SIGMA: Semantic-complete Graph Matching for Domain Adaptive Object Detection
(*Supplementary Material*)

Wuyang Li Xinyu Liu Yixuan Yuan*
City University of Hong Kong
{wuyangli2, xliu423}-c@my.cityu.edu.hk yxyuan.ee@cityu.edu.hk

A. Sensitivity Analysis

A.1. Parameter Sensitivity

As shown in Table 1, we analyze the sensitivity in terms of the adaptation intensity $\lambda_{1,2}$, where λ_1 works on the node classification loss and λ_2 controls the intensity of structure-aware matching loss. We first try a group of consistent parameters $\{0.05, 0.1, 0.2\}$ for $\lambda_{1,2}$ (1^{st} to 3^{rd} lines), finding that decreasing the values leads to a significant performance drop compared with our main settings ($\lambda_{1,2} = 0.1$). By fixing λ_1, increasing and decreasing λ_2 sightly decrease the overall performance, demonstrating that our setting ($\lambda_2 = 0.1$) is optimal. By fixing λ_2, decreasing λ_1 shows a significant negative impact on the framework while increasing it gives some further improvements. These results demonstrate that the larger intensity on the node loss contributes to establishing a better graphical space for the graph-matching-based adaptation.

λ_1	λ_2	$\text{mAP}_{0.5:0.95}$	$\text{mAP}_{0.5}$	$\text{mAP}_{0.75}$
0.05	0.05	22.8	42.2	21.4
0.1	0.1	24.0	43.5	**23.5**
0.2	0.2	**24.2**	43.3	23.3
0.1	0.05	23.2	42.9	23.0
0.1	0.2	23.5	43.3	23.1
0.05	0.1	22.3	42.0	21.8
0.2	0.1	**24.2**	**43.7**	23.1

Table 1. Comparison results on Cityscapes→Foggy Cityscapes (%) of different settings of λ_1 and λ_2. We set $\lambda_{1,2} = 0.1$ in the experiments of the manuscript as 2^{nd} line.

A.2. Position Sensitivity

We further investigate the position to deploy the Node Discriminator (ND) to align the matched nodes, and record the comparison results in Table 2. We compare three settings for the node alignment, i.e., P1: semantic-complete nodes $\mathcal{V}_{s/t}$ (without the hallucination nodes), P2: enhanced nodes after graph convolution $\tilde{\mathcal{V}}_{s/t}$, and P3: the nodes after Cross Graph Interaction (CGI) $\hat{\mathcal{V}}_{s/t}$. It can be observed that performing the alignment on the semantic-complete nodes (P1) achieves the best results with well-aligned node pairs. Besides, we find a significant performance drop on P3 because the proposed CGI will exchange information across domains, confusing the discriminator and harming the adversarial alignment. Hence, aligning nodes in P1 is optimal in the proposed method as the setting in our manuscript.

Pos.	prsn	rider	car	truc	bus	train	moto	bike	mAP
P1	**46.9**	**48.4**	**63.7**	27.1	50.7	35.9	**34.7**	**41.4**	**43.5**
P2	43.9	46.0	57.0	**29.7**	**53.9**	**39.7**	34.6	39.6	43.0
P3	44.0	45.4	57.2	25.2	48.4	26.8	27.5	38.7	39.2

Table 2. Comparison results on Cityscapes→Foggy Cityscapes (%) by deploying the ND on different nodes, i.e., semantic-complete nodes $\mathcal{V}_{s/t}$ (P1), enhanced nodes after graph convolution $\tilde{\mathcal{V}}_{s/t}$ (P2), and the nodes after cross graph interaction $\hat{\mathcal{V}}_{s/t}$ (P3).

A.3. Normalization Sensitivity

The proposed method transforms the visual feature to the graphical space (V2G) with a projection module (Fc-Norm-ReLU-Fc). Hence, we present a comparison among different projection strategies with different normalization (Norm) tricks, including Group Normalization (commonly used in the FCOS [16] detection head), Batch Normalization (commonly used in the ResNet [4] backbone network), and Layer Normalization [1], as shown in Table 3. Our projection design with Layer Normalization works better on node embedding than other common settings, preserving node-based correspondence and achieving the best adaptation result (43.5% mAP).

