10k to Cityscapes, a popular domain transfer benchmark, as an example for detailed illustration. As shown in Fig.3, the bounding boxes with a very small scale are hard to detect even when we lower the confidence threshold to zero. These missing bounding boxes will act as false negatives during pseudo labeling, and the proportion of false negatives is more than 50%, harming the performance of self-training. Fortunately, there exist considerable true positives with very high confidence ($\geq$0.9), which can provide supervision cues for unsupervised pseudo label optimization. Although there are a small amount of false positives mixed with true positives in the region of high confidence, their influence is limited compared with the massive false negatives. Nevertheless, false positives suppression is still somewhat important and we leave this as our future work. In this paper, we mainly delve into unsupervised false negatives mining problem.

Methods	Source	AP of car (K→C)	AP of car (S→C)
Source only	✓	41.9	39.9
SW-Faster [6]	✓	37.9	42.3
Noise Labeling [23]	✓	43.0	42.6
DA-Detection [18]	✓	43.9	-
AT-Faster [7]	✓	42.1	42.8
Coarse-to-Fine [9]	✓	43.8	-
SFOD [22]	×	44.6	42.9
S&M w/o FNS+FNM	×	41.9	43.0
S&M w/o FNM	×	48.1	46.7
S&M	×	**49.7**	**48.2**

Table 1. Results of two domain adaptation tasks, including K-to-C and S-to-C.

Methods	Source	tru	car	rid	per	tra	mot	bic	bus	mAP
Source only	✓	21.1	43.7	39.1	31.3	25.3	22.3	36.3	37.9	32.1
SW-Faster [6]	✓	23.7	47.3	42.2	32.3	27.8	28.3	35.4	41.3	34.8
Noise Labeling [23]	✓	35.1	42.1	49.2	30.1	45.2	27.0	26.9	36.0	36.4
DA-Detection [18]	✓	24.3	54.4	45.5	36.0	25.8	29.1	35.9	44.1	36.9
CR-DA-DET [8]	✓	27.2	49.2	43.8	32.9	36.4	30.3	34.6	45.1	37.4
Coarse-to-Fine [9]	✓	43.2	37.4	52.1	34.7	34.0	46.9	29.9	30.8	38.6
AT-Faster [7]	✓	23.7	50.0	47.0	34.6	38.7	33.4	38.8	43.3	38.7
SFOD [22]	×	27.9	51.7	44.7	33.2	21.3	28.6	37.3	45.9	36.3
S&M w/o FNS+FNM*	×	28.7	48.7	38.5	28.6	24.8	23.2	33.9	44.7	34.4
S&M w/o FNM*	×	29.5	52.7	45.5	34.2	**36.9**	29.8	38.7	46.3	37.9
S&M*	×	**29.6**	**52.9**	**49.0**	**38.6**	26.2	**34.8**	**41.7**	**50.0**	**39.7**

Table 2. Results of adaptation from C-to-F. $*$ denotes pseudo labeling in the defoggy data while training and testing in the foggy data.

Methods	Source	tru	car	rid	per	tra	mot	bic	bus	mAP
Source only	✓	22.6	50.6	32.5	34.8	-	24.2	25.6	21.9	30.1
SW-Faster [6]	✓	15.2	45.7	29.5	30.2	-	17.1	21.2	18.4	25.3
CR-DA-DET [8]	✓	19.5	46.3	31.3	31.4	-	17.3	23.8	18.9	26.9
SFOD [22]	×	20.6	50.4	32.6	32.4	-	18.9	25.0	23.4	29.0
S&M w/o FNS+FNM	×	25.9	50.6	33.4	34.3	-	23.0	30.9	26.3	31.8
S&M w/o FNM	×	29.7	**52.4**	**37.6**	**42.1**	-	25.0	32.6	26.7	34.4
S&M	×	**30.0**	52.0	36.7	41.7	-	**25.3**	**33.2**	**27.5**	**34.6**

Table 3. Results of adaptation from C-to-B.

3. SIMULATION-AND-MINING FRAMEWORK

3.1. False Negatives Simulation

The pseudo boxes with very high confidence are assumed to be reliable true positives, which is a reasonable premise to start our S&M self-training framework. Given the pseudo boxes with confidence scores annotated by the pre-trained model, they can be filtered by a very high confidence threshold δ to obtain $\hat{D}_t = \{x_t^i, \hat{y}_t^i\}_{i=1}^{N_t}$, where x_t^i is the i-th target data and $\hat{y}_t^i$ is the corresponding pseudo label. Without any specific statement, δ is set as 0.9 by default.

