SAGA: Semantic Anchor-Guided Alignment for Multi-Source Domain Adaptive Object Detection

Supplementary Material

This supplementary material offers detailed descriptions of the datasets and the three configurations used for multi-source domain adaptive object detection. It also reports the per-class AP values under the cross-time and mixed-domain settings, as outlined in Sec. A. Additional implementation details are provided in Sec. B, while further ablation studies, analyses, and visualizations are presented in Sec. C and Sec. D.

A. Datasets

A.1. Datasets

In our work, we employed five datasets to establish the multi-source domain adaptive object detection (MSDAOD) scenario. These datasets are detailed below and summarized in Tab. A.

1. BDD100k – BDD100K [15] is a large-scale and diverse driving dataset containing 70,000 training and 10,000 testing images collected under various conditions such as Daytime, Night, and Dusk/Dawn. This diversity makes it well-suited for domain adaptation tasks.

2. Cityscapes – Cityscapes [4] focuses on urban street scenes for autonomous driving research. It offers 2,975 training images and 500 testing images.

3. KITTI – The KITTI dataset [6] is widely used in self-driving research. It consists of images and corresponding sensor data captured from a moving vehicle in urban environments, providing 7,481 RGB training images.

4. MS COCO – MS COCO [8] is a well-known benchmark in computer vision, featuring a large-scale dataset with significant appearance variation. It contains around 330,000 images annotated with 80 object categories.

5. Synscapes – Synscapes [12] is a synthetic dataset tailored for autonomous driving. Its high variability makes it a valuable resource for evaluating our approach, and it includes 25,000 training images.

A.2. Class-wise AP

We also report the detailed class-wise AP of SAGA on the Cross-Time and Mixed Adaptation settings in Tab. B and Tab. C respectively. It is worth noting that the Cross-camera Adaptation setting involves only a single class; hence, this is not applicable.

B. More Implementation Details

B.1. About Faster-RCNN

To ensure a fair comparison with previous studies that adopt Faster R-CNN [10], we conduct our evaluation using the same detection framework. Following the settings in [13] and [1], we employ Faster R-CNN as the base detector within our SAGA architecture and implement it using Detectron2. Depending on the experimental configuration, VGG16 [11] pre-trained on ImageNet [5] is used as the backbone network. In alignment with the Faster R-CNN implementation incorporating ROI-Align [7], all input images are resized such that their shorter side is 600 pixels while preserving the aspect ratio. The confidence threshold is set to 0.8.

As described in Section 3.2 of the main text, training proceeds in two stages. In the second stage, SAGA is trained using source domain labels for 10k iterations. And the learned weights are transferred to initialize both the Teacher and Student models for mutual learning, which continues for an additional 50k iterations. Throughout the entire training process, we use a fixed learning rate of 0.04 without applying learning rate decay. The optimization is performed using Stochastic Gradient Descent (SGD).

For data augmentation, we apply random horizontal flipping as the weak augmentation, while strong augmentations include random color jittering, grayscaling, Gaussian blurring, and patch cutout. The exponential moving average (EMA) smoothing coefficient for the Teacher model is set to 0.9996. All experiments are conducted on four NVIDIA V100 GPUs with a total batch size of 8 and are implemented in PyTorch.

For the $\mathcal{L}_{ca}$, unlike DINO, which employs a bipartite matching strategy to obtain one-to-one positive and negative sample assignments, we retain the one-to-many positive and negative sample assignment strategy used in original Faster R-CNN. All other implementation details remain identical to those in the DINO-based detector setup.

C. More Ablation and Further Analysis

To further analyze the effect of different components and hyperparameters on the SAGA, we conduct extensive ablation studies in this section.

C.1. Number of Bidirectional Fusion Layers

In SAGA, we design bidirectional fusion layers consisting of multiple fusion layers to achieve deep interaction be-

Setting	Src.1	# Img.	Src.2	# Img.	Src.3	# Img.	Target	# Img.
Cross-Time	Day	36,728	Night	27,971	-	-	Dawn	5,027
Cross-Camera	Cityscapes	2,831	Kitty	6,684	-	-	Day	36,728
Mixed Domain	Cityscapes	2,975	COCO	71,745	Synscapes	25,000	Day	36,728

Table A. Summary of the different MSDAOD settings used in our work.

