

## A. Architectures for Domain-aware Attention

Table 1. Architectures for Domain-aware Attention module.

Features	Inter-domain Attention Branch				Intra-domain Attention Branch			
Input	$\mathbb{T}$ $K \times 1024$	$\mathbf{f}_i$ $B \times H \times W \times C$			$\mathbb{T}_s$ $K \times 1024$	$\mathbf{f}_s$ $B \times H \times W \times C$		$\mathbb{T}_t$ $K \times 1024$
Layer		$\mathbf{f}_p = \mathcal{P}(\mathbf{f}_i)$ $B \times H \times W \times 1024$	$\mathbf{f} = \mathcal{C}(\mathbf{f}_i)$ $B \times H \times W \times C$		$\mathbf{f}_p^s = \mathcal{P}^s(\mathbf{f}_s)$ $B \times H \times W \times 1024$	$\mathbf{f}_s^c = \mathcal{C}^c(\mathbf{f}_s)$ $B \times H \times W \times C$		$\mathbf{f}_p^t = \mathcal{P}^t(\mathbf{f}_t)$ $B \times H \times W \times 1024$
Layer	$\mathbf{w} = SEAttention(\mathbb{T}, \mathbf{f}_p)$ $B \times H \times W \times 1$				$\mathbf{w}_s = SEAttention(\mathbb{T}_s, \mathbf{f}_p^s)$ $B \times H \times W \times 1$			$\mathbf{w}_t = SEAttention(\mathbb{T}_t, \mathbf{f}_p^t)$ $B \times H \times W \times 1$
Output	$\mathbf{f}' = \mathbf{w} \cdot \mathbf{f} + \mathbf{f}_i$ $B \times H \times W \times C$							
Output					$\mathbf{f}'_s = \mathbf{w}_s \cdot \mathbf{f}_s^c + \mathbf{f}'$ $B \times H \times W \times C$		$\mathbf{f}'_t = \mathbf{w}_t \cdot \mathbf{f}_t^c + \mathbf{f}'$ $B \times H \times W \times C$	

We provide detailed architectures for the proposed domain-aware attention module in Table 1, including the size of input and output features for each layer.

## B. Compared to SOTA methods

We present representative state-of-the-art DAOD approaches for comparison, including feature alignment, semi-supervised learning and VLM-based domain alignment methods. This section provides a more detailed comparison including more methods as well as architectural details.

**Cross-Weather Adaptation Scenario** As shown in Table 2 (C→F), the proposed SEEN-DA outperforms all compared methods in terms of mAP and advances SOTA by 1.6%, from 55.9% to 57.5%. Specifically, our method improves performance over six categories (*i.e.* person, rider, car, truck, train, and bicycle) ranging from 0.8% to 3.1%. From the perspective of baseline, RegionCLIP [28] fine-tuned on the source domain suffers a 4.0% performance drop compared to zero-shot. This suggests that directly fine-tuning on the source domain destroys the highly generalized semantic information provided by the VLM, leading to insufficient domain-specific semantic information on the target domain. By freezing the visual encoder and tuning the domain-aware attention module, the proposed SEEN-DA shows remarkable improvements of 8.9% over the source-only variant and 4.9% over the zero-shot. These results indicate the effectiveness of the SEEN-DA in eliminating redundant information and supplementing domain-specific semantic information.

**Cross-FOV Adaptation Scenario** Table 2(K→C) reports result for KITTI→Cityscapes. SEEN-DA achieves SOTA performance of 67.1% mAP, gaining an improvement of 5.7%. And SEEN-DA outperforms the source-only baseline by 8.0%, showing great efficiency.

**Sim-to-Real Adaptation Scenario** Table 2 (S→C) shows that the proposed method achieves the best results of 66.8% mAP, outperforming the previous best entry HT [6] 65.5% with 1.3%. And the baseline is improved by SEEN-DA of 7.9% on source-only condition and 6.0% on zero-shot, validating that the domain-aware attention can efficiently improve the discriminability of the visual encoder in new scenarios.

**Cross-Style Adaptation Scenario** Additionally, we assess SEEN-DA on the more challenging Cross-Style adaptation, where the semantic hierarchy has broader discrepancies. In Table 3, SEEN-DA peaks with 47.9% mAP and improves six categories (aeroplane, bird, boat, bus, sheep and train). SEEN-DA also improves the RegionCLIP with 5.2% mAP, verifying that the method is effective under challenging domain shifts and in multi-class problem scenarios.