As we studied above, prior knowledge about the relation between true positives and false negatives is the same semantic information but the difference in appearance. Briefly speaking, false negatives can be regarded as hard sample and essentially, a low visual quality transformation of true positives. Given this tip, we attempt to augment true positives to simulate false negatives. The straightforward truth is that you know what you have seen. If the model is trained with some true positives with similar appearance with false negatives during training, the related false negatives can be continuously dug out from the background class. Since false negatives usually act as the small-scale and occluded transformation of true positives, following SFOD [22], we perform false negatives simulation via Mosaic augmentation [2], which merges four different images into one image with random resize and occlusion. To enhance false negatives simulation, we perform stronger Mosaic augmentation with a more aggressive zoom-out range.

3.2. False Negatives Mining

After optimizing the model via false negatives simulation, it is necessary to mine back the reliable false negatives from background to foreground via re-filtering. However, in the context of UDA, there is no annotated validation set to select the best model trained. To tackle this problem, Temporal Model Committee (TMC) is set up to combine experiences of multiple models from different timestamps. We consider that every piece of history has a positive impact on the present and regard the timestamp T as a hyper-parameter. In this paper, we twist T=5 for all adaptation tasks simultaneously. f_n denotes the output of the model at n-th timestamp with the input of augmented image $\tilde{x}_t^i$. Since the TMC is expected to preserve the original performance as lower bound, the final pseudo boxes $Y(x_t^i)$ can be formulated as,

$$Y(x_t^i) = F(\bigcup_{n=1}^{T} f_n(\tilde{x}_t^i) \cup S_{source}) \quad (1)$$

where S_{source} is the pseudo boxes provided by the source domain pre-trained model. $F(\cdot)$ is a filtering function to filter out objects with low confidence by the threshold $\delta = 0.9$. To obtain the final pseudo boxes, *Non-Maximum Suppression* (NMS) is adopted to $Y(x_t^i)$, which are exploited to drive the next round training for false negatives simulation as is illustrated in Fig.2. To avoid over-fitting to noisy labels, in each round of false negative simulation, only the updated pseudo labels are inherited while the optimized model by false negatives mining is discarded.

4. EXPERIMENTS

4.1. Experimental Setup

Datasets: Five public datasets are utilized in our experiments. **KITTI (K)** [24] is a popular dataset for autonomous driving which are manually collected with 7,481 labeled images. **Sim10k (S)** [25] is from a computer game Grand Theft Auto (GTAV) with 10k images. **Cityscapes (C)** [26] focuses on the high variability of outdoor street scenes from different cities with 2,975 training images and 500 validation images. **Foggy Cityscapes (F)** [27] is built upon the Cityscapes [26], which simulates three levels of foggy weather. **BDD100k (B)** [28] consists of 100k images. Only a subset of images labeled as daytime are utilized here, which is consistent with other comparison methods [4, 6, 8, 22].

Implementation Details: For a fair comparison, we follow exactly the same experimental setting as [4, 8, 22]. Re-

Methods	GL	TT	tru	car	rid	per	tra	mot	bic	bus	mAP
S&M w/o FNS+FNM	×	×	21.1	45.7	40.1	32.3	26.3	23.3	37.3	38.9	28.5
S&M w/o FNM	×	×	18.7	44.4	40.7	33.1	11.6	25.8	38.4	33.3	30.1
S&M	×	×	21.5	44.5	42.2	33.4	16.9	28.4	38.1	33.1	31.0
S&M w/o FNS+FNM	✓	✓	28.2	49.3	41.0	28.7	22.3	26.8	34.3	43.9	33.5
S&M w/o FNM	✓	✓	**30.3**	52.6	45.0	34.0	17.8	30.5	39.6	46.1	36.6
S&M	✓	✓	28.1	52.4	45.6	33.8	24.8	30.9	38.2	49.2	36.9
S&M w/o FNS+FNM	✓	×	28.7	48.7	38.5	28.6	24.8	23.2	33.9	44.7	34.4
S&M w/o FNM	✓	×	29.5	52.7	45.5	34.2	**36.9**	29.8	38.7	46.3	37.9
S&M	✓	×	29.6	**52.9**	**49.0**	**38.6**	26.2	**34.8**	**41.7**	**50.0**	**39.7**

Table 4. Ablation study on defoggy. GL: generate pseudo labels in defoggy data, TT: train and test in defoggy data.