Setting	**Source**	**Detectors**	**Method**	**Bike**	**Bus**	**Car**	**Motor**	**Person**	**Rider**	**Light**	**Sign**	**Train**	**Truck**	**mAP**
Source Only	D	DINO	DINO [16]	44.8	48.5	82.6	39.8	66.7	43.2	67.2	72.5	-	61.5	54.0
	N			50.3	55.7	77.3	23.1	61.2	28.5	60.5	62.3	-	48.3	46.6
	D+N			49.9	66.9	83.9	42.9	68.6	35.2	71.2	74.4	-	65.0	55.6
Single Source	D	DINO	DA-Faster [3]	53.1	62.3	83.5	39.9	67.7	36.9	68.9	73.4	-	63.1	54.9
			UBT [9]	47.4	64.0	83.5	40.2	68.0	42.7	68.4	73.6	-	62.8	55.1
Single Source	N	DINO	DA-Faster [3]	52.1	56.1	79.2	32.3	63.4	26.4	62.2	65.1	-	50.1	48.7
			UBT [9]	52.1	57.5	78.4	26.5	62.0	29.5	62.5	63.6	-	49.4	48.2
Single Source	D+N	DINO	DA-Faster [3]	50.5	67.5	84.8	41.0	69.0	39.8	72.1	74.3	-	64.9	56.4
			UBT [9]	51.8	68.7	85.4	44.6	69.7	39.0	72.8	**75.8**	-	66.2	57.4
MSDA	D+N	Faster-RCNN	DMSN [14]	36.5	54.3	55.5	20.4	36.9	27.7	26.4	41.6	-	50.8	35.0
		Faster-RCNN	TRKP [13]	48.4	56.3	61.4	22.5	41.5	27.0	41.1	47.9	-	51.9	39.8
		Faster-RCNN	PMT [2]	55.3	59.8	67.6	29.9	47.6	32.7	46.3	56.0	-	57.7	45.3
		Faster-RCNN	AICA [1]	56.1	61.0	69.2	31.9	51.8	39.8	49.2	59.0	-	61.0	47.9
		Faster-RCNN	SAGA (Ours)	57.3	62.5	70.3	33.5	53.1	41.2	52.0	60.7	-	62.4	49.3
		DINO	PMT-DINO†	52.8	68.9	85.4	45.4	69.9	39.6	73.3	75.7	-	66.6	57.8
		DINO	**SAGA (Ours)**	**57.5**	**71.7**	**85.4**	**52.8**	**71.3**	**50.0**	**74.0**	75.3	-	**70.4**	**60.8**
Oracle	BDD100K	DINO	DINO [16]	57.7	67.8	85.2	40.5	70.7	42.3	72.8	74.8	-	66.0	57.8

Table B. Class-wise AP of SAGA compared against the Source Only, DAOD, MSDAOD, and Oracle in the cross-time setting. Source domains are daytime (D) and night (N) subsets and the target is always Dusk/Dawn of BDD100K. † represents the results we reproduced.