Table 2. Comparison (%) with existing methods on Cross-Weather adaptation Cityscapes→Foggy Cityscapes (C→F), Cross-Fov adaptation KITTI→Cityscapes (K→C) and Sim-to-Real adaptation SIM10K→Cityscapes (S→C).

Methods	Arch.	C→F								K→C		S→C
		Person	Rider	Car	Truck	Bus	Train	Motor	Bicycle	mAP	mAP	mAP
DA-Faster [3]	FR	29.2	40.4	43.4	19.7	38.3	28.5	23.7	32.7	32.0	41.9	38.2
VDD [25]	FR	33.4	44.0	51.7	33.9	52.0	34.7	34.2	36.8	40.0	-	-
DSS [23]	FR	42.9	51.2	53.6	33.6	49.2	18.9	36.2	41.8	40.9	42.7	44.5
MeGA [21]	FR	37.7	49.0	52.4	25.4	49.2	46.9	34.5	39.0	41.8	43.0	44.8
SCAN [14]	FCOS+Graph	41.7	43.9	57.3	28.7	48.6	48.7	31.0	37.3	42.1	45.8	52.6
TIA [27]	FR+CycleGAN	52.1	38.1	49.7	37.7	34.8	46.3	48.6	31.1	42.3	44.0	-
LRA [19]	FR	45.6	47.1	59.7	31.2	52.4	44.6	28.1	39.5	43.5	49.4	55.7
SIGMA [15]	FCOS+Graph	44.0	43.9	60.3	31.6	50.4	51.5	31.7	40.6	44.2	45.8	53.7
SIGMA++ [16]	FCOS+Graph	46.4	45.1	61.0	32.1	52.2	44.6	34.8	39.9	44.5	49.5	57.7
CIGAR [18]	FCOS+Graph	46.1	47.3	62.1	27.8	56.6	44.3	33.7	41.3	44.9	48.5	58.5
OADA [26]	FCOS+Teacher	47.8	46.5	62.9	32.1	48.5	50.9	34.3	39.8	45.4	47.8	59.2
MTM [24]	DETR+Teacher	51.0	53.4	67.2	37.2	54.4	41.6	38.4	47.7	48.9	-	58.1
CSDA [8]	FCOS+Graph	46.6	46.3	63.1	28.1	56.3	53.7	33.1	39.1	45.8	48.6	57.8
HT [6]	FCOS+Teacher	52.1	55.8	67.5	32.7	55.9	49.1	40.1	50.3	50.4	60.3	65.5
AT [17]	FR+Teacher	56.3	51.9	64.2	38.5	45.5	55.1	<b>54.3</b>	35.0	50.9	-	-
SOCER [4]	FR+Teacher	51.7	57.7	68.6	38.2	51.6	47.5	41.6	51.7	51.1	-	63.8
DSD-DA [7]	FR+Teacher	49.0	59.6	65.3	35.7	61.0	46.5	43.9	57.3	52.3	49.3	52.5
CAT [10]	FR+Teacher	44.6	57.1	63.7	40.8	66.0	49.7	44.9	53.0	52.5	-	-
NSA-UDA [29]	FR+Teacher+Graph	50.3	60.1	67.7	37.4	57.4	46.9	47.3	54.3	52.7	55.6	56.3
REACT [12]	FR	51.4	57.9	67.4	37.7	58.4	52.8	44.6	54.6	53.1	59.5	58.6
DA-Pro [11]	FR+VLM	55.4	62.9	70.9	40.3	<b>63.4</b>	54.0	42.3	58.0	55.9	61.4	62.9
RegionCLIP [28](Zero-Shot)	FR+VLM	51.8	59.0	67.4	36.8	59.5	50.6	39.7	55.9	52.6	59.5	60.8
RegionCLIP [28](Source-Only)	FR+VLM	49.6	55.0	63.2	34.1	55.6	48.3	36.0	47.0	48.6	59.1	58.9
SEEN-DA (Ours)	FR+VLM	<b>58.5</b>	<b>64.5</b>	<b>71.7</b>	<b>42.0</b>	61.2	<b>54.8</b>	47.1	<b>59.9</b>	<b>57.5</b>	<b>67.1</b>	<b>66.8</b>