*Yixuan Yuan is the corresponding author.
This work was supported by Hong Kong Research Grants Council (RGC) General Research Fund 11211221 (CityU 9043152).

Pos.	prsn	rider	car	truc	bus	train	moto	bike	mAP
GN	45.7	44.9	63.1	24.8	48.3	43.2	32.6	40.9	42.9
BN	46.1	42.8	61.7	**27.6**	45.5	34.8	32.0	38.0	41.0
LN	**46.9**	**48.4**	**63.7**	27.1	**50.7**	**35.9**	**34.7**	**41.4**	**43.5**

Table 3. Comparison results on Cityscapes→Foggy Cityscapes (%) of different normalization strategies in the vision-to-graph (V2G) transformation.

B. Discussion

B.1. Baseline Selection

Two-stage v.s. single-stage baselines. Two-stage object detectors, e.g., Faster RCNN [10], consist of a feature extractor, a Region Proposal Network (RPN) and a detection head for classification and regression. These approaches first adopt RPN on image features to obtain Region of Interests (RoIs), and then perform detection based on these region proposals. Differently, single-stage object detectors [9, 16] only contain a feature extractor and detection head, and these approaches directly make prediction on image features without RPN.

Reasons for the singe-stage baseline. In this paper, we mainly focus on the domain adaptation for singe-stage object detectors as lots of recently published works [2,5–8,15], and we select the single-stage detector as the baseline because of the following two main reasons.

1) Discarding RPN. Most adaptation works [18, 20, 21] perform adaptation on both image features and RoI representations, which highly rely on the RPN and are limited to the two-stage detectors. In contrast, our method achieves fine-grained adaptation only using image features and totally discards the RPN, yielding enormous potentials to be generalized to different baselines. Hence, we use the single-stage baseline free of RPN in our method to demonstrate the advantages without bells and whistles.

2) Fair comparison. The fairness and agreement of the benchmark comparison have been proven in recently published literature [2, 5, 7, 8, 15] for single-stage object detectors due to the comparable source only results and adaptation gains. Besides, we also report the fair adaptation gains in benchmark comparison to demonstrate our effectiveness in terms of domain adaptation. Moreover. most of the latest adaptation works [2,5,7,8,15] are based on the single-stage detectors [9, 16], and we aim to present a comparison with them using same baseline model.

Potentials for the two-stage extension. We psropose a Graph-embedded Semantic Completion module (GSC) to complete the mismatched semantics and leverage a Bipartite Graph Matching adaptor (BGM) to achieve fine-grained adaptation on image features. These two modules are totally independent of the detection baseline types and can be effortlessly extended to different baselines by deploying on the features extracted from backbone networks.

B.2. Limitation

Though the proposed model could achieve state-of-the-art results, it may have some failure cases (Figure 1) due to the limited visual features. As shown in 1^{st} and 2^{nd} row, we find that our method may miss and wrongly detect some distant objects obscured by heavy fog, e.g., the missing truck (1^{st} row) and the wrongly detected person (2^{nd} row), due to the poor visual features caused by the tiny scale (long distance) and low-quality appearance (heavy fog). This problem can be solved from two aspects, i.e., improving visual representations and compensating for visual features with other cues. On the one hand, we can use more robust backbone networks, e.g., ResNet-101 [4], to obtain better features than the VGG-16 backbone [12]. On the other hand, we can establish graph matching between visual and linguistic cues [19] to compensate for the limited visual features with extra semantics.

Figure 1. Illustration of some failure examples compared between (a) the proposed SIGMA framework and (b) ground-truth.