Methods	TMC	AP of car
S&M w/o FNS+FNM	×	40.1
S&M w/o FNM	×	47.8
S&M	×	48.4
S&M	✓	**49.7**

Table 5. Ablation study on TMC (K-to-C).

cently, the anchor setting of Faster-RCNN for domain adaptation is changed from $\{8, 16, 32\}$ to $\{4, 8, 16, 32\}$. We follow this setting directly. To further validate our framework, three settings are included in our experiments, denoted as "S&M w/o FNS+FNM", "S&M w/o FNM" and "S&M". Specially, "S&M w/o FNS+FNM" means that we only use $\delta = 0.9$ to generate pseudo labels and train the network without any other operations. Note that the cycle time of 1 is enough to achieve saturated improvement.

4.2. Comparison with State-of-The-Arts

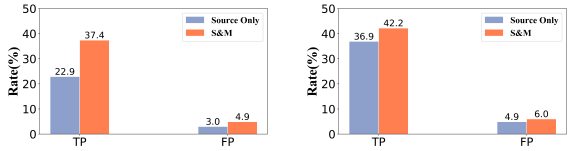
Adaptation to A New Sense (K-to-C): We investigate the ability of our method to adapt from one real dataset to another real dataset. Table 1 shows the comparison of the average precision (AP) on the car category. When FNS is used alone, AP can be increased from 41.9% to 48.1% which is even better than SFOD [22], demonstrating the effectiveness of the stronger Mosaic augmentation. The AP can be further improved to 49.7% by adopted S&M, achieving state-of-the-art performance.

Adaptation from Synthetic to Real World (S-to-C): Note that there is a performance gap between synthetic and real data. The effectiveness of the proposed method in this scenario is investigated here. Compared with recent methods [4, 6, 23, 18, 7, 9] with access to source domain, Table 1 shows the superiority of our S&M method, reaching 48.2% in terms of AP.

Adaptation from Normal to Foggy Weather (C-to-F): In this part, we study the environment adaptation from normal to foggy weather. Table 2 shows the comparison among various methods [4, 6, 23, 18, 8, 9, 7, 22]. The same defogging method as [18, 22] is performed to improve the image quality of the target data. Experimental results shows that our FNS module can boost the performance of "S&M w/o FNS+FNM" among all categories by 3.5% on average. When FNM is further introduced, our algorithm can achieve to state-of-the-art by 39.7% in terms of the mean average precision (mAP).

a) Source Only b) S&M w/o FNS+FNM c) S&M w/o FNM d) S&M

Fig. 4. Visualizations on the adaptation of Normal-to-Foggy.



a) Sim10k to Cityscapes b) Cityscapes to BDD100k

Fig. 5. The proportion of TP and FP to the ground truth.

Adaptation to Large-Scale Dataset (C-to-B): As shown in Table 3, sporadic methods challenge this task, but our framework can still achieve state-of-the-art performance. Especially when S&M are equipped with FNS and FNM sequentially, the mAP can reach 34.4% and 34.6%.

4.3. Ablation Study

Ablation Study on Defoggy Operation: Table 4 shows the ablation study in the adaptation task from Cityscapes to Foggy Cityscapes. It is the best setting to generate pseudo label in defoggy target data while training and testing in the original foggy target data, which implies that the enhanced images benefit the generation of pseudo label.

Ablation Study on TMC: Table 5 illustrates that TMC can combine the advantages of different temporal models to generate more robust pseudo labels.

Qualitative Results: The detection visualization of different experimental settings are compared in Fig.4. The "source only" mainly detects some clear objects. As FNS and FNM proceed, more and more blocked and small-scale false negatives are detected gradually. Fig.5 further visualizes the improvement that true positives get increased while false positives are almost invariant compared with initial labels from "source only".

5. CONCLUSIONS

In this paper, we first study the domain generalization differentiation phenomenon in object detection, which motives us to modify self-training framework. Then, we propose a simple yet effective S&M to solve source data-free UDA object detection problem. Experimental results on four adaptation tasks demonstrate that the proposed framework can achieve state-of-the-art performance easily. We hope our solution can bring a new inspiration to the domain adaptation community.

