Improving Transferability for Domain Adaptive Detection Transformers
Supplementary Material

Kaixiong Gong
Beijing Institute of Technology
Beijing, China
kxgong@bit.edu.cn

Shuang Li*
Beijing Institute of Technology
Beijing, China
shuangli@bit.edu.cn

Shugang Li
Beijing Institute of Technology
Beijing, China
shugangli@bit.edu.cn

Rui Zhang
Beijing Institute of Technology
Beijing, China
zhangrui20@bit.edu.cn

Chi Harold Liu
Beijing Institute of Technology
Beijing, China
liuchi02@gmail.com

Qiang Chen
Baidu VIS
Beijing, China
chenqiang13@baidu.com

A MORE IMPLEMENTATION DETAILS

A.1 Backbone Adversarial Alignment

Our method conducts adversarial domain alignment on the features of CNN backbone for achieving domain-invariant based features. The feature maps from different stages of CNN backbone are utilized for adversarial alignment to ensure the detection performance of objects of various scales, i.e., the feature maps of stage C_3, C_4, C_5 and an extra feature map generated by applying a 3×3 convolution with stride 2 on the feature map from C_5. All these features maps from four stages are fed into a shared CNN based domain discriminator depicted in the following. The domain discriminator and the CNN backbone network play a minimax game for adversarially aligning the backbone features from two domains.

A.2 Structure of Domain Discriminator

In this work, a CNN based domain discriminator is utilized for performing adversarial alignment, which enhances the cross-domain performance of detectors. The domain discriminator is a two-layer convolutional neural network. The structure is shown in Table 1.

Table 1: Structure of the domain discriminator.

Domain Discriminator
Convolutional 256×512, kernel=1, stride=1
ReLU
Convolutional 512×2, kernel=1, stride=1

A.3 Data Pre-processing

Following Deformable DETR, we randomly flip, resize, and crop the images. And, we adopt color jitter and gaussian blur for data augmentation to enhance the robustness of models.

B EXPERIMENT ON CONDITIONAL DETR

We build our method on Conditional DETR [4] to verify whether the proposed method can boost the cross-domain performance of other DETR-style detectors. Conditional DETR differs from Deformable DETR [6] in two aspects: 1) Conditional DETR exploits one CNN backbone feature map instead of the multi-scale feature maps used in Deformable DETR and 2) it still uses the global dense attention with proposed conditional cross-attention mechanism, while Deformable DETR introduces a sparse attention scheme that only queries a small set of keys.

B.1 Implementation Details

We implement our method based on the code released by the authors. The backbone feature alignment is only performed on the single-scale backbone features. The CNN backbone is a ResNet-50 [2] network pre-trained on ImageNet [1]. The default learning rates are utilized for training the detector with Adam optimizer, 1×10^{-4} for the transformer and 1×10^{-5} for the backbone. We set the learning rate for the domain discriminator as 2×10^{-3}. The experiments are conducted on the Cityscapes → Foggy Cityscapes.

B.2 Results and Analysis

Experimental results are presented in Table 2. One can observe that Conditional DETR is an inferior baseline in domain shift settings compared with Deformable DETR. The interpretation might be that the single-level feature map fails to capture the patterns of small objects and is more fragile to the domain shift. Even though Conditional DETR offers a relatively low initial performance, the proposed alignment modules still substantially facilitate its performance, reaching 39.8 mAP. This manifests the effectiveness and versatility of our method.

C SENSITIVITY TO PRE-TRAINING

Pre-training the network model using source data is typically used for obtaining a reliable initialization for pseudo labels [3]. [3] pre-trains the detector for 12k iterations. Following [3], we pre-train the Deformable DETR model for 10k iterations. In addition, to exam the impact of the iterations of source pre-training, we vary the pre-training iterations from 2k to 10k. The performance results of our method based on various pre-trained models are presented in Fig. 1. One can observe that the final model performance reaches saturation if pre-training the model on the source domain for more iterations, since the initial pseudo labels are more reliable, yielding better detection performance. In addition, when only pre-training the source model for 4k or 6k iterations, the model performances

Table 2: Results of our method based on Conditional DETR [4] on Cityscapes → Foggy Cityscapes. C-DETR is the abbreviation for Conditional DETR.

Method	Detector	person	rider	car	truck	bus	train	mcycle	bicycle	mAP
Conditional DETR (source)	C-DETR	24.4	25.7	38.6	12.1	23.8	10.3	11.6	22.8	21.2
Ours	C-DETR	36.7	44.4	57.5	27.2	40.7	41.9	32.2	38.1	39.8

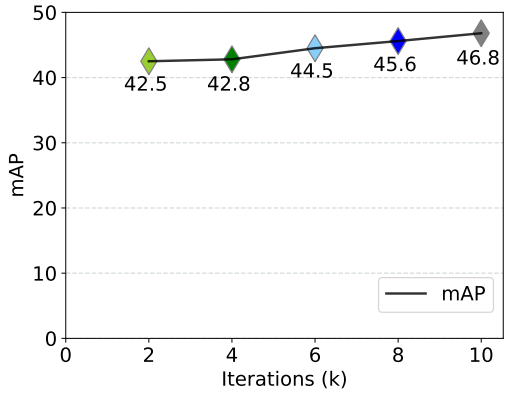

Figure 1: mAPs of our method based on various pre-training models of different training iterations. Experiments are conducted on Cityscapes → Foggy Cityscapes.

are still competitive. These results validate that our method is not sensitive to the pre-training iterations.