C. Implementation Details

C.1. Discriminator Architecture

As shown in Table 4, we present the detailed architecture of the adversarial alignment module in our SIGMA framework, which includes the loss terms $\mathcal{L}_{GA}$ and $\mathcal{L}_{NA}$. We adopt image-level global alignment [3] using the Global Discriminator as [3, 5–7, 15, 17, 18, 20]. Then, we introduce a node discriminator to align well-match graph nodes, as illustrated in the bottom part of Table 4. Considering the graph nodes refactor the image-level spatial correspondence with edge connections, we replace the convolution layers with fully-connected layers. Besides, we change the Group Normalization (GroupNorm) with Layer Normalization (LayerNorm) due to the advantage of operating the node-based representation, as in Sec. A.3.

Global Discriminator [5]
Gradient Reversal Layer (GRL)
Conv $256 \times 3 \times 3$, stride 1 $\rightarrow$ GroupNorm $\rightarrow$ ReLU
Conv $256 \times 3 \times 3$, stride 1 $\rightarrow$ GroupNorm $\rightarrow$ ReLU
Conv $256 \times 3 \times 3$, stride 1 $\rightarrow$ GroupNorm $\rightarrow$ ReLU
Conv $256 \times 3 \times 3$, stride 1 $\rightarrow$ GroupNorm $\rightarrow$ ReLU
Conv $1 \times 3 \times 3$, stride 1
Node Discriminator (ours)
Gradient Reversal Layer (GRL)
Fc 256 $\rightarrow$ LayerNorm $\rightarrow$ ReLU
Fc 256 $\rightarrow$ LayerNorm $\rightarrow$ ReLU
Fc 256 $\rightarrow$ LayerNorm $\rightarrow$ ReLU
Fc 1 $\rightarrow$ LayerNorm $\rightarrow$ ReLU

Table 4. Architectures of the adversarial alignment modules.

C.2. Implementation and Training

1) Different blocks. The non-linear projection layer used in the vision-to-graph (V2G) transformation is deployed with a Fc-LayerNorm-ReLU-Fc block, and the classifier for node classification is Fc-ReLU-Fc.

2) Dropout rate. The dropout rate is set 0.1 for the edge-drop [11] to avoid the potential visual bias.

3) Spectral clustering. For the learning of the graph-guided memory bank, we perform spectral clustering if the number of nodes is larger than 5 to ensure the clustering reliability. Besides, we replace the Laplacian affinity [14] with K-Nearest Neighbor (K=5) in the clustering algorithm, which reduces the time-consuming significantly.

4) End-to-end training. Our method can achieve end-to-end training without the warm-up stage. We utilize halved source nodes as the placeholders if no nodes appear in the target domain to train our matching module and introduce extra 10,000 iterations for training, which can achieve the same results as the warm-up-included strategy.

5) Multiple matching. The detailed implementation of the multiple-matching ablation study (in Table 5 of our manuscript) is as follows,

$$\mathcal{L}_{mat} = Loss[sigmoid(\mathbf{M}_{\text{aff}}), \mathbf{Y}_{\mathbf{\Pi}}], \tag{1}$$

where $\mathbf{M}_{\text{aff}}$ is the node affinity without adopting Instance Normalization and the Sinkhorn [13] layer, and $Loss[A, B]$ can be selected as the BCE and MSE loss to evaluate the difference between A and B.