Setting	**Source**	**Detectors**	**Method**	**Person**	**Car**	**Rider**	**Truck**	**Motor**	**Bicycle**	**Bus**	**mAP**
Source Only	C	DINO	DINO [16]	37.4	52.7	23.3	12.9	11.9	17.9	10.7	23.8
Single Source			DA-Faster [3]	23.6	58.6	27.2	15.4	14.6	13.6	17.1	24.3
Single Source			UBT [9]	42.6	56.0	26.0	16.8	16.6	21.9	15.8	28.0
Source Only	C+M	DINO	DINO [16]	60.7	75.1	22.8	43.5	37.9	40.4	46.6	46.7
Source Combined		DINO	DA-Faster [3]	49.8	61.3	25.4	24.3	23.6	27.5	35.8	35.4
Source Combined		DINO	UBT [9]	60.2	72.8	5.0	44.0	39.1	41.4	47.3	44.3
MSDA		Faster-RCNN	TRKP [13]	39.2	53.2	32.4	28.7	25.5	31.1	37.4	35.3
MSDA		Faster-RCNN	PMT [2]	41.1	53.5	31.2	31.9	33.7	34.9	44.6	38.7
MSDA		Faster-RCNN	AICA [1]	43.3	58.1	33.3	35.1	33.7	38.6	45.2	41.0
MSDA		Faster-RCNN	SAGA (Ours)	45.1	57.6	35.5	37.0	35.8	40.5	46.7	42.6
MSDA		DINO	PMT-DINO†	61.2	75.6	23.3	44.0	38.4	40.9	47.1	47.2
MSDA		DINO	SAGA (Ours)	61.6	74.7	32.6	46.3	44.2	38.6	51.2	49.9
Source Only	C+M+S	DINO	DINO [16]	60.8	75.2	25.9	44.4	39.7	41.2	46.9	47.7
Source Combined		DINO	DA-Faster [3]	27.0	68.0	26.3	34.3	18.2	27.2	34.5	33.6
Source Combined		DINO	UBT [9]	60.3	74.8	8.3	44.9	37.4	43.8	48.4	45.4
MSDA		Faster-RCNN	TRKP [13]	40.2	53.9	31.0	30.8	30.4	34.0	39.3	37.1
MSDA		Faster-RCNN	PMT [2]	43.3	54.1	32.0	32.6	35.1	36.1	44.8	39.7
MSDA		Faster-RCNN	AICA [1]	44.9	59.2	**33.8**	33.5	38.3	39.9	46.5	42.3
MSDA		Faster-RCNN	SAGA (Ours)	46.9	59.7	32.6	38.0	41.3	43.1	47.8	44.2
MSDA		DINO	PMT-DINO†	62.4	76.8	27.5	46.0	41.3	42.8	48.5	49.3
MSDA		DINO	**SAGA (Ours)**	**63.4**	**77.0**	26.1	**50.1**	**47.5**	**45.5**	**56.3**	**52.3**
Oracle	BDD100K	DINO	DINO [16]	74.4	86.0	57.4	67.2	52.5	55.6	66.3	65.6

Table C. Class-wise AP of SAGA compared against the Source Only, DAOD, MSDAOD, and Oracle in the mixed adaptation setting. Source domains are Cityscapes(C), MS COCO(M), and Synscapes(S) datasets while the Daytime domain of BDD100K is the target domain. † represents the results we reproduced.

Table D. Effect of the number of bidirectional fusion Layers in SAGA.

# Layer	Cross Time mAP	Cross Camera AP	Mixed Domain mAP
1	59.4	66.6	51.0
2	60.2	67.2	51.7
3 (Default)	**60.8**	**67.5**	**52.3**

Table E. Ablation study on fusion direction in SAGA with three fusion layers on Mixed Domain settings.

Fusion Direction	mAP	Description
Detector → VFM (single-direction)	50.8	Inject task-specific cues into VFM
VFM → Detector (single-direction)	50.6	Transfer semantic priors to detector
Bi-directional (↔)	**52.3**	Mutual feature refinement

Table F. Performance analysis of DINOv2 version ablation study in the cross-time setting. Each number indicates the $AP_{50}(\%)$.

Source Only	DINOv2_vitb14	DINOv2_vitl14	DINOv2_vitg14
55.6	59.9	60.8	**61.5**

tween the detector features and semantic anchors. As shown in Tab. D, increasing the number of fusion layers improves detector performance, with three layers yielding the best results. Adding more layers enables deeper interaction between the semantic anchors and the encoder features, allowing the detector to acquire general visual knowledge from the vision foundation model, while the semantic anchors, in turn, learn more task-specific representations from the detector.

C.2. On Fusion Ways.