Table 3. Comparison (%) with existing methods on Cross-Style adaptation task Pascal VOC→Clipart

Methods	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	Tv	mAP
UaDAN [9]	35.0	<b>73.7</b>	41.0	24.4	21.3	69.8	53.5	2.3	34.2	61.2	31.0	<b>29.5</b>	47.9	63.6	62.2	61.3	13.9	7.6	48.6	23.9	40.2
TFD [22]	27.9	64.8	28.4	29.5	25.7	64.2	47.7	13.5	47.5	50.9	50.8	21.3	33.9	60.2	65.6	42.5	15.1	40.5	45.5	48.6	41.2
DBGL [1]	28.5	52.3	34.3	32.8	38.6	66.4	38.2	25.3	39.9	47.4	23.9	17.9	38.9	78.3	61.2	51.7	26.2	28.9	56.8	44.5	41.6
FGRR [2]	30.8	52.1	35.1	32.4	42.2	62.8	42.6	21.4	42.8	58.6	33.5	20.8	37.2	81.4	66.2	50.3	21.5	29.3	<b>58.2</b>	47.0	43.3
UMT [5]	39.6	59.1	32.4	35.0	45.1	61.9	48.4	7.5	46.0	<b>67.6</b>	21.4	<b>29.5</b>	48.2	75.9	70.5	<b>56.7</b>	25.9	28.9	39.4	43.6	44.1
SIGMA [15]	40.1	55.4	37.4	31.1	54.9	54.3	46.6	23.0	44.7	65.6	23.0	22.0	42.8	55.6	67.2	55.2	32.9	<b>40.8</b>	45.0	58.6	44.5
ATMT [13]	37.5	63.4	37.9	29.8	45.1	62.7	41.2	19.5	43.7	57.4	22.9	25.3	39.6	87.1	70.9	50.6	29.1	32.2	58.4	50.5	45.2
CIGAR [18]	35.2	55.0	39.2	30.7	60.1	58.1	46.9	31.8	47.0	61.0	21.8	26.7	44.6	52.4	68.5	54.4	31.3	38.8	56.5	63.5	46.2
TIA [27]	42.2	66.0	36.9	37.3	43.7	<b>71.8</b>	49.7	18.2	44.9	58.9	18.2	29.1	40.7	<b>87.8</b>	67.4	49.7	27.4	27.8	57.1	50.6	46.3
SIGMA++ [16]	36.3	54.6	40.1	31.6	58.0	60.4	46.2	<b>33.6</b>	44.4	66.2	25.7	25.3	44.4	58.8	64.8	55.4	36.2	38.6	54.1	<b>59.3</b>	46.7
CMT [20]	39.8	56.3	38.7	39.7	<b>60.4</b>	35.0	<b>56.0</b>	7.1	<b>60.1</b>	60.4	<b>35.8</b>	28.1	<b>67.8</b>	84.5	<b>80.1</b>	55.5	20.3	32.8	42.3	38.2	47.0
RegionCLIP [28](Zero-Shot)	38.1	70.4	48.8	37.3	44.8	55.8	43.5	14.4	48.2	47.8	14.3	18.3	58.3	78.4	67.9	22.2	30.1	16.9	48.4	50.2	42.7
SEEN-DA (Ours)	<b>44.1</b>	73.4	<b>54.7</b>	<b>47.1</b>	45.1	<b>76.0</b>	51.6	20.4	51.7	53.0	18.5	17.3	61.8	86.8	72.2	22.8	<b>37.7</b>	21.1	<b>58.9</b>	52.7	<b>47.9</b>

### C. Sensitivity on $\mathcal{L}_{adv}$

Table 4. Sensitivity to hyper-parameters of initialization of  $\lambda_{adv}$ .

Cityscapes→FoggyCityscapes						
$\lambda_{adv}$	0.01	0.05	0.1	0.5	1.0	10.0
mAP	56.5	57.1	57.5	55.3	53.6	52.8

To select hyper-parameters for the adversarial loss in inter-domain attention branch, we perform experiments of different choices of the weight value  $\lambda_{adv}$ . We conduct the experiment on SEEN-DA on Cityscapes→FoggyCityscapes adaptation



Figure 1. Visual Features w/ and w/o domain-aware attention

scenarios, as shown in Table 4. Initialized with 0.01, it suffers from insufficient alignment and obtains limited performance of 56.5% mAP. Increasing the  $\lambda_{adv}$ , SEEN-DA peaks 57.5% when  $\lambda_{adv} = 0.1$ , and further increasing the hyper-parameter leads to significant performance degradation. Therefore, we set  $\lambda_{adv}$  to 0.1.