Algorithm 1 Semantic-complete Graph Matching

Input:
$\mathcal{I}_{s/t}$: source and target images
$\mathcal{Y}_s$: source annotations
$\lambda_{1,2}$: hyperparameters in the loss function
Output:
Domain adaptive object detector Θ
for $l = 1$ **to** $maxiter$ **do**
extract image features $\mathcal{F}_{s/t}$ witn backbone networks;
generate global alignment loss $\mathcal{L}_{GA}$ on $\mathcal{F}_{s/t}$;
send $\mathcal{F}_{s/t}$ to the detection head to generate $\mathcal{L}_{det}$ with $\mathcal{F}_s$ and classification maps $\mathcal{M}_t$ with $\mathcal{F}_t$;
Graph-embedded Semantic Completion (GSC)
perform V2G transformation to obtain nodes $\mathcal{V}_{s/t}^{raw}$;
generate node alignment loss $\mathcal{L}_{NA}$;
perform DNC for semantic-complete nodes $\mathcal{V}_{s/t}$;
establish graphs $\mathcal{G}_{s/t}$ and perform GCN for $\tilde{\mathcal{V}}_{s/t}$;
update GMB with enhanced nodes $\tilde{\mathcal{V}}_{s/t}$;
Bipartite Graph Matching (BGM)
perform CGI obtaining $\hat{\mathcal{V}}_{s/t}$ and generate loss $\mathcal{L}_{node}$;
perform SNA matrix learning to obtain $\tilde{\mathbf{M}}_{\text{aff}}$;
generate graph matching $\mathcal{L}_{mat}$;
Network Parameter Updating
use $\mathcal{L} = \lambda_1\mathcal{L}_{node} + \lambda_2\mathcal{L}_{mat} + \mathcal{L}_{NA} + \mathcal{L}_{GA} + \mathcal{L}_{det}$ to update network parameters with backpropagation;
end for
return Domain adaptive object detector Θ;

C.3. Optimization Pipeline

The overall optimization pipeline of the proposed SIGMA framework is shown in Algorithem 1. Given the source and target images $\mathcal{I}_{s/t}$, source annotations $\mathcal{Y}_s$, and some predefined hyperparameters $\lambda_{1,2}$, we implement the SIGMA framework to obtain a domain adaptive object detector Θ with $maxiter$ iterative training.

D. Qualitative Results

D.1. Matching Visualization

As shown in Figure 2, we visualize the learned doubly stochastic node affinity matrix $\tilde{\mathbf{M}}_{\text{aff}}$ and the ground-truth matrix $\mathbf{Y}_{\mathbf{\Pi}}$ (Refer to Figure 2 of the manuscript for better understanding.). Each activated entry $\tilde{\mathbf{M}}_{\text{aff}}^{i,j}$ represents a matched node pair across domains, and each activated entry $\mathbf{Y}_{\mathbf{\Pi}}^{i,j} = 1$ (marked in red) indicates that the source node $\hat{v}_s^i$ and the target counterpart $\hat{v}_t^j$ are in the same category. Based on the proposed structure-aware matching loss, each source node successfully find an optimal target node in the same category adaptively and match it to achieve graph-matching-based adaptation.

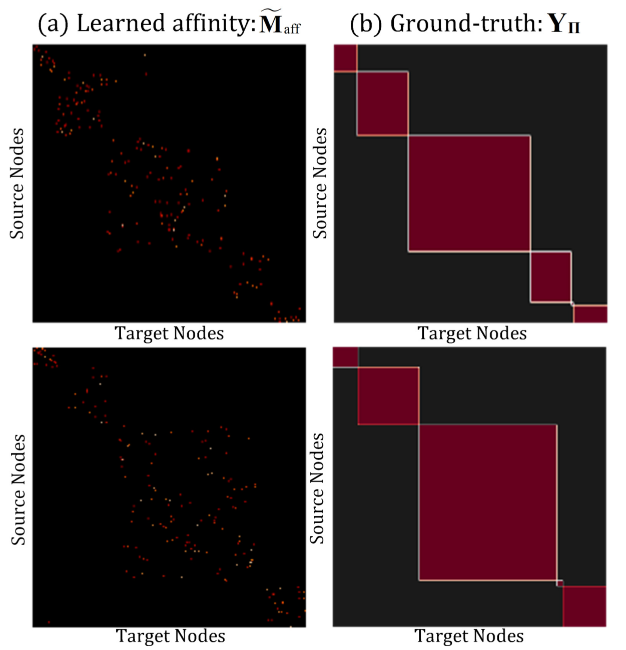

Figure 2. Illustration of (a) the learned doubly stochastic affinity matrix $\tilde{\mathbf{M}}_{\text{aff}}$ and (b) the ground-truth $\mathbf{Y}_{\mathbf{\Pi}}$. Each activated entry $\tilde{\mathbf{M}}_{\text{aff}}^{i,j}$ represents an adaptive matching between the source node $\hat{v}_s^i$ and target node $\hat{v}_t^j$. Each positive entry $\mathbf{Y}_{\mathbf{\Pi}}^{i,j}$ (marked in red) indicates that the node $\hat{v}_s^i$ and $\hat{v}_t^j$ are in the same category.

