DA-DETR: Domain Adaptive Detection Transformer with Information Fusion (Supplementary Materials)

A. Discussions

A.1. Parameter Analysis

We study the balance weight λ as defined in Eq.10 in the main text as well as the sensitivity of the parameter K that determines the number of split groups in the proposed Split-Merge Fusion (SMF).

Parameter K. For parameter K, we study its sensitivity by changing it from 1 to 64. Since the number of channels C should be divisible by the number of groups K, we select $K = (1, 4, 8, 16, 32, 64)$, where $K = 1$ means no splitting and $K = 64$ means each group of feature contains 4 channels. The experiments are conducted over domain adaptation tasks Cityscapes $\rightarrow$ Foggy Cityscapes and SIM 10k $\rightarrow$ Cityscapes, and Tables 1 and 2 show the experiment results. It can be seen that the performance of DA-DETR is quite tolerant to parameter K and the best performance is obtained when $K = 32$.

	K (the number of groups in SMF)					
Scenario	1	4	8	16	32	64
Weather	41.7	42.3	43.1	43.3	**43.5**	43.2

Table 1. Parameter K affects domain adaptation in scenario Normal weather to Foggy weather: Cityscapes $\rightarrow$ Foggy Cityscapes (in mAP).

	K (the number of groups in SMF)					
Scenario	1	4	8	16	32	64
Scene	52.3	53.1	53.6	54.1	**54.7**	54.2

Table 2. Parameter K affects domain adaptation in scenario Synthetic scene to Real scene: SIM 10k $\rightarrow$ Cityscapes (in mAP).

Balance weight λ. For balance weight λ, we study its sensitivity by changing it from 0.01 to 1.0. The experiments are conducted over the tasks Cityscapes $\rightarrow$ Foggy Cityscapes and SIM 10k $\rightarrow$ Cityscapes. As shown in Tables 3 and 4, the performance of DA-DETR is quite tolerant to λ and the best detection performance is obtained when λ is set at 0.1.

	λ					
Scenario	0.01	0.02	0.05	0.1	0.5	1.0
Weather	43.0	42.9	43.3	**43.5**	43.2	43.1

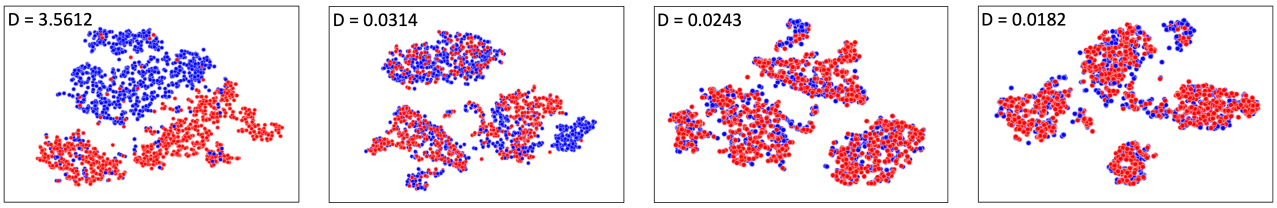
Table 3. Parameter λ affects domain adaptation in scenario Normal weather to Foggy weather: Cityscapes $\rightarrow$ Foggy weather (in mAP).

	λ					
Scenario	0.01	0.02	0.05	0.1	0.5	1.0
Scene	54.1	53.8	54.2	**54.7**	54.2	54.1

Table 4. Parameter λ affects domain adaptation in scenario Real scene to Synthetic scene: SIM 10k $\rightarrow$ Cityscapes (in mAP).

A.2. Effectiveness of CNN-Transformer Blender (CTBlender)

We study how the proposed CTBlender helps to align cross-domain features over task Cityscapes $\rightarrow$ foggy cityscapes. As Figs. 1(b) and (c) show, the direct alignment with two types of features including CNN features f and Transformer features p helps to reduce inter-domain distance as compared with that of the original features (without domain adaptation) as shown in Fig. 1(a). With the proposed CTBlender, the generated source and target features are better aligned (with smaller inter-domain distance D) as shown in Fig. 1(d), demonstrating that CTBlender helps to align cross-domain features effectively.



(a) Original features (b) Alignment (CNN features) (c) Alignment (Transformer features) (d) Adapted by DA-DETR

Figure 1. The t-SNE [8] visualization of feature representations under the scenario Normal weather to Foggy weather (Cityscapes $\rightarrow$ Foggy cityscapes): Red points represent source features and blue points represent target features. D denotes the distance between source and target feature representations as measured by Maximum Mean Discrepancy [4]. Direct alignment by CNN features or Transformer features helps to reduce inter-domain distance as shown in (b) and (c). The proposed DA-DETR can further reduce inter-domain distance clearly as shown in (d).

