Novel Scenes & Classes: Towards Adaptive Open-set Object Detection (*Supplementary Material*)

Wuyang Li[1] Xiaoqing Guo[1,2] Yixuan Yuan[1,3,*]

[1]City University of Hong Kong [2]University of Oxford [3]The Chinese University of Hong Kong

wuyangli2-c@my.cityu.edu.hk xiaoqing.guo@eng.ox.ac.uk yxyuan@ee.cuhk.edu.hk

In this supplementary material, we provide extensive experimental justifications and discussions to clarify the proposed method, which consists of three sections.

Sec. A: Quantitative Comparison.

- The comparison with OSOD and DAOD settings;
- Further analysis in graph and motif designs;
- Sensitivity analysis of hyperparameters;
- Model efficiency comparison.

Sec. B: Clarification and Discussion.

- The difference with other related tasks;
- Clarifying more technical details;
- Detailed AOOD benchmark setup.

Sec. C: Qualitative Comparison.

A. Quantitative Comparison

A.1. Benchmark Comparison

Comparison under Open Set Object Detection (OSOD) settings. To further evaluate the novel-class detection capacity, we compare the performance under the OSOD setting (without the domain gap), as shown in Table 1. We observe that PROSER [25] and OW-DETR [9] will sacrifice obvious mAP$_b$, while OpenDet [10] and the proposed SOMA can both boost the base-class accuracy. Moreover, SOMA achieves the best 51.84%, 48.87%, 61.77%, 41.32% mAP$_b$ and 4.51%, 10.3%, 11.64%, 9.96% AR$_n$ on all tasks, surpassing state-of-the-art OW-DETR [9] comprehensively. Our method works well in different scenarios.

Method	Set	mAP$_b$↑	AR$_n$↑	WI↓	AOSE↓
DDETR [26]$_{ICLR'21}$	het-sem	50.03	0.00	1.344	714
PROSER [25]$_{CVPR'21}$		50.32	2.44	0.997	296
OpenDet [10]$_{CVPR'22}$		50.85	3.47	1.031	297
OW-DETR [9]$_{CVPR'22}$		48.32	2.48	1.147	286
SOMA (ours)		**51.84**	**4.51**	**0.946**	**205**
DDETR [26]$_{ICLR'21}$	hom-sem	46.77	0.00	4.012	1693
PROSER [25]$_{CVPR'21}$		45.51	7.41	3.978	1030
OpenDet [10]$_{CVPR'22}$		47.02	8.36	3.923	996
OW-DETR [9]$_{CVPR'22}$		46.86	7.34	4.095	1000
SOMA (ours)		**48.87**	**10.37**	**3.883**	**904**
DDETR [26]$_{ICLR'21}$	freq-dec	59.46	0.00	1.631	675
PROSER [25]$_{CVPR'21}$		59.62	10.07	1.496	448
OpenDet [10]$_{CVPR'22}$		60.56	9.79	1.306	392
OW-DETR [9]$_{CVPR'22}$		58.75	7.69	**1.231**	**342**
SOMA (ours)		**61.77**	**11.64**	1.313	371
DDETR [26]$_{ICLR'21}$	freq-inc	39.47	0.00	5.299	3049
PROSER [25]$_{CVPR'21}$		39.44	3.46	5.015	1962
OpenDet [10]$_{CVPR'22}$		40.74	3.20	4.428	1523
OW-DETR [9]$_{CVPR'22}$		38.92	3.26	4.532	938
SOMA (ours)		**41.32**	**9.96**	**4.231**	**647**

Table 1. Comparison results on Cityscapes with OSOD evaluation on five novel classes. This setting does not consider novel scenes.

Method	SFA [20]	AQT [12]	O^{2}net [7]	MTTrans [23]	SOMA
Reference	MM21	IJCAI22	MM22	ECCV22	Ours
C→F	41.3	47.1	46.8	45.4	**47.9 (+0.8)**
C→B	28.9	29.4	30.5	32.6	**33.0 (+0.4)**

Table 2. Comparison results on Cityscapes→Foggy Cityscapes (C→F) and Cityscapes→BDD100k (C→B) with DAOD settings. This setting does not consider novel classes.

