Multi-Granularity Alignment Domain Adaptation for Object Detection Supplementary Materials

A. Omni-scale Gated Fusion for Faster-RCNN

In the Faster-RCNN detector [6], we use the backbone features with the stride 16 to collect the feature maps F. Different from FCOS [10], Faster-RCNN [6] is a two-stage object detection method. Therefore we directly use the RPN as the coarse detection heads to predict the candidate boxes $\tilde{b}$. To adapt to multi-scale objects $\tilde{b}$, we construct *low-resolution*, *mid-resolution* and *high-resolution* streams in the omni-scale gated fusion module (see Figure 1). Each stream contains convolutional layers with different kernels to extract features of objects, where the 3×3 convolutional layer with stride 2 is used to expand the receptive field.

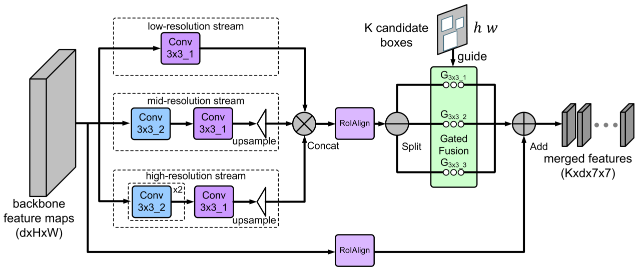

Figure 1. Omni-scale gated fusion module in the Faster-RCNN detector [6]. "3x3_2" in the blue rectangles denotes the 3×3 convolutional layer with stride 2. d is the number of channels of feature maps.

Since the RPN in Faster-RCNN [6] only predicts sparse detections of the top K proposals, we first concatenate 3 feature maps and extract the corresponding RoI features $F \in \mathbb{R}^{3 \times K \times d \times 7 \times 7}$ by the ROIAlign operation, where d denotes the dimension of the feature. Given the width w and height h of coarse detections, we can determine the corresponding gating mask from and merge them into $M \in \mathbb{R}^{K \times d \times 7 \times 7}$.

B. Real-to-Artistic Adaptation Results

As presented in Table 1 and 2, we report the detection accuracy for each category on Clipart and Watercolor [4]. It can be seen that our method outperforms other state-of-the-art algorithms. Specifically, our method achieves the best performance on 7 out of 20 categories from PASCAL VOC [2] to Clipart [4] and 3 out of 6 categories from PASCAL VOC [2] to Watercolor [4] respectively.

C. Visual Adaptive Detection Results

We provide some adaptation object detection results in Figure 2 and 3. It indicates that our method can achieve state-of-the-art performance in various complex scenarios.

References

[1] Yanhua Cheng, Rui Cai, Zhiwei Li, Xin Zhao, and Kaiqi Huang. Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation. In *CVPR*, pages 1475–1483, 2017.

[2] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010. 1

[3] Zhenwei He and Lei Zhang. Domain adaptive object detection via asymmetric tri-way faster-rcnn. In *ECCV*, volume 12369, pages 309–324, 2020. 2

[4] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*, pages 5001–5009, 2018. 1

[5] Congcong Li, Dawei Du, Libo Zhang, Longyin Wen, Tiejian Luo, Yanjun Wu, and Pengfei Zhu. Spatial attention pyramid network for unsupervised domain adaptation. In *ECCV*, volume 12358, pages 481–497, 2020. 2

[6] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *TPAMI*, 39(6):1137–1149, 2017. 1

[7] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. Gated fusion network for single image dehazing. In *CVPR*, pages 3253–3261, 2018.

[8] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, pages 6956–6965, 2019. 2

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

Method	Detector	Backbone	aero	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	
Baseline	Faster-RCNN	ResNet-101	35.6	52.5	24.3	23.0	20.0	43.9	32.8	10.7	30.6	11.7	
SW-DA [8]	Faster-RCNN	ResNet-101	26.2	48.5	32.6	33.7	38.5	54.3	37.1	18.6	34.8	58.3	
SCL [9]	Faster-RCNN	ResNet-101	**44.7**	50.0	33.6	27.4	42.2	55.6	38.3	19.2	37.9	69.0	
ATF [3]	Faster-RCNN	ResNet-101	41.9	67.0	27.4	**36.4**	41.0	48.5	42.0	13.1	39.2	**75.1**	
PD [12]	Faster-RCNN	ResNet-101	41.5	52.7	**34.5**	28.1	**43.7**	58.5	41.8	15.3	40.1	54.4	
SAPNet [5]	Faster-RCNN	ResNet-101	27.4	**70.8**	32.0	27.9	42.4	63.5	47.5	14.3	**48.2**	46.1	
Our	Faster-RCNN	ResNet-101	35.5	64.6	27.8	34.5	41.6	**66.4**	**49.8**	**26.8**	43.6	56.7	
			table	dog	horse	bike	person	plant	sheep	sofa	train	tv	mAP
Baseline	Faster-RCNN	ResNet-101	13.8	6.0	36.8	45.9	48.7	41.9	16.5	7.3	22.9	32.0	27.8
SW-DA [8]	Faster-RCNN	ResNet-101	17.0	12.5	33.8	65.5	61.6	52.0	9.3	24.9	54.1	49.1	38.1
SCL [9]	Faster-RCNN	ResNet-101	30.1	26.3	34.4	67.3	61.0	47.9	21.4	26.3	50.1	47.3	41.5
ATF [3]	Faster-RCNN	ResNet-101	**33.4**	7.9	41.2	56.2	61.4	**50.6**	**42.0**	25.0	53.1	39.1	42.1
PD [12]	Faster-RCNN	ResNet-101	26.7	**28.5**	37.7	75.4	63.7	48.7	16.5	**30.8**	54.5	48.7	42.1
SAPNet [5]	Faster-RCNN	ResNet-101	31.8	17.9	**43.8**	68.0	68.1	49.0	18.7	20.4	55.8	51.3	42.2
ours	Faster-RCNN	ResNet-101	24.3	20.9	43.2	**84.3**	**74.2**	41.1	17.4	27.6	**56.5**	**57.6**	**44.8**