A.3. Discriminator Analysis

In the proposed CTBlender, the fused features are fed to the discriminator C_d for inter-domain alignment via adversarial learning. The discriminator C_d simply consists of two 1x1 convolutional layers as in the prior work [2]. We conduct new experiments to discuss how different discriminator structures affect the model performance. The experiments are conducted over domain adaptation tasks Cityscapes $\rightarrow$ Foggy Cityscapes on DETR [13]. As Table 5 shows, the proposed DA-DETR is tolerant to different discriminator structures and can achieve superior domain adaptation performance consistently.

Methods	Architecture of Discriminator	mAP
DA-DETR	3 1x1 conv. layers [11]	42.8
DA-DETR	3 3x3 conv. layers and 1 linear layer [10]	43.1
DA-DETR	1 linear layer [7]	42.3
DA-DETR	2 1x1 conv. layers (default)	**43.5**

Table 5. Different discriminator designs vs DA-DETR performance over task Cityscapes $\rightarrow$ Foggy Cityscapes.

A.4. Comparison with Other UDA Methods

Besides adversarial alignment, we compare the proposed DA-DETR with another two recent UDA-based detection approaches, $i.e.$, self-training approach that iteratively pseudo-labels target samples in network training and image translation approach that mitigates domain gaps by modifying source images to have target-like image styles. Table 6 shows experimental results. It can be observed that the proposed DA-DETR achieves competitive performance as compared with the compared self-training methods and image-translation methods. Note many prior studies combine different UDA approaches. For example, [3, 12] exploits both adversarial alignment and self-training. [5, 6] combine image translation and adversarial

alignment for UDA. In addition, self-training and image translation are often more complicated and computational intensive. For example, self-training usually involves multiple training iterations with online/offline pseudo-labeling. Image translation usually requires a separate process to train specific image translation models before UDA, etc.

Cityscapes → Foggy cityscapes											
Method	Type	Backbone	person	rider	car	truck	bus	train	mcycle	bicycle	mAP
DETR [13]	None	ResNet-50	37.7	39.1	44.2	17.2	26.8	5.8	21.6	35.5	28.5
DAM [6]	IT & AL	ResNet-50	46.7	43.5	60.2	24.3	36.8	28.3	27.3	42.1	38.7
Progressive DA [5]	IT & AL	ResNet-50	49.2	48.9	60.1	22.5	40.8	31.6	25.7	42.5	40.2
UMT [3]	AL & ST	ResNet-50	49.5	49.2	62.1	26.1	43.2	29.5	31.2	43.9	41.8
CST [12]	AL & ST	ResNet-50	49.3	48.7	59.3	26.3	41.2	34.6	27.8	42.8	41.3
SimROD [9]	ST	ResNet-50	48.8	48.9	61.5	**27.7**	41.9	35.9	28.9	43.2	42.1
DA-DETR	AL	ResNet-50	**49.9**	**50.0**	**63.1**	24.0	**45.8**	**37.5**	**31.6**	**46.3**	**43.5**

Table 6. Comparing the proposed DA-DETR with self-training methods and image-translation methods. IT, AL and ST stand for image translation, adversarial learning and self-training, respectively.

B. Qualitative Results

We perform qualitative experiments as well and Fig. 2 shows experimental results. We can observe that the baseline model DETR [13] produces a number of false detection due to domain gaps. State-of-the-art domain adaptation method [7] generates more precise bounding boxes but tends to miss many small objects. The proposed DA-DETR adapts well under all four scenarios and can detect more small objects with less false alarms as illustrated.