Comparison under Domain Adaptive Object Detection (DAOD) settings. To evaluate and justify the effectiveness of the cross-domain adaptation, we further conduct DAOD experiments on Cityscapes→Foggy Cityscapes (C→F) and Cityscapes→BDD100k [22] (C→B) following the standard DAOD setting [20] and implement SOMA on the latest DAOD baselines to make a comparison with existing DETR-based DAOD methods. SOMA outperforms the existing best entry (AQT [13]) with 0.8% on C→F setting, and also surpasses SOTA method (MTTrans [23]) on C→B with 0.4% mAP respectively. This indicates that introducing adequate high-order evidence can greatly enhance the domain alignment with more accurate pseudo-labels, verifying the

[*]Corresponding author.

This work was supported by Hong Kong Research Grants Council General Research Fund 11211221, 14204321, 14220622, and Innovation and Technology Commission-Innovation and Technology Fund ITS/100/20.

practical design of the proposed method.

A.2. Further Analysis

Justifying densely connected base-class centers. We explore establishing graphs with densely connected base-class centers (*dense*), as shown in Table 3. Our default design (*farthest*) outperforms the dense counterpart on both het-sem (*left*) and hom-sem (*right*) settings significantly, verifying our reasonable design. The reason may be that the farthest connection can avoid the semantic-level bias for two similar classes, improving the class-irrelevant property learning for reliable novel-class learning. Moreover, the T-SNE visualization in the main paper also supports the effectiveness of the farthest design in the graph establishment.

Extension to higher-order motifs. We further explore 4^{th} and 5^{th} order motifs, as shown in the Table 4. Compared with 3^{rd} order, higher-order reduces open-set errors (AOSE) but sacrifices AR$_n$ due to the more strict motif selection constraint in the topological evidence aspect. Hence, increasing the order of motifs can select higher-quality motifs with better semantic discriminability (better AOSE and mAP$_b$), but may reduce the number of selected motifs with a worse AR$_n$. Our design (3^{rd}-order) achieves a satisfactory trade-off between the base and novel classes.

A.3. Sensitivity Analysis

The weight of loss functions. The detailed experimental analysis on the λ_1 for $\mathcal{L}_{STL}$ and λ_2 for $\mathcal{L}_{SNL}$ is shown in Table 5. We control one of the hyper-parameters and change the other to analyze its sensitivity respectively. As for λ_1 with $\{0.05, 0.1, 0.2\}$ three settings, we find that base-class results are robust while setting a larger weight can further improve novel-class performance, *e.g.*, $\lambda_1 = 0.02$ gives a better 4.15% AR$_n$. Our default setting ($\lambda_1 = 0.1$) can achieve a satisfactory trade-off between the base and novel classes on all four evaluation metrics. As for λ_2, we find that setting a larger and smaller value both affects the performance negatively, which verifies our optimal setting.

The learning of the semantic bank. As shown in Table 6, we analyse the semantic bank **Q** in terms of α (controlling the learning speed) and β (scaling the standard diversion). As for the α with $\{0.001, 0.01, 0.1\}$, the setting with a too large ($\alpha = 0.1$) or small ($\alpha = 0.001$) value has some negative influence with 44.29% and 43.21% mAP$_b$ respectively, compared with our optimal setting ($\alpha = 0.01$) with 45.5% mAP$_b$. Considering the three settings $\{1.0, 2.0, 3.0\}$, setting the scaling factor β with 1.0 and 2.0 give similar results with robust performance. Using a too large value $\beta = 3.0$ reduces the performance on mAP$_b$ and AR$_n$, which may be caused by the inaccurate distribution estimation. Hence, the sensitivity analysis verifies the robustness of our method and the appropriate hyperparameter tuning.

	het-sem				hom-sem			
Graph	mAP$_b$↑	AR$_n$↑	WI↓	AOSE↓	mAP$_b$↑	AR$_n$↑	WI↓	AOSE↓
dense	43.76	3.55	0.777	1218	43.07	8.07	2.883	3021
farthest	**45.55**	**4.08**	**0.526**	**649**	**43.37**	**8.42**	**2.281**	**2886**

Table 3. Comparison of the densely connected graph (*dense*) and the proposed farthest connected design (*farthest*) on Cityscapes→ Foggy Cityscapes with 5 novel classes.

3^{rd} order (default)			4^{th} order			5^{th} order		
mAP$_b$↑	AR$_n$↑	AOSE↓	mAP$_b$↑	AR$_n$↑	AOSE↓	mAP$_b$↑	AR$_n$↑	AOSE↓
45.55	**4.08**	649	**46.20**	3.82	452	45.92	3.77	**429**

Table 4. Comparison results on Cityscapes→ Foggy Cityscapes (het-sem. 5 novel-class) with the motifs in different orders.