D.2. Qualitative Comparison

We present more qualitative comparisons among (a) source only, (b) EPM [5], (c) the proposed SIGMA, and (d) ground-truth in Figure 3-6. Our method can eliminate some missing errors (false-negative cases) and avoid some wrong classification cases (false-positive cases) compared with the class-agnostic method EPM [5], which verifies the effectiveness of aligning class-conditional distributions.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 1

[2] Chaoqi Chen, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. I3net: Implicit instance-invariant network for adapting one-stage object detectors. In *CVPR*, pages 12576–12585, 2021. 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] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 2

[5] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *ECCV*, pages 733–748, 2020. 2, 3, 4, 5, 6, 7, 8

[6] Chuang Lin, Zehuan Yuan, Sicheng Zhao, Peize Sun, Changhu Wang, and Jianfei Cai. Domain-invariant disentangled network for generalizable object detection. In *ICCV*, pages 8771–8780, October 2021. 2

[7] Muhammad Akhtar Munir, Muhammad Haris Khan, M Saquib Sarfraz, and Mohsen Ali. Synergizing between self-training and adversarial learning for domain adaptive object detection. 2021. 2

[8] Rindra Ramamonjison, Amin Banitalebi-Dehkordi, Xinyu Kang, Xiaolong Bai, and Yong Zhang. Simrod: A simple adaptation method for robust object detection. In *ICCV*, pages 3570–3579, 2021. 2

[9] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

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

[11] Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. Dropedge: Towards deep graph convolutional networks on node classification. *ICLR*, 2020. 3

[12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[13] Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. math. stat.*, 35:876–879, 1964. 3

[14] X Yu Stella and Jianbo Shi. Multiclass spectral clustering. In *ICCV*, volume 2, pages 313–313. IEEE Computer Society, 2003. 3

[15] Kun Tian, Chenghao Zhang, Ying Wang, Shiming Xiang, and Chunhong Pan. Knowledge mining and transferring for domain adaptive object detection. In *ICCV*, pages 9133–9142, October 2021. 2

[16] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, pages 9627–9636, 2019. 1, 2

[17] 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, June 2021. 2

[18] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *CVPR*, pages 12355–12364, 2020. 2

[19] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, pages 14393–14402, 2021. 2

[20] Yixin Zhang, Zilei Wang, and Yushi Mao. Rpn prototype alignment for domain adaptive object detector. In *CVPR*, pages 12425–12434, June 2021. 2

[21] Yangtao Zheng, Di Huang, Songtao Liu, and Yunhong Wang. Cross-domain object detection through coarse-to-fine feature adaptation. In *CVPR*, pages 13766–13775, 2020. 2