Tab. E shows that unidirectional fusion, either from detector to VFM or from VFM to detector, provides only limited gains. In contrast, the bi-directional design yields the best performance by allowing mutual feature refinement between the two branches. This confirms that cross-branch semantic interaction, rather than one-way information flow, is critical for effective domain adaptation.

C.3. Alterative Choices of Different Dinov2 Version

In SAGA, the vision foundation model can be chosen from different versions of DINOv2. As shown in Tab. F, we conducted experiments using various DINOv2 variants (vit-b14, vit-l14, and vit-l14) as the vision foundation model. With the increase in DINOv2 model size, the detector achieves its highest performance of 61.5% AP_{50} when using DINOv2_vit-g14. This improvement can be attributed to the stronger generalization capability of larger DINOv2 models, which helps the detector learn more universal and transferable representations, thereby boosting overall performance.

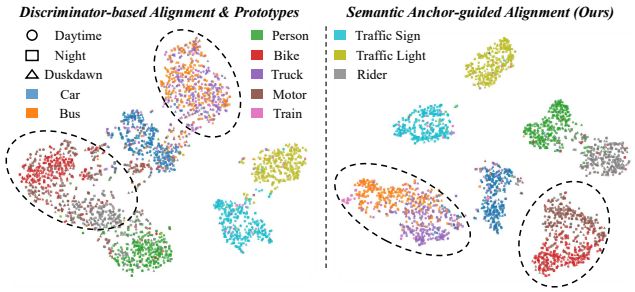

Figure A. T-SNE visualization of the features produced by the discriminator-based alignment & prototypes method and Ours. Each ellipse in the figure encloses instances from semantically similar categories across different domains. Compared with the discriminator-based alignment and prototype methods, our approach significantly alleviates semantic confusion across domains, achieving better separation between semantically similar but distinct categories, while improving the alignment of the same category across different domains.

D. Qualitative Results

D.1. T-SNE Visualization

After training, for each category, we randomly sample the same number of features from the two source and target domain for T-SNE visualization, as shown in Fig. A. It can be observed that, compared with discriminator-based alignment and prototype methods, our method better separates similar categories (truck and bus, bike and motor), effectively mitigating semantic confusion.

D.2. Detection Results

As illustrated in Fig. B, we present more qualitative comparisons among (a) *Source Only*, (b) *Multiple Discriminators (MD) + prototypes*, (c) *Ours*, and (d) *Ground-truth*. Our method can eliminate some missing errors and avoid some wrong classification cases compared with the other settings, which verifies the effectiveness of semantic anchor-guided alignment.

References

[1] Atif Belal, Akhil Meethal, Francisco Perdigon Romero, Marco Pedersoli, and Eric Granger. Attention-based class-conditioned alignment for multi-source domain adaptation of object detectors. *arXiv preprint arXiv:2403.09918*, 2024. 1, 2

[2] Atif Belal, Akhil Meethal, Francisco Perdigon Romero, Marco Pedersoli, and Eric Granger. Multi-source domain adaptation for object detection with prototype-based mean teacher. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1277–1286, 2024. 2