λ_1	λ_2	mAP$_b$ ↑	AR$_n$ ↑	WI↓	AOSE↓
.05	.01	45.37	2.98	0.532	717
.1	.01	**45.55**	4.08	**0.526**	**649**
.2	.01	44.96	**4.15**	0.589	710
.1	.005	44.53	3.67	0.566	683
.1	.01	**45.55**	**4.08**	**0.526**	**649**
.1	.02	44.68	3.52	0.741	832

Table 5. Comparison results on Cityscapes→Foggy Cityscapes about different loss function weight settings on λ_1 and λ_2.

	α			β		
	0.001	0.01	0.1	1.0	2.0	3.0
mAP$_b$ ↑	44.29	**45.55**	43.21	**45.55**	45.39	44.02
AR$_n$ ↑	4.02	**4.08**	3.81	4.08	**4.09**	3.21

Table 6. Comparison results on Cityscapes→Foggy Cityscapes about the hyper-parameter α and β in semantic bank learning.

Training time (s/iter)↓			Inference time (s/iter)↓		
Baseline	OW-DETR	SOMA	Baseline	OW-DETR	SOMA
0.5191	0.8169	0.6725	**0.3950**	0.4029	**0.3950**

Table 7. Comparison of training time and inference speed.

A.4. Model Efficiency

We compare the training and inference time (s/iter) with DDETR [26] (Baseline), OW-DETR [9] (SOTA counterpart), and the proposed SOMA. Compared with the baseline model, SOMA achieves comparable training time and does not sacrifice inference speed. Moreover, SOMA works better than OW-DETR in both aspects with satisfactory model efficiency. Note that the proposed SOMA is a parameter-free method, which doesn't introduce any extra computation cost in the inference stage. Hence, *the considerable performance improvement of the proposed SOMA framework does not rely on additional model parameters*.

B. Clarification and Discussion

B.1. The Difference with Existing Task Settings

The relationship between AOOD and Universal Domain Adaptive Object Detection (UniDAOD). Though UniDAOD [18] also considers the novel classes in the cross-domain scenes, there are significant differences between the proposed AOOD and UniDAOD, which can be summarized into the three aspects: *1) Task setting.* AOOD aims to conduct high-quality detection in both base and novel classes while UniDAOD only detects base-class objects and does not detect novel objects. The critical advantage of AOOD is its capacity to detect abnormal objects. *2) Methodology Design.* AOOD needs to mitigate the mutual influence between base and novel objects to ensure good performance in both aspects. UniDAOD focuses on better base-class learning by eliminating the influence of novel objects. *3) Evaluation Metrics.* AOOD follows the strict evaluation in both base-class detection (mAP$_b$) and novel-class detection (AR$_n$, WI, and AOSE), while UniDAOD only needs to evaluate the base-class performance with mAP$_b$.

The difference between AOOD and Open-Set Object Detection (OSOD). OSOD [3, 10], a.k.a, OOD in object detection [5, 4] aim to detect base and novel-class objects in a labeled domain, while AOOD needs to perform the same detection in an unlabeled domain with significant domain shift. In OSOD, the objects out of the base-class distribution can be uniformly considered as novel-class objects for the model training. Differently, AOOD is a more challenging setting since both novel-class and cross-domain objects are embedded out of the labeled base-class distribution, preventing reliable model learning.

The difference between AOOD and Open Vocabulary Detection (OVD). OVD [8, 24, 6] requires additional labels (image captions) with linguistic cues for both base and novel classes, and classifies each novel class within this labeled domain. In contrast, AOOD does not rely on any labeled novel-class cues, and must detect novel-class objects as unknowns in an unlabeled novel domain.

The difference between AOOD and related classification settings. Assuming the base-class set Ω_b, novel-class set Ω_n and background-class set Ω_{bg}, we define the class space of a source and target domain with C^s and C^t, respectively, to clarify the difference. Adaptive open-set classification has not been well benchmarked in literature, while some highly related settings are summarized below.

