Supplementary Material for *"Target-Relevant Knowledge Preservation for Multi-Source Domain Adaptive Object Detection"*

In this supplementary material, we provide more implementation details of the detector in Sec. A, detailed experimental results for the settings of Cross Time Adaptation and Mixed Domain Adaptation in Sec. B, visualization results of the HTRM module in Sec. C as well as discussion on limitations of our approach in Sec. D.

A. More Implementation Details

In this section, we provide more implementation details about the network structure of the teacher detector TeDet$(\cdot)$. Since the student detector StDet$(\cdot)$ shares the same structure as the teacher detector, we therefore only describe the details of TeDet$(\cdot)$. Without loss of generality, we consider TeDet$(\cdot)$ with the AMSD module for two source domains. As shown in Fig. A, TeDet$(\cdot)$ consists of the VGG-16 backbone, RPN, RoI Align, RoI feature extractor, *GRL* and the multiple heads, where their configurations and the sizes of channels/feature maps are also displayed.

Images from each source domain are applied to train the corresponding head and perform adversarial learning on the other heads. Given an image from the target domain, the multiple heads make predictions simultaneously based on proposals from the shared RPN. On each proposal, the predicted classification and regression results from multi-heads are aggregated by averaging before non-maximum suppression. We implement the overall training process of the teacher-student framework based on the open source[1] of UBT [3]. In all experiments, we adopt VGG-16 [8] pre-trained on ImageNet [1] as the backbone.

B. Detailed Experimental Results

In this section, we display more experimental results for the settings of Cross Time Adaptation in Sec. B.1 and Extension to Mixed Domain Adaptation in Sec. B.2, respectively.

B.1. Cross Time Adaptation

As demonstrated in Table A, we report the AP of all categories on the BDD100K *dawn/dusk* subset. By following [11], the result on the category 'train' is not reported. The proposed TRKP approach outperforms the other counterparts for most categories. Both AMSD and HTRM improve the detection performance for almost all the categories and achieve the best result when they are combined.

B.2. Extension to Mixed Domain Adaptation

More results for the setting of Mixed Domain Adaptation are summarized in Table B, where the category "train" with very few instances is ignored as in [9]. With more available sources, the detection performance is consistently improved for most categories except for "rider", where the results drop when introducing more data. The reason behind probably lies in the huge domain gap and category shift between the source domains w.r.t. the 'rider' class. Despite of that, our method reaches the best results in most cases, showing its effectiveness.

C. Visualization of HTRM

To display the effectiveness of the HTRM module, we demonstrate the images with different target-relevance weights in the Cross Time Adaptation setting on the BDD100K dataset.

Recall that the source domains consist of images from *Daytime* and *Night*, and the target domain from *Dawn/Dust*. As shown in Fig. B, the source image with a larger weight α clearly has a more similar appearance to those from the target, in regard of the illumination condition.

D. Discussion on limitations.

The existing study [11] considers two sources (cross camera and cross time). Although we extend it to a harder case with three sources, the experimental setting of multi-source DAOD is still at street views. We will consider more source domains and larger domain gaps to further improve the generality in our future work.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1