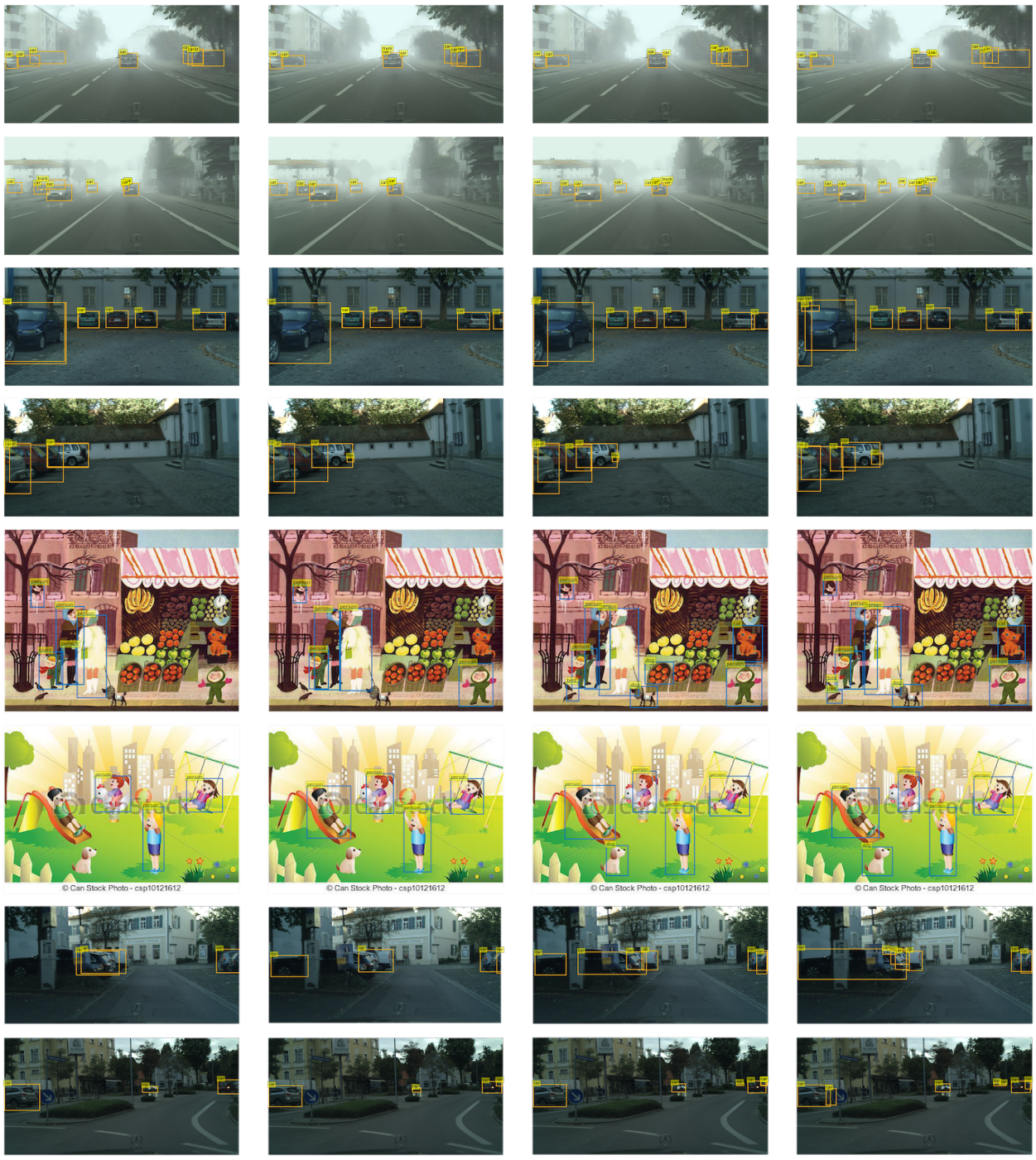


Figure 2. Qualitative comparison of DA-DETR with DETR [13] and SAP [7] over four domain adaptive detection benchmarks including Cityscapes → Foggy cityscapes as in the first and second rows, SIM 10k → Cityscapes as in the third and forth rows, PASCAL VOC → Clipart1k as in the fifth and sixth rows and KITTI → Cityscapes as in last two rows, respectively. DA-DETR outperforms DETR and SAP consistently by detecting more accurate boxes across all sample images.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 4

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

[3] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4091–4101, 2021. 2, 3

[4] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012. 2

[5] Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih Hung, Hung-Yu Tseng, Maneesh Singh, and Ming-Hsuan Yang. Progressive domain adaptation for object detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 749–757, 2020. 2, 3

[6] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12456–12465, 2019. 2, 3

[7] Congcong Li, Dawei Du, Libo Zhang, Longyin Wen, Tiejian Luo, Yanjun Wu, and Pengfei Zhu. Spatial attention pyramid network for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 481–497. Springer, 2020. 2, 3, 4

[8] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 2

[9] Rindra Ramamonjison, Amin Banitalebi-Dehkordi, Xinyu Kang, Xiaolong Bai, and Yong Zhang. Simrod: A simple adaptation method for robust object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3570–3579, 2021. 3

[10] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6956–6965, 2019. 2

[11] Zhiqiang Shen, Harsh Maheshwari, Weichen Yao, and Marios Savvides. Scl: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses. *arXiv preprint arXiv:1911.02559*, 2019. 2

[12] Ganlong Zhao, Guanbin Li, Ruijia Xu, and Liang Lin. Collaborative training between region proposal localization and classification for domain adaptive object detection. In *European Conference on Computer Vision*, pages 86–102. Springer, 2020. 2, 3

[13] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2, 3, 4