- Open-set domain adaptation (OSDA) [16, 17, 15], assumes $C^s = \Omega_b$ and $C^t = \Omega_b \cup \Omega_n$.
- Universal domain adaptation (UniDA) [21] follows the assumption with $C^s = \Omega_b \cup \Omega_n^s$ and $C^t = \Omega_b \cup \Omega_n^t$ with the inconsistent novel-class splitting $\Omega_n^s \neq \Omega_n^t$.
- Partial domain adaptation (PDA) [1] is defined as $C^s = \Omega_b$, $C^t \subsetneq \Omega_b$.

Set	Base-class	Novel-class $\{3, 4, 5\}$
het-sem	car truck bus	person motor train person motor train bicycle person motor train bicycle rider
hom-sem	person bicycle bus	car truck train car truck train motor car truck train motor rider
freq-dec	person car rider	bicycle train truck bicycle train truck motor bicycle train truck motor bus
freq-inc	motor truck bus	person train car person train car bicycle person train car bicycle rider

Table 8. Detailed class splitting settings for Cityscapes→ Foggy Cityscapes and Cityscapes→ BDD100k benchmarks.

The differences between the proposed AOOD and the aforementioned tasks lie in three aspects. 1) Different from all classification tasks, AOOD considers a more challenging and real-world friendly setting: $C^s = \Omega_b \cup \Omega_n^s \cup \Omega_{bg}^s$ and $C^t = \Omega_b \cup \Omega_n^t \cup \Omega_{bg}^t$, with the involved unlabeled background class. 2) Different from UniDA, there are no labeled novel classes in the source domain for AOOD. Moreover, AOOD is a more practical setting without the $\Omega_n^s \neq \Omega_n^t$ constraint since it formulates the existence of the same/different novel objects in two domains. 3) Unlike OSDA and PDA, AOOD allows novel classes to appear everywhere (both domains) to simulate the real-world scenario.

B.2. Technical Details

The basic object detector (DDETR). Given batched image input I, DDETR [26] first adopts a feature extractor to obtain multi-level image features X. With the image-level feature $X \in \mathbb{R}^{D\times H\times W}$ extracted from the backbone, DDETR [26, 2] formulates each feature point $\{x_i\}_{i=1}^{W\times H}, x_i \in \mathbb{R}^D$ as a feature token and send them to the deformable transformer encoder (with position encoding) to conduct the self-attention operation. Then, in the deformable transformer decoder, $N = 100$ predefined object queries $Q_{raw} \in \mathbb{R}^{N\times D}$ are introduced to conduct cross-attention with feature tokens. After that, we obtain N decoded object queries Q with the information of X.

How to obtain matched object queries. Given the decoded object queries Q, we follow DETR [2, 26] by sending object queries into feedforward networks to obtain the class and bounding box predictions. Then, the bipartite matching [2] is performed to match each ground-truth object with the optimal decoded object query. After that, the matched object query is denoted as Q_m, while the rest unmatched counterparts are denoted as Q_{um} in the main paper.

B.3. Experimental Setup

Cityscapes→Foggy Cityscapes/BDD100k. We present the detailed base and novel class splittings according to four protocols in Table 8, which considers the *semantic overlapping* and *instance frequency* for the following reasons.

1) The novel-class semantic is diverse in the real world, which may (not) be overlapped with base classes. Recently-published open-set research [19] has pointed out that both aspects (with and without semantic overlapping) are meaningful in the real world. However, most existing OSOD works [10, 11, 9, 14] only consider a single aspect, *i.e.*, the novel-class objects having minor semantic overlapping with base classes. Hence, we consider both aspects with heterogeneous and homogeneous semantics for strict evaluation in the proposed AOOD benchmark.

2) The novel-class scale is diverse in the real world, which may (not) be the majority of a scene. However, all existing OSOD research has ignored this vital open-set property in the real world. Hence, to break through this barrier, we count the number of objects in each class and select the most and the least three classes as base classes to simulate this real-world diversity.

After splitting the base and novel classes, we follow existing OSOD works [10] by considering different numbers of novel classes for strict evaluation, establishing $\{3, 4, 5\}$ three sub-tasks. We remove the image containing the novel-class objects that haven't been defined in the current sub-task[1] to avoid wrong and inaccurate evaluations.