Table 1. Real-to-Artistic adaptation detection results from PASCAL VOC to Clipart.

Method	Detector	Backbone	bike	bird	car	cat	dog	person	mAP
Baseline	Faster-RCNN	ResNet-101	68.8	46.8	37.2	32.7	21.3	60.7	44.6
SW-DA [8]	Faster-RCNN	ResNet-101	82.3	55.9	46.5	32.7	35.5	66.7	53.3
SCL [9]	Faster-RCNN	ResNet-101	82.2	55.1	51.8	39.6	38.4	64.0	55.2
ATF [3]	Faster-RCNN	ResNet-101	78.8	**59.9**	47.9	41.0	34.8	66.9	54.9
PD [12]	Faster-RCNN	ResNet-101	**95.8**	54.3	48.3	**42.4**	35.1	65.8	56.9
SAPNet [5]	Faster-RCNN	ResNet-101	81.1	51.1	53.6	34.3	39.8	71.3	55.2
ours	Faster-RCNN	ResNet-101	87.6	49.9	**56.9**	37.4	**44.6**	**72.5**	**58.1**

Table 2. Real-to-Artistic adaptation detection results from PASCAL VOC to Watercolor.

[10] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: fully convolutional one-stage object detection. In *ICCV*, pages 9626–9635, 2019. 1

[11] Bairui Wang, Lin Ma, Wei Zhang, Wenhao Jiang, Jingwen Wang, and Wei Liu. Controllable video captioning with POS sequence guidance based on gated fusion network. In *ICCV*, pages 2641–2650, 2019.

[12] Aming Wu, Yahong Han, Linchao Zhu, and Yi Yang. Instance-invariant domain adaptive object detection via progressive disentanglement. *TPAMI*, 2021. 2

[13] 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.