References

[1] S. Ren, K. He, R. Girshick, , and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks.," *NIPS*, 2015.

[2] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection.," *In arXiv:2004.10934*, 2020.

[3] S. Liu, D. Huang, and Y.Wang, "Receptive field block net for accurate and fast object detection.," *ECCV*, 2018.

[4] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. V. Gool, "Domain adaptive faster r-cnn for object detection in the wild," *In CVPR*, pp. 3339¨C–3348, 2018.

[5] W. Chen, Y. Guo, S. Yang, Z. Li, and Y. Zhuang, "Box re-ranking: Unsupervised false positive suppression for domain adaptive pedestrian detection," 2021.

[6] K. Saito, Y. Ushiku, T. Harada, and K Saenko, "Strong-weak distribution alignment for adaptive object detection," *CVPR*, 2019.

[7] Z. He and L. Zhang, "Domain adaptive object detection via asymmetric tri-way faster-rcnn," *ECCV*, 2020.

[8] C. Xu, X. Zhao, X. Jin, and X. Wei, "Exploring categorical regularization for domain adaptive object detection.," *CVPR*, 2020.

[9] Zheng Y, Huang D, Liu S, and et al., "Cross-domain object detection through coarse-to-fine feature adaptation.," *CVPR*, 2020.

[10] Zhu X, Pang J, and et al. Yang C, "Adapting object detectors via selective cross-domain alignment.," *CVPR*, 2019.

[11] Kim S, Choi J, and et al. Kim T, "Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection.," *ICCV*, 2019.

[12] Ganlong Zhao, Guanbin Li, Ruijia Xu, , and Liang Lin., "Collaborative training between region proposal localization and classification for domain adaptive object detection.," *In European Conference on Computer Vision*, pp. 86¨C–102, 2020.

[13] X. Zhu, J. Pang, C. Yang, J. Shi, and D. Lin, "Adapting object detectors via selective cross-domain alignment.," *CVPR*, 2019.

[14] Inoue N, Furuta R, Yamasaki T, and et al., "Cross-domain weakly-supervised object detection through progressive domain adaptation.," *CVPR*, 2018.

[15] T. Kim, M. Jeong, S. Kim, S. Choi, and C. Kim., "Diversify and match: A domain adaptive representation learning paradigm for object detection.," *CVPR*, 2019.

[16] Q. Cai, Y. Pan, C.-W. Ngo, X. Tian, L. Duan, and T. Yao., "Exploring object relation in mean teacher for cross-domain detection.," *CVPR*, 2019.

[17] A. Tarvainen and H. Valpola., "Mean teachers are better role models: Weight-averaged consistency targets improve semisupervised deep learning results.," *NIPS*, 2017.

[18] H.-K. Hsu, C.-H. Yao, Y.-H. Tsai, W.-C. Hung, H. Tseng, M. Singh, and M.-H. Yang, "Progressive domain adaptation for object detection.," *WACV*, 2020.

[19] A. L. Rodriguez and K. Mikolajczyk., "Domain adaptation for object detection via style consistency.," *BMCV*, 2019.

[20] S. Kim, J. Choi, T. Kim, and C. Kim., "Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection.," *ICCV*, 2019.

[21] Yang Zou, Zhiding Yu, Xiaofeng Liu, B. V. K. Vijaya Kumar, and Jinsong Wang., "Confidence regularized self-training.," *ICCV*, pp. 5981¨C–5990, 2019.

[22] Li X, Chen W, Xie D, and et al., "A free lunch for unsupervised domain adaptive object detection without source data.," *AAAI*, 2021.

[23] A. Khodabandeh, M.and Vahdat, M. Ranjbar, and W. G. Macready, "A robust learning approach to domain adaptive object detection.," *ICCV*, 2019.

[24] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite.," *CVPR*, 2012.

[25] M. Johnson-Roberson, R. Barto, C.; Mehta, S. N. Sridhar, K. Rosaen, and R. Vasudevan, "Driving in the matrix: Can virtual worlds replace humangenerated annotations for real world tasks?," *ICRA*, 2017.

[26] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding.," *CVPR*, 2016.

[27] C. Sakaridis, D. Dai, and L. V. Gool, "Semantic foggy scene understanding with synthetic data.," *IJCV*, 2018.

[28] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving video database with scalable annotation tooling.," *In arXiv:1805.04687*, 2018.