Pascal VOC→Clipart. We follow the base-/novel-class splitting [14] by considering the first 10 classes in alphabetical order as the base class. The rest classes are divided into $|\Omega_n| \in \{6, 8, 10\}$ classes to serve as the considered novel-class set Ω_n, yielding three different sub-tasks. Note that we remove the images containing the novel-class $\notin \Omega_n$ for a fair evaluation in each sub-task, which can prevent the correctly-detected novel-class objects from being wrongly evaluated as a false-positive prediction.

C. Qualitative Comparison

We present more qualitative comparisons among (a) DDETR [26], (b) OW-DETR [9], (c) the proposed SOMA in Figure 1-2. We highlight the detected open-set objects by the proposed method. Our method can achieve better open-set detection compared with the low-order method OW-DETR [9], which verifies the effectiveness of our motif-based high-order solution. Moreover, compared with DDETR [9] and OW-DETR [26], our method gives more high-quality detection for base classes, demonstrating the strength in cross-domain adaptation. Our method performs better on both base and novel objects, fitting for the real-world scenario with great potential.

[1] The objects have corresponding labels defined in the dataset but are not considered as novel classes $\notin \Omega_n$ in the current sub-task.

References

[1] Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. Partial adversarial domain adaptation. In *ECCV*, pages 135–150, 2018. 3

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 3

[3] Akshay Dhamija, Manuel Gunther, Jonathan Ventura, and Terrance Boult. The overlooked elephant of object detection: Open set. In *WACV*, pages 1021–1030, 2020. 3

[4] Xuefeng Du, Gabriel Gozum, Yifei Ming, and Yixuan Li. Siren: Shaping representations for detecting out-of-distribution objects. In *Neurips*, 2022. 3

[5] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *arXiv preprint arXiv:2202.01197*, 2022. 3

[6] Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. Towards open vocabulary object detection without human-provided bounding boxes. *arXiv preprint arXiv:2111.09452*, 2021. 3

[7] Kaixiong Gong, Shuang Li, Shugang Li, Rui Zhang, Chi Harold Liu, and Qiang Chen. Improving transferability for domain adaptive detection transformers. In *ACMMM*, pages 1543–1551, 2022. 1

[8] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *ICLR*, 2022. 3

[9] Akshita Gupta, Sanath Narayan, KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer. In *CVPR*, 2022. 1, 2, 4, 6, 7

[10] Jiaming Han, Yuqiang Ren, Jian Ding, Xingjia Pan, Ke Yan, and Gui-Song Xia. Expanding low-density latent regions for open-set object detection. In *CVPR*, 2022. 1, 3, 4

[11] Yusuke Hosoya, Masanori Suganuma, and Takayuki Okatani. More practical scenario of open-set object detection: Open at category level and closed at super-category level. *arXiv preprint arXiv:2207.09775*, 2022. 4

[12] Wei-Jie Huang, Yu-Lin Lu, Shih-Yao Lin, Yusheng Xie, and Yen-Yu Lin. Aqt: Adversarial query transformers for domain adaptive object detection. In *IJCAI*, 2022. 1

[13] Wei-Jie Huang, Yu-Lin Lu, Shih-Yao Lin, Yusheng Xie, and Yen-Yu Lin. Aqt: Adversarial query transformers for domain adaptive object detection. In *31st International Joint Conference on Artificial Intelligence, IJCAI 2022*, pages 972–979. IJCAI, 2022. 1

[14] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *CVPR*, pages 5830–5840, 2021. 4

[15] Wuyang Li, Jie Liu, Bo Han, and Yixuan Yuan. Adjustment and alignment for unbiased open set domain adaptation. In *CVPR*, pages 24110–24119, June 2023. 3

[16] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *ICCV*, pages 754–763, 2017. 3

[17] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *ECCV*, pages 153–168, 2018. 3

[18] Wenxu Shi, Lei Zhang, Weijie Chen, and Shiliang Pu. Universal domain adaptive object detector. In *ACM MM*, pages 2258–2266, 2022. 3

[19] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. *arXiv preprint arXiv:2110.06207*, 2021. 4

[20] Wen Wang, Yang Cao, Jing Zhang, Fengxiang He, Zheng-Jun Zha, Yonggang Wen, and Dacheng Tao. Exploring sequence feature alignment for domain adaptive detection transformers. In *ACM MM*, pages 1730–1738, 2021. 1

[21] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In *CVPR*, pages 2720–2729, 2019. 3

[22] 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. In *CVPR*, June 2020. 1