#### D. Sensitivity on $\lambda_t$

Table 5. Sensitivity to hyper-parameters of initialization of  $\lambda_t$ .

Cityscapes→FoggyCityscapes					
$\lambda_t$	0.1	0.5	1.0	2.0	5.0
mAP	56.9	57.3	57.5	56.3	55.8

We also study the sensitivity of weight value  $\lambda_t$  for the classification loss with pseudo labels, as shown in Table 5. As the weight value increases, the performance peaks with  $\lambda_t = 1.0$  and then appears to decline. Therefore, we set  $\lambda_{adv}$  to 1.0.

#### E. Feature Visualization

To further verify the effectiveness of domain-aware attention, we provide feature visualization. As shown in Fig. 1(c), compared with (b) baseline, our domain-aware attention highlights the features of the object region and suppresses the background area, improving the discriminability of visual features.

#### F. Analysis on Domain Tokens

The [Domain] token is manually defined to provide coarse domain information, with single word like [real] or multi words like [real-world life scenario]. Meanwhile, we use learnable prompts  $[v^s][v^t]$  to adaptively learn fine-grained multiple domain differentiating factors. In Table 6, compared with hand-crafted prompts, our learnable prompts perform well on both simple and detailed [Domain] tokens, showing effectiveness in learning complex domain factors in P→C. Therefore, the selection for [Domain] token is less important in SEEN-DA.

Table 6. Results (%) on P→C with different [Domain] tokens.

"[Domain]" token for Source	Real-world life scenarios	Real	Pascal VOC	-
"[Domain]" token for Target	Cartoon characters in art paintings	Art	Clipart	-
"A photo of [Class] in [Domain]"	47.5	47.1	46.5	46.0
" $[v^s][v^t]$ [Class][Domain]"	47.9	47.9	47.8	47.6

#### G. Error Bars

We provide error bars in Table 7. The error bars are captured by multiple running with given experimental conditions.

Table 7. mAP(%) on four benchmarks.

Cross-Weather	Cross-FoV	Sim-to-Real	Cross-Style
57.5( $\pm 0.3$ )	67.1( $\pm 0.2$ )	66.8( $\pm 0.4$ )	47.9( $\pm 0.2$ )

## H. Limitation

Though effective, the proposed SEEN-DA is specially designed for the domain adaptive object detection task, where a labelled source domain and a unlabelled target domain are needed. Currently, the method cannot deal with the setting of multiple source domains or no target domain. We plan to resolve these problems in our future research.