[1] https://github.com/facebookresearch/unbiased-teacher

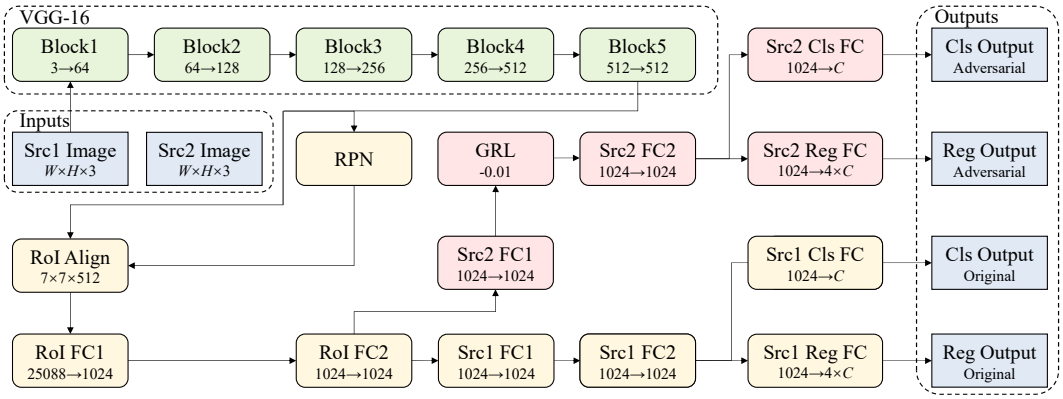

Figure A. Illustration of the detailed network architecture of the teacher detector TeDet$(\cdot)$ with the AMSD module for two source domains. The configuration and the sizes of channels/feature maps are also presented. "Block" stands for the convolutional network layers of VGG [8] and "FC" refers to the fully-connected layer. "$W \times H$" and "C" indicate the image size and the number of object categories, respectively.

Setting	Source	Method	bike	bus	car	motor	person	rider	light	sign	train	truck	mAP
Source Only	D	FRCNN [5]	35.1	51.7	52.6	9.9	31.9	17.8	21.6	36.3	-	47.1	30.4
	N		27.9	32.5	49.4	15.0	28.7	21.8	14.0	30.5	-	30.7	25.0
	D+N		31.5	46.9	52.9	8.4	29.5	21.6	21.7	34.3	-	42.2	28.9
Single Source	D	SW [6]	34.9	51.2	52.7	15.1	32.8	23.6	21.6	35.6	-	47.1	31.4
		SCL [7]	29.1	51.3	52.8	17.2	32.0	19.1	21.8	36.3	-	47.2	30.7
		GPA [10]	36.6	52.1	53.1	15.6	33.0	23.0	21.7	35.4	-	48.0	31.8
		CRDA [9]	32.8	51.4	53.0	15.4	32.5	22.3	21.2	35.4	-	47.9	31.2
		UMT [2]	39.7	52.3	56.1	14.2	35.7	23.7	31.5	42.2	-	42.4	33.8
		UBT [3] (Baseline)	37.4	52.3	56.6	14.3	35.0	22.9	31.1	40.3	-	42.6	33.2
Single Source	N	SW [6]	31.4	38.2	51.0	9.9	29.5	22.2	18.7	32.5	-	35.7	26.9
		SCL [7]	25.3	31.7	49.3	8.9	25.8	21.2	15.0	28.6	-	26.2	23.2
		GPA [10]	32.7	38.3	51.8	14.1	29.0	21.5	17.1	31.1	-	40.0	27.6
		CRDA [9]	32.3	45.1	51.6	7.2	29.2	24.9	19.9	33.0	-	41.1	28.4
		UMT [2]	37.9	18.4	50.4	8.8	24.7	11.6	15.1	30.1	-	19.4	21.6
		UBT [3] (Baseline)	42.7	18.8	52.5	8.2	26.5	20.0	19.7	29.5	-	23.7	24.2
Source Combined	D+N	SW [6]	29.7	50.0	52.9	11.0	31.4	21.1	23.3	35.1	-	44.9	29.9
		SCL [7]	33.9	47.8	52.5	14.0	31.4	23.8	22.3	35.4	-	45.1	30.9
		GPA [10]	31.7	48.8	53.9	20.8	32.0	21.6	20.5	33.7	-	43.1	30.6
		CRDA [9]	25.3	51.3	52.1	17.0	33.4	18.9	20.7	34.8	-	47.9	30.2
		UMT [2]	42.3	48.1	56.4	13.5	35.3	26.9	31.1	41.7	-	40.1	33.5
		UBT [3] (Baseline)	40.5	49.9	56.4	14.5	33.7	23.6	30.4	40.0	-	41.6	33.1
MSDA	D+N	MDAN [12]	37.1	29.9	52.8	15.8	35.1	21.6	24.7	38.8	-	20.1	27.6
		M^{3}SDA [4]	36.9	25.9	51.9	15.1	35.7	20.5	24.7	38.1	-	15.9	26.5
		DMSN [11]	36.5	54.3	55.5	20.4	36.9	**27.7**	26.4	41.6	-	50.8	35.0
		HTRM (Ours)	41.6	50.9	58.3	21.5	37.6	24.7	35.3	43.6	-	41.3	35.5
		AMSD (Ours)	44.0	55.3	60.1	17.7	39.8	26.7	37.9	46.9	-	51.2	38.0
		TRKP (Ours)	**48.4**	**56.3**	**61.4**	**22.5**	**41.5**	27.0	**41.1**	**47.9**	-	**51.9**	**39.8**
Oracle	BDD100K	FRCNN [5]	27.2	39.6	51.9	12.7	29.0	15.2	20.0	33.1	-	37.5	26.6