[23] Jinze Yu, Jiaming Liu, Xiaobao Wei, Haoyi Zhou, Yohei Nakata, Denis Gudovskiy, Tomoyuki Okuno, Jianxin Li, Kurt Keutzer, and Shanghang Zhang. Cross-domain object detection with mean-teacher transformer. In *ECCV*, 2022. 1

[24] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021. 3

[25] Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Learning placeholders for open-set recognition. In *CVPR*, 2021. 1

[26] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *ICLR*, 2020. 1, 2, 3, 4, 6, 7

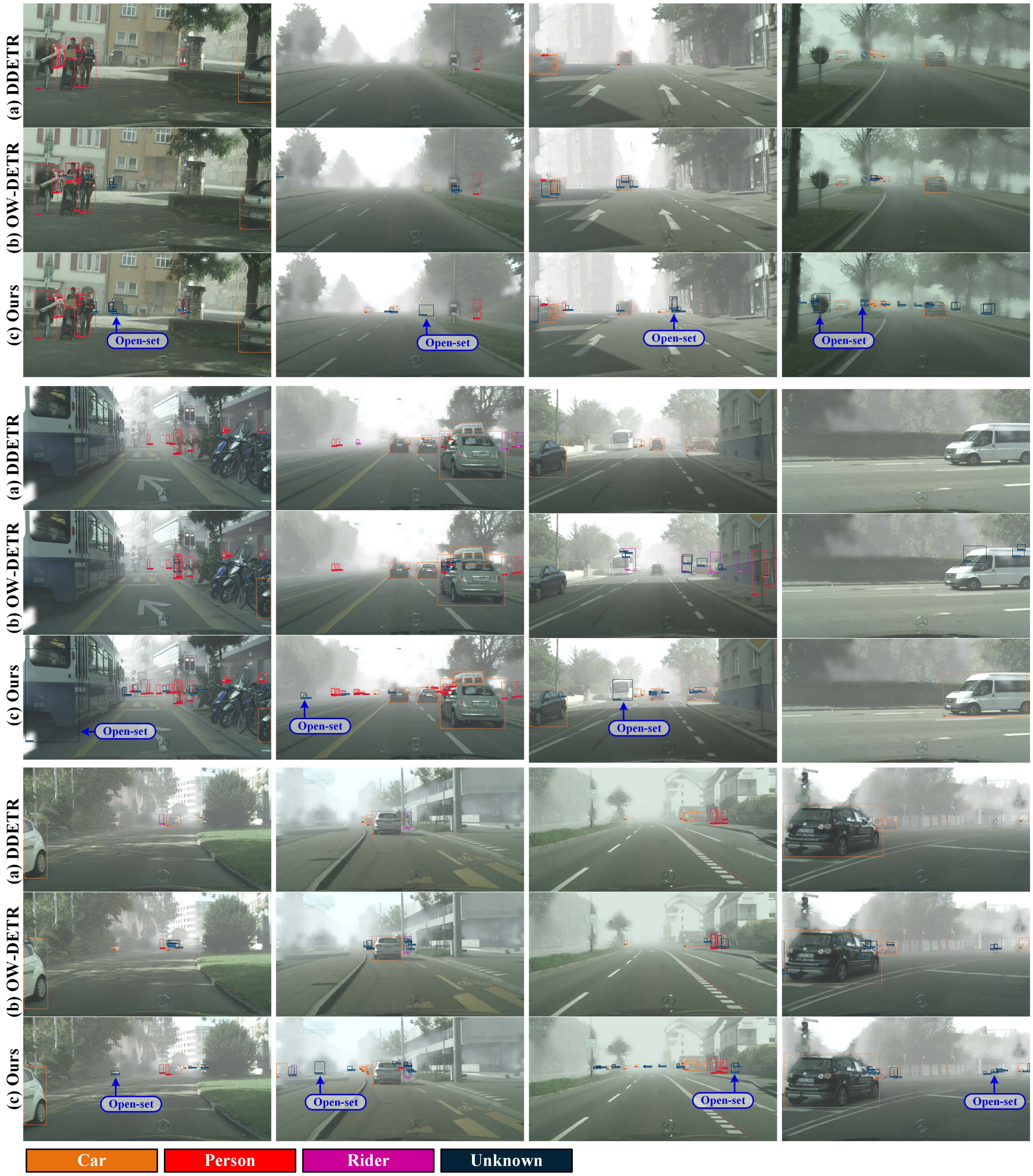

Figure 1. Qualitative results on the Cityscapes→Foggy Cityscapes AOOD benchmark of (a) DDETR baseline [26], (b) OW-DETR [9], (c) the proposed SOMA. (Zooming in for best view.)