D MORE ATTENTION VISUALIZATION

Object-Aware Alignment module is introduced for facilitating the domain alignment on the features of CNN backbone while emphasizing the importance of the foreground regions. Then the detector would put more attention on those object areas thus learning the adaptive object patterns across domains. We present more visualization results for the attention of the backbone network as shown in Fig. 2. From it, one can observe that our method obviously concentrates on those foreground regions, compared with the source model and the global alignment.

E MORE QUALITATIVE RESULTS

Fig. 3 shows the detection results of the source model, SFA [5] and our method, along with the ground-truth bounding boxes. Our method can significantly reduce the false negatives and yield more precise bounding boxes due to the comprehensive feature alignment on the CNN backbone features that contain more details for small objects, and the optimal transport alignment on decoder features that preserve the location information.

REFERENCES

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*. 248–255.
[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.
[3] Junguang Jiang, Baixu Chen, Jianmin Wang, and Mingsheng Long. 2021. Decoupled Adaptation for Cross-Domain Object Detection. *arXiv preprint arXiv:2110.02578* (2021).
[4] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. 2021. Conditional detr for fast training convergence. In *ICCV*. 3651–3660.
[5] Wen Wang, Yang Cao, Jing Zhang, Fengxiang He, Zheng-Jun Zha, Yonggang Wen, and Dacheng Tao. 2021. Exploring Sequence Feature Alignment for Domain Adaptive Detection Transformers. In *ACM MM*. 1730–1738.
[6] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2020. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *ICLR*.

Figure 2: More illustration results of the attention maps of CNN backbone features for testing samples. This figure exhibits the attention maps of source model, global alignment (GA) and our object-aware alignment (OAA). From top to bottom, we show the results on Cityscapes → Foggy Cityscapes, Cityscapes → BDD100k and Sim10k → Cityscapes.

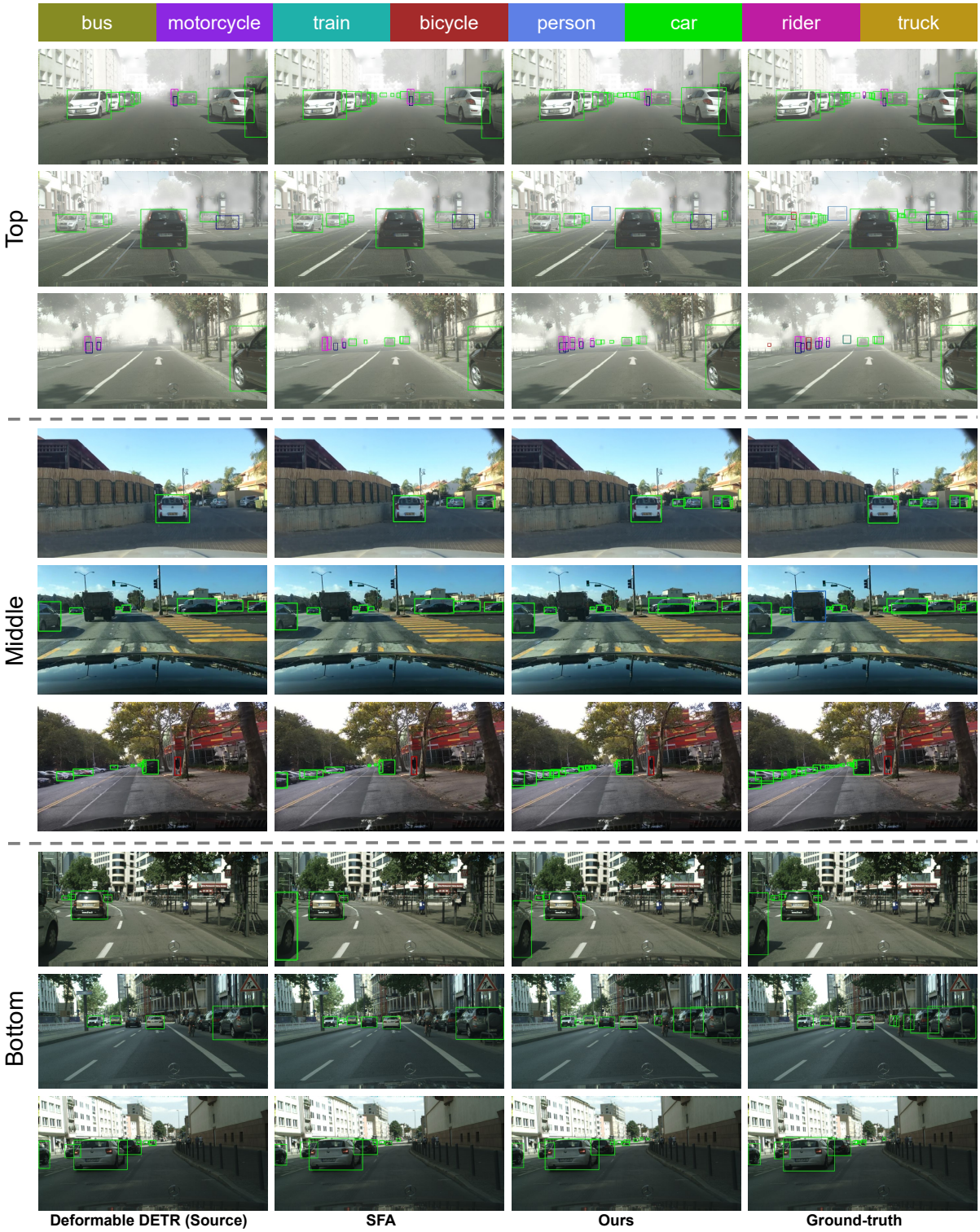

Figure 3: More qualitative results. We visualize the detection results of source model, SFA [5] and our method. From top to bottom, we present the results on Cityscapes → Foggy Cityscapes, Cityscapes → BDD100k and Sim10k → Cityscapes.