## References

- [1] Chaoqi Chen, Jiongcheng Li, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. Dual bipartite graph learning: A general approach for domain adaptive object detection. In *ICCV*, pages 2703–2712, 2021. 2
- [2] Chaoqi Chen, Jiongcheng Li, Hong-Yu Zhou, Xiaoguang Han, Yue Huang, Xinghao Ding, and Yizhou Yu. Relation matters: foreground-aware graph-based relational reasoning for domain adaptive object detection. *TPAMI*, 45(03):3677–3694, 2023. 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 *CVPR*, pages 3339–3348, 2018. 2
- [4] Yiming Cui, Liang Li, Jiehua Zhang, Chenggang Yan, Hongkui Wang, Shuai Wang, Heng Jin, and Li Wu. Stochastic context consistency reasoning for domain adaptive object detection. In *ACMMM*, pages 1331–1340, 2024. 2
- [5] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *CVPR*, pages 4091–4101, 2021. 2
- [6] Jinhong Deng, Dongli Xu, Wen Li, and Lixin Duan. Harmonious teacher for cross-domain object detection. In *CVPR*, pages 23829–23838, 2023. 1, 2
- [7] Yongchao Feng, Shiwei Li, Yingjie Gao, Ziyue Huang, Yanan Zhang, Qingjie Liu, and Yunhong Wang. Dsd-da: Distillation-based source debiasing for domain adaptive object detection. In *ICML*, 2024. 2
- [8] Changlong Gao, Chengxu Liu, Yujie Dun, and Xueming Qian. Csd: Learning category-scale joint feature for domain adaptive object detection. In *ICCV*, pages 11421–11430, 2023. 2
- [9] Dayan Guan, Jiaxing Huang, Aoran Xiao, Shijian Lu, and Yanpeng Cao. Uncertainty-aware unsupervised domain adaptation in object detection. *TMM*, 24:2502–2514, 2021. 2
- [10] Mikhail Kennerley, Jian-Gang Wang, Bharadwaj Veeravalli, and Robby T Tan. Cat: Exploiting inter-class dynamics for domain adaptive object detection. In *CVPR*, pages 16541–16550, 2024. 2
- [11] Haochen Li, Rui Zhang, Hantao Yao, Xinkai Song, Yifan Hao, Yongwei Zhao, Ling Li, and Yunji Chen. Learning domain-aware detection head with prompt tuning. *NeurIPS*, 36, 2024. 2
- [12] Haochen Li, Rui Zhang, Hantao Yao, Xin Zhang, Yifan Hao, Xinkai Song, and Ling Li. React: Remainder adaptive compensation for domain adaptive object detection. *TIP*, 33:3735–3748, 2024. 2
- [13] Kai Li, Curtis Wigington, Chris Tensmeyer, Vlad I. Morariu, Handong Zhao, Varun Manjunatha, Nikolaos Barmpalios, and Yun Fu. Improving cross-domain detection with self-supervised learning. In *CVPR*, pages 4745–4754, 2023. 2
- [14] Wuyang Li, Xinyu Liu, Xiwen Yao, and Yixuan Yuan. Scan: Cross domain object detection with semantic conditioned adaptation. In *AAAI*, volume 6, page 7, 2022. 2
- [15] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *CVPR*, pages 5291–5300, 2022. 2
- [16] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma++: Improved semantic-complete graph matching for domain adaptive object detection. *TPAMI*, 45(07):9022–9040, 2023. 2
- [17] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *CVPR*, pages 7581–7590, 2022. 2
- [18] Yabo Liu, Jinghua Wang, Chao Huang, Yaowei Wang, and Yong Xu. Cigar: Cross-modality graph reasoning for domain adaptive object detection. In *CVPR*, pages 23776–23786, 2023. 2
- [19] Zhengquan Piao, Linbo Tang, and Baojun Zhao. Unsupervised domain-adaptive object detection via localization regression alignment. *TNNLS*, 35(11):15170–15181, 2024. 2
- [20] VS Vibashan, Poojan Oza, and Vishal M Patel. Instance relation graph guided source-free domain adaptive object detection. In *CVPR*, pages 3520–3530, 2023. 2
- [21] Vibashan Vs, Vikram Gupta, Poojan Oza, Vishwanath A Sindagi, and Vishal M Patel. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *CVPR*, pages 4516–4526, 2021. 2

- [22] Haoan Wang, Shilong Jia, Tiejiong Zeng, Guixu Zhang, and Zhi Li. Triple feature disentanglement for one-stage adaptive object detection. In *AAAI*, pages 5401–5409, 2024. [2](#)
- [23] Yu Wang, Rui Zhang, Shuo Zhang, Miao Li, Yangyang Xia, XiShan Zhang, and ShaoLi Liu. Domain-specific suppression for adaptive object detection. In *CVPR*, pages 9603–9612, 2021. [2](#)
- [24] Weixi Weng and Chun Yuan. Mean teacher detr with masked feature alignment: A robust domain adaptive detection transformer framework. In *AAAI*, pages 5912–5920, 2024. [2](#)
- [25] Aming Wu, Rui Liu, Yahong Han, Linchao Zhu, and Yi Yang. Vector-decomposed disentanglement for domain-invariant object detection. In *ICCV*, pages 9342–9351, 2021. [2](#)
- [26] Jayeon Yoo, Inseop Chung, and Nojun Kwak. Unsupervised domain adaptation for one-stage object detector using offsets to bounding box. In *ECCV*, pages 691–708, 2022. [2](#)
- [27] Liang Zhao and Limin Wang. Task-specific inconsistency alignment for domain adaptive object detection. In *CVPR*, pages 14217–14226, 2022. [2](#)
- [28] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, pages 16793–16803, 2022. [1](#), [2](#)
- [29] Wenzhang Zhou, Heng Fan, Tiejian Luo, and Libo Zhang. Unsupervised domain adaptive detection with network stability analysis. In *ICCV*, pages 6986–6995, 2023. [2](#)