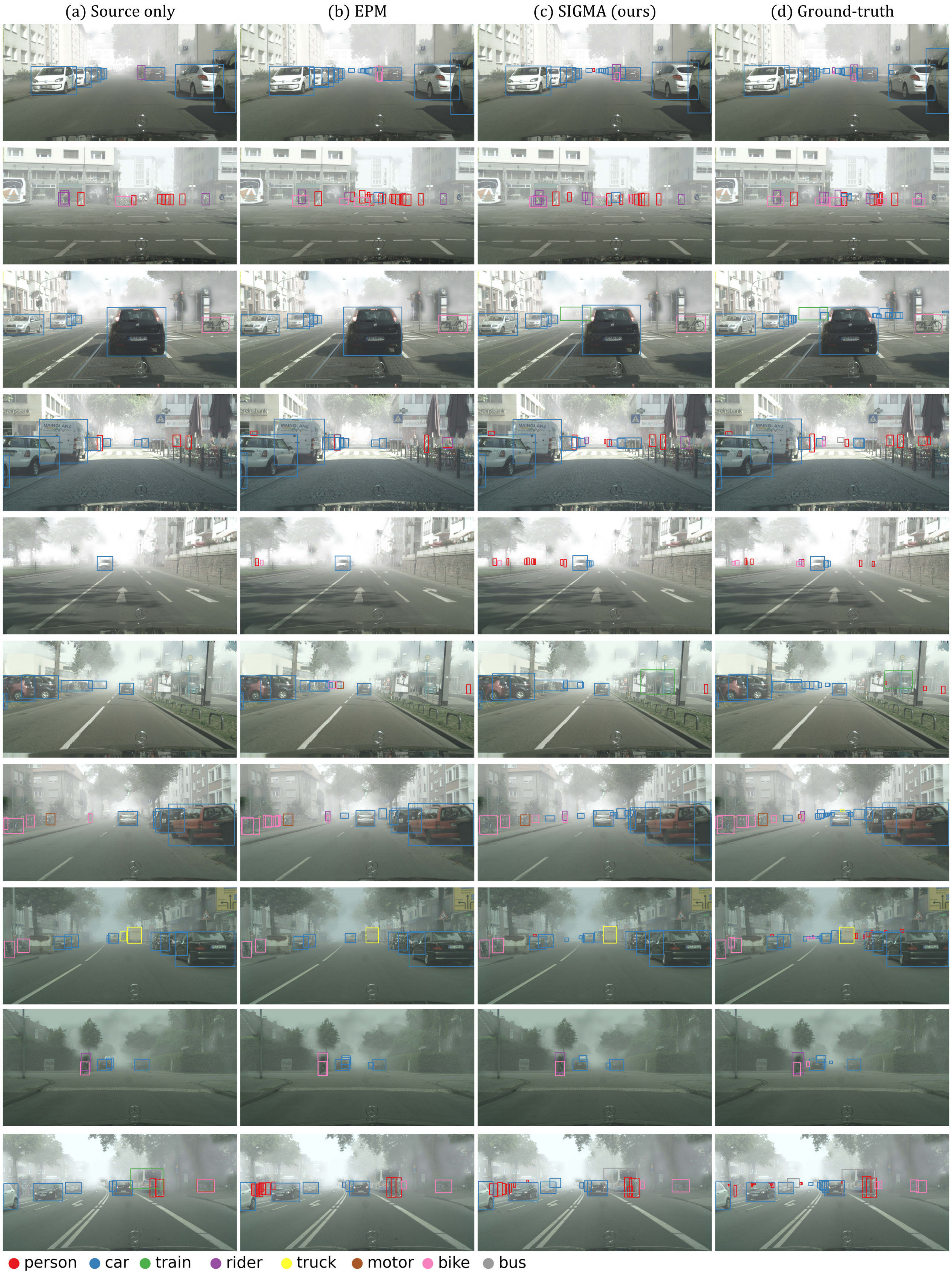

Figure 3. Qualitative reustks on the Cityscapes→Foggy Cityscapes adaptation scenario of (a) the source only model, (b) EPM [5], (c) the proposed SIGMA, and (d) Ground-truth. (Zooming in for best view.)

Figure 4. Qualitative results on the Cityscapes→Foggy Cityscapes adaptation scenario of (a) the source only model, (b) EPM [5], (c) the proposed SIGMA, and (d) Ground-truth. (Zooming in for best view.)

Figure 5. Qualitative results on the Cityscapes→Foggy Cityscapes adaptation scenario of (a) the source only model, (b) EPM [5], (c) the proposed SIGMA, and (d) Ground-truth. (Zooming in for best view.)

Figure 6. Qualitative results on the Cityscapes→Foggy Cityscapes adaptation scenario of (a) the source only model, (b) EPM [5], (c) the proposed SIGMA, and (d) Ground-truth. (Zooming in for best view.)