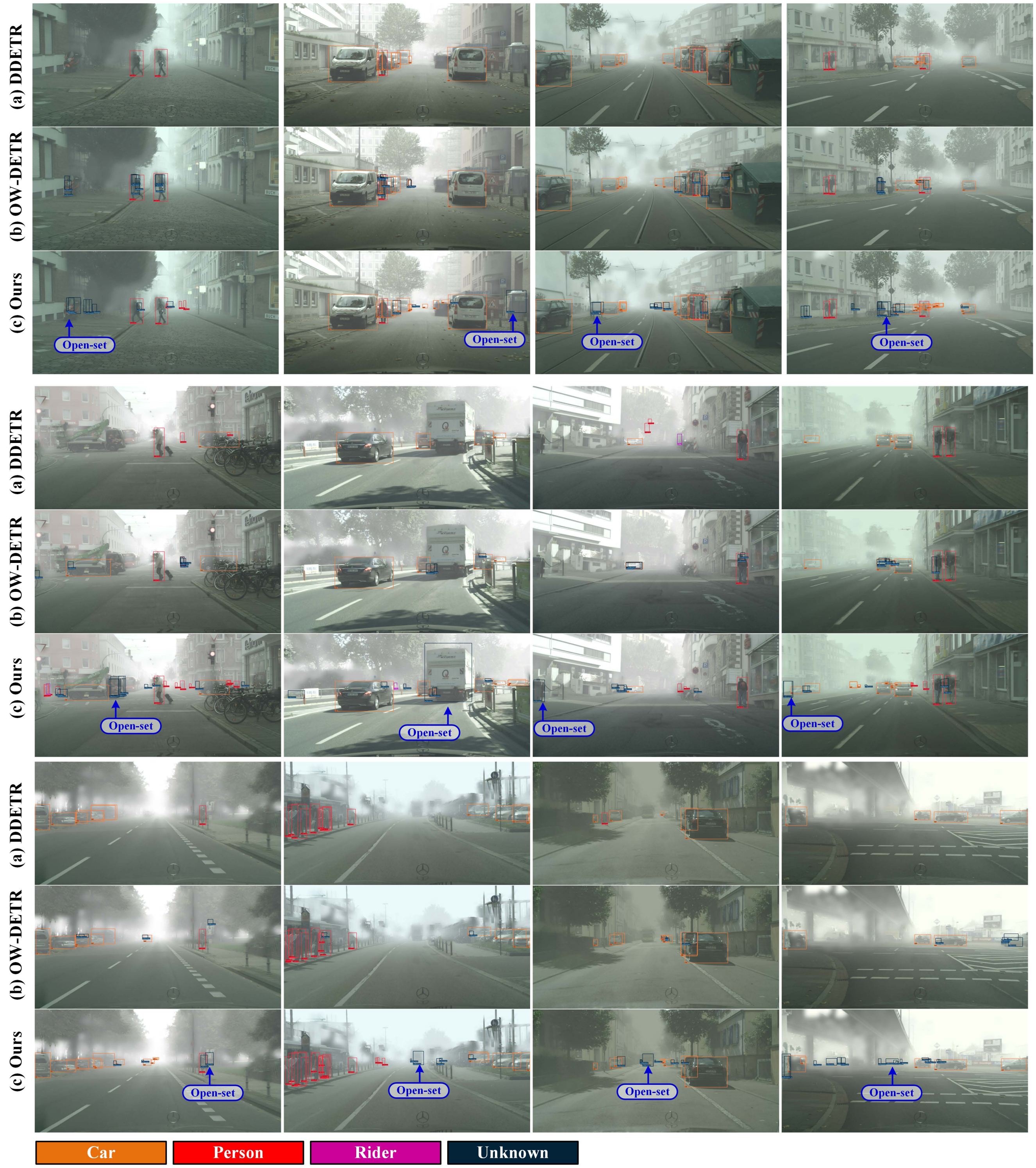

Figure 2. Qualitative results on the Cityscapes→Foggy Cityscapes AOOD benchmark of (a) DDETR baseline [26], (b) OW-DETR [9], (c) the proposed SOMA. (Zooming in for best view.)