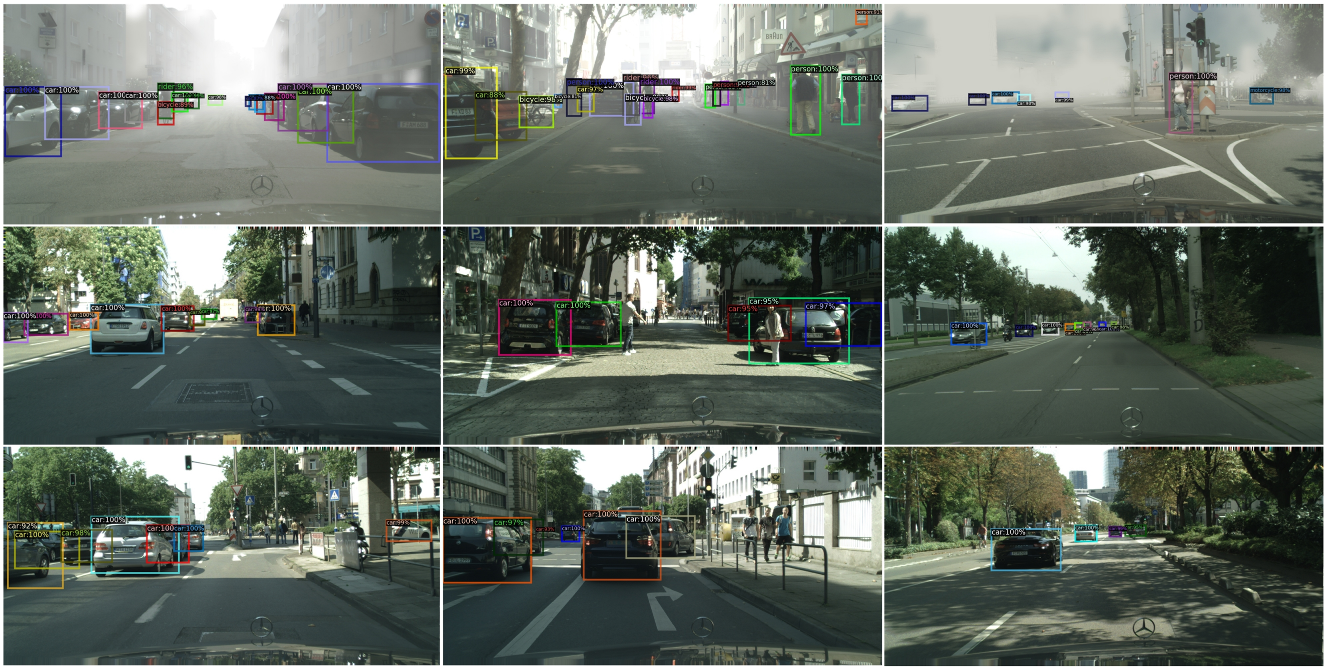

Figure 2. Weather adaptation and synthetic-to-real adaptation object detection results. From top to down: Cityscapes→FoggyCityscapes, Sim10k→Cityscapes and KITTI→Cityscapes.

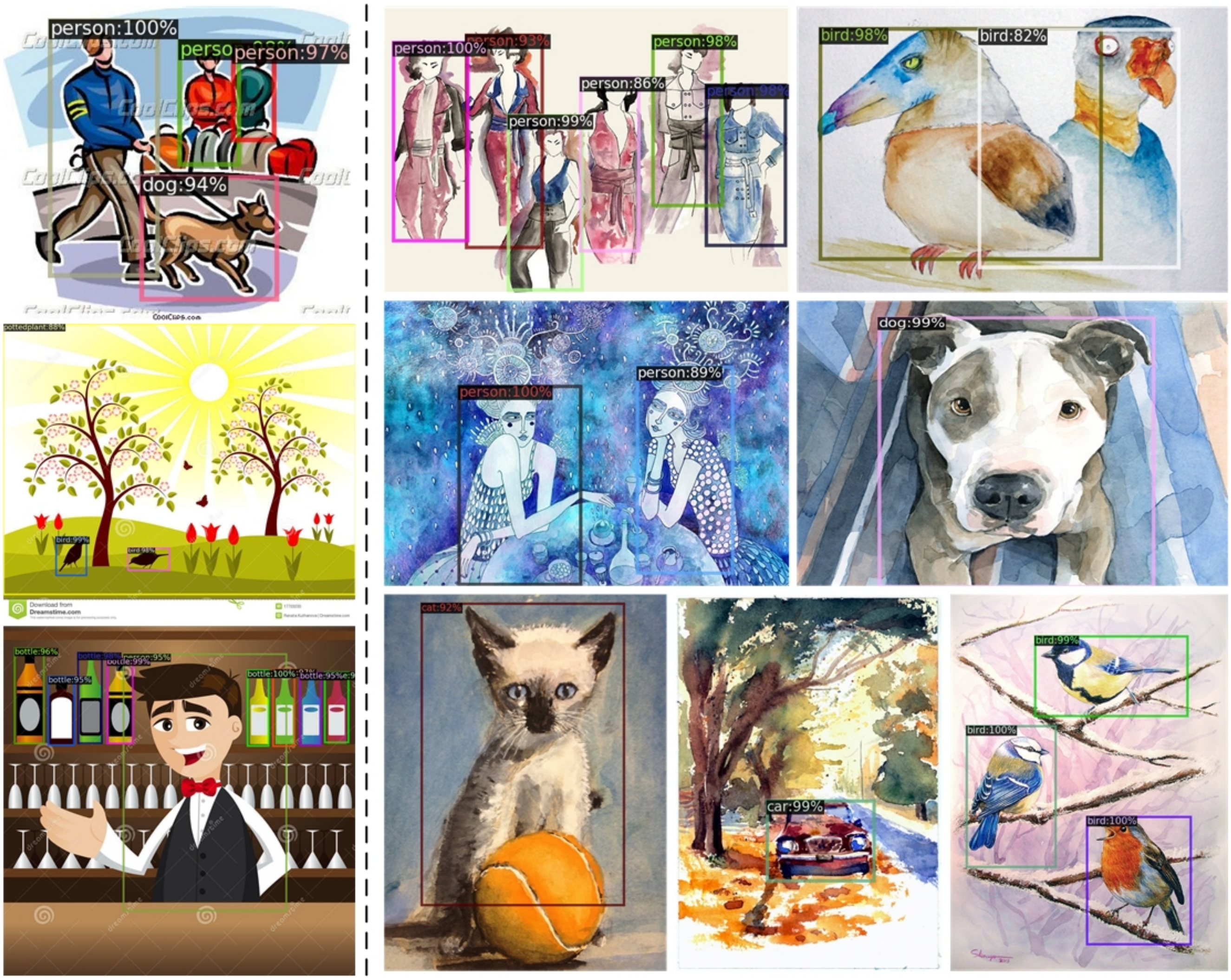

Figure 3. Real-to-Artistic adaptation object detection results. From left to right: PASCAL VOC→Clipart and PASCAL VOC→WaterColor.