[3] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on*

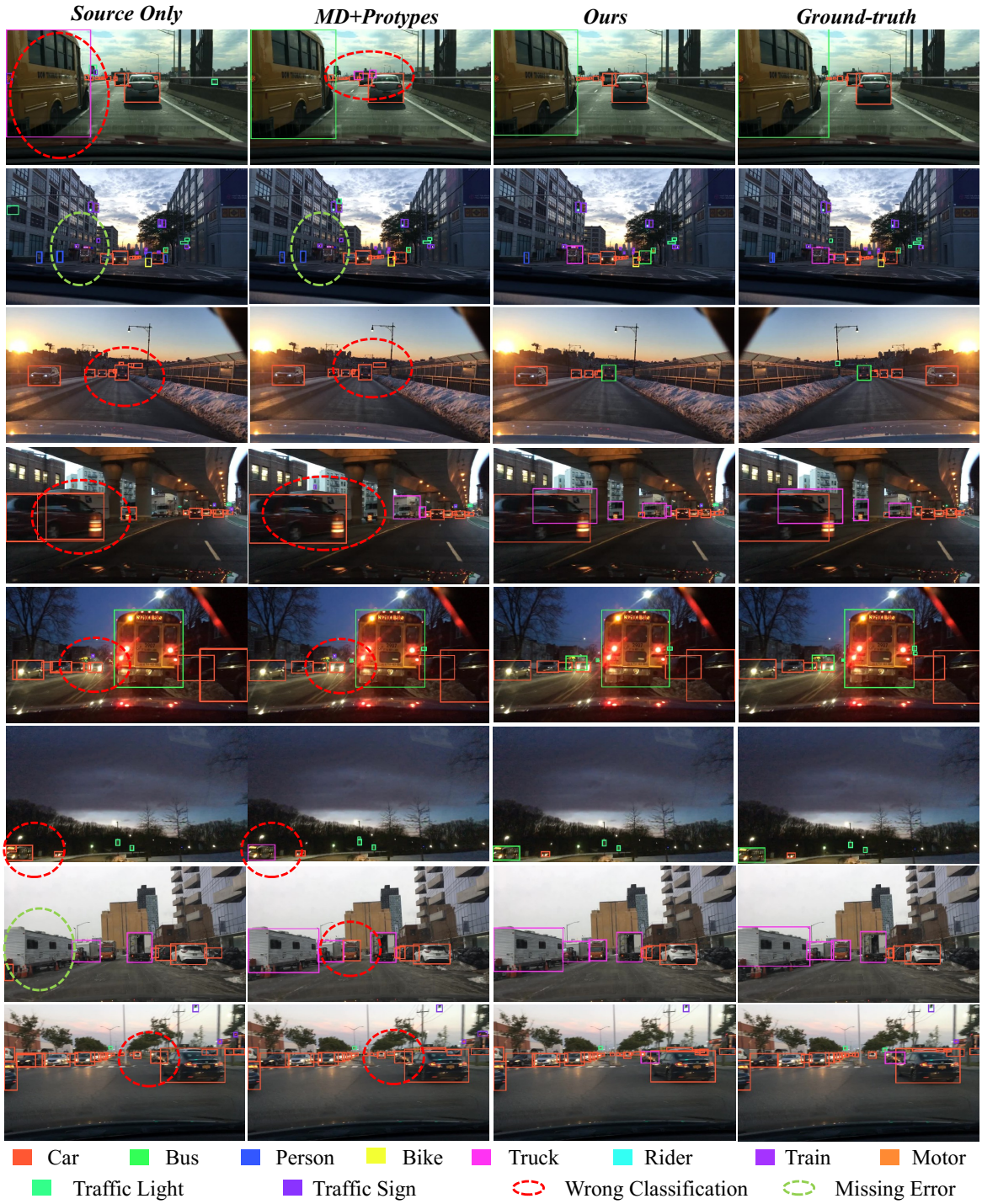

Figure B. Qualitative results on the ***Cross Time*** adaptation scenario of (a) the *Source Only* model, (b) *Multiple Discriminators (MD) + prototypes*, (c) *Ours*, and (d) *Ground-truth*. (Zooming in for best view.).

computer vision and pattern recognition, pages 3339–3348, 2018. 2

[4] 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, 2016. 1

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image

database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[6] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 1

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 1

[9] 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. 2

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016. 1

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[12] Magnus Wrenninge and Jonas Unger. Synscapes: A photorealistic synthetic dataset for street scene parsing, 2018. 1

[13] Jiaxi Wu, Jiaxin Chen, Mengzhe He, Yiru Wang, Bo Li, Bingqi Ma, Weihao Gan, Wei Wu, Yali Wang, and Di Huang. Target-relevant knowledge preservation for multi-source domain adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5301–5310, 2022. 1, 2

[14] Xingxu Yao, Sicheng Zhao, Pengfei Xu, and Jufeng Yang. Multi-source domain adaptation for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3273–3282, 2021. 2

[15] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning, 2020. 1

[16] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 2