Table A. Detailed results for the setting of Cross Time Adaptation. 'D' and 'N' indicate the *daytime* and *night* subsets of BDD100K. mAP (%) for all the classes and detailed AP (%) of each individual category on BDD100K *dawn/dusk* are reported. Best in bold.

[2] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *CVPR*, pages 4091–4101, 2021. 2

[3] 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. 1, 2, 3

[4] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, pages 1406–1415, 2019. 2

[5] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015. 2, 3

[6] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive ob-

Setting	Source	Method	person	car	train	rider	truck	motor	bicycle	bus	mAP
Source Only	C	FRCNN [5]	26.9	44.7	-	22.1	17.4	17.1	18.8	16.7	23.4
Single Source	C	UBT [3] (Baseline)	37.8	50.9	-	**38.2**	21.3	19.9	29.9	10.9	29.7
Source Only	C+M	FRCNN [5]	35.2	49.5	-	26.1	25.8	18.9	26.1	26.5	29.7
Source Combined	C+M	UBT [3] (Baseline)	30.7	28.0	-	3.9	11.2	19.2	17.8	18.7	18.5
MSDA	C+M	**HTRM (Ours)**	34.6	48.3	-	20.2	21.7	26.7	32.0	34.1	31.1
MSDA	C+M	**AMSD (Ours)**	38.6	52.1	-	28.2	22.9	24.9	28.5	33.3	32.6
MSDA	C+M	**TRKP (Ours)**	39.2	53.2	-	32.4	28.7	25.5	31.1	37.4	35.3
Source Only	C+M+S	FRCNN [5]	36.6	49.0	-	22.8	24.9	26.9	28.4	27.7	30.9
Source Combined	C+M+S	UBT [3] (Baseline)	32.7	39.6	-	6.6	21.2	21.3	25.7	28.5	25.1
MSDA	C+M+S	**HTRM (ours)**	37.7	50.2	-	20.5	**32.7**	27.0	30.4	35.7	33.5
MSDA	C+M+S	**AMSD (ours)**	40.1	52.8	-	25.3	25.9	29.1	31.8	36.2	34.5
MSDA	C+M+S	**TRKP (ours)**	**40.2**	**53.9**	-	31.0	30.8	**30.4**	**34.0**	**39.3**	**37.1**
Oracle	BDD100K	FRCNN [5]	35.3	53.9	-	33.2	46.3	25.6	29.3	46.7	38.6

Table B. Detailed results for the setting of Mixed Domain Adaptation. 'C'/'M'/'S' indicate Cityscapes/MS COCO/Synscapes, respectively. mAP (%) and detailed AP (%) of each category on BDD100K *daytime* are reported.

Figure B. Visualization of the source images ranked by weights generated via HTRM on BDD100K. With a larger target-relevance weight α, the corresponding source image appears more similar to the images from the target domain, *i.e. dawn/dusk*.

ject detection. In *CVPR*, pages 6956–6965, 2019. 2

[7] Zhiqiang Shen, Harsh Maheshwari, Weichen Yao, and Marios Savvides. SCL: towards accurate domain adaptive object detection via gradient detach based stacked complementary losses. *arXiv preprint*, 1911.02559, 2019. 2

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1, 2

[9] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *CVPR*, pages 11721–11730, 2020. 1, 2

[10] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *CVPR*, pages 12352–12361, 2020. 2

[11] Xingxu Yao, Sicheng Zhao, Pengfei Xu, and Jufeng Yang. Multi-source domain adaptation for object detection. In *ICCV*, 2021. 1, 2

[12] Han Zhao, Shanghang Zhang, Guanhang Wu, José M. F. Moura, João Paulo Costeira, and Geoffrey J. Gordon. Adversarial multiple source domain adaptation. In *NeurIPS*, pages 8568–8579, 2018. 2