Supplementary Material
Bidirectional alignment for domain adaptive detection with transformers

Liqiang He[1], Wei Wang[2], Albert Chen[2], Min Sun[2], Cheng-Hao Kuo[2], and Sinisa Todorovic[1]

[1]Oregon State University, Corvallis, OR, USA
{heli, sinisa}@oregonstate.edu
[2]Amazon, Bellevue, WA, USA
{wweiwan, aycchen, minnsun, chkuo}@amazon.com

This supplementary material complements our main paper with the following additional descriptions and results:

1) BiADT architecture.
2) Sensitivity to loss weights.
3) Ablation of DyHinge loss.
4) Ablation of Mutual information loss.
5) Additional visualizations of our results.

1. BiADT Architecture

Fig. 1 illustrates our BiADT which consists of 6 encoder layers and 6 decoder layers. The input image is passed to the backbone to extract CNN features. Then, in the transformer, the image features along with their positional and domain embeddings are encoded by the 6 stacked encoder layers. The encoder aligns the image domain-invariant features $\mathcal{L}^Y_-$, and makes the image domain-specific features distinct $\mathcal{L}^Y_+$, while simultaneously minimizing the mutual information $\mathcal{L}^Y_{\text{MI}}$ between these two kinds of features. In the decoder, object queries decode the object representations from the encoded image token sequence, and apply the same alignment strategy, *i.e.* minimize the domain alignment losses $\mathcal{L}^X_-, \mathcal{L}^X_+, \mathcal{L}^X_{\text{MI}}$ on object tokens. Finally, a detection loss $\mathcal{L}^X_{\text{det}}$ is applied on the object tokens. We use two "domain-heads" to calculate the losses of $\mathcal{L}_+$ and $\mathcal{L}_-$, respectively. The domain heads $F_{\mathcal{I}}$ and $F_{\mathcal{D}}$ are detailed in Fig. 2. Both are used for proper feature projection and output a probability value between $[0, 1]$.

2. Loss weights sensitivity analysis

Table 4 in the main paper shows the performance contribution of different modules. Eq. (15) in the main paper uses empirically optimized values for the weights λ of each loss function. Tab. 1 shows our sensitivity analysis of these weights in Eq. (15). We evaluate different value combinations for λ^Y_-, λ^Y_+ and λ^X_-, λ^X_+ in terms of mAP. As expected, the weights affect the final accuracy (mAP). From Table 4 in the main paper, we observe the image token alignment improves the mAP score more significantly. Accordingly, Tab. 1 shows that varying the values of λ^Y_-, λ^Y_+ also cause larger variations in performance than the other weights. The best values for the λs are presented in the 6th row of Tab. 1, where both the values of (λ^Y_+ and λ^X_+) are 10 times smaller than λ^Y_- and λ^X_-. The "+" alignment makes the domain-specific features more distinct (absorb domain specific features), and the "−" alignment uses the GRL to enforce the features to become more domain-invariant.

Row	λ^Y_-	λ^Y_+	mAP
1	0.5	0.5	35.9
2	0.5	0.1	38.1
3	0.5	0.01	38.8
4	0.1	0.5	40.7
5	0.1	0.1	42.0
6	0.1	0.01	**44.4**
7	0.01	0.5	40.5
8	0.01	0.1	39.7
9	0.01	0.01	38.3

(a) Image-token alignment

Row	λ^X_-	λ^X_+	mAP
1	0.5	0.5	38.3
2	0.5	0.1	37.8
3	0.5	0.01	40.6
4	0.1	0.5	38.4
5	0.1	0.1	41.3
6	0.1	0.01	**42.9**
7	0.01	0.5	39.6
8	0.01	0.1	41.4
9	0.01	0.01	40.1

(b) Object-token alignment

Table 1: Loss weights sensitivity analysis

3. Ablation of DyHinge loss

Tab. 2 presents a detailed analysis of the performance contribution of DyHinge loss in comparison to BCE (bianry cross-entropy) loss and Hinge loss. As shown in Tab. 2, BCE and Hinge loss with default margin 1.0 have similar performance contributions. However, when we reduce the hinge loss margin to 0.5, the mAP drops from 48.4 to 47.9. This suggests that adjusting an overall margin value

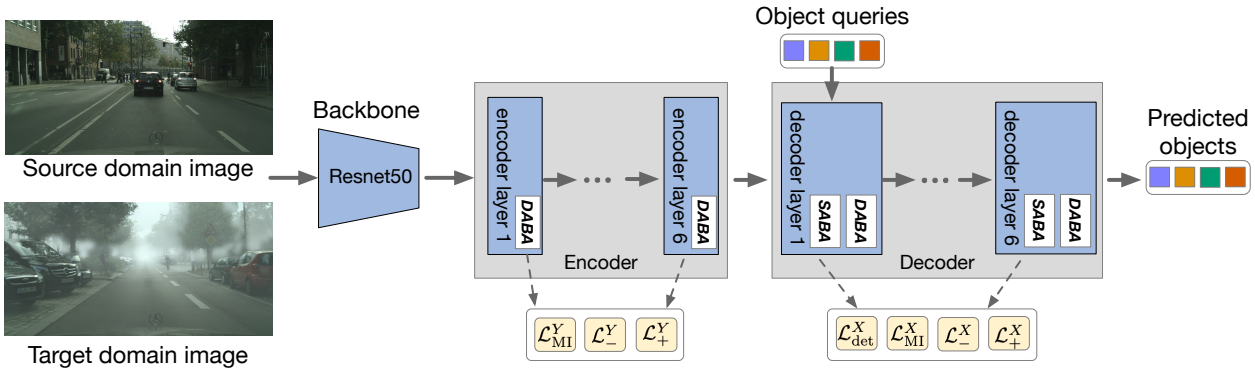
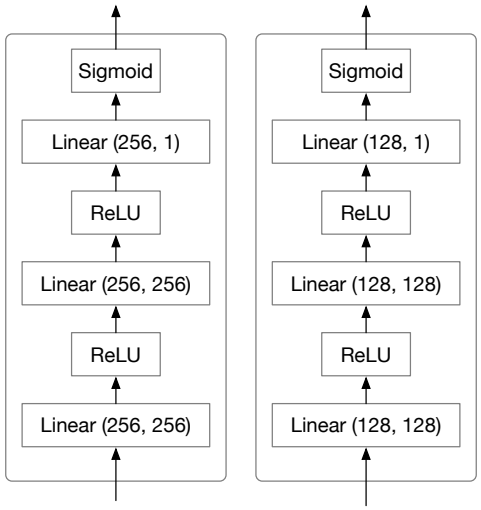


Figure 1: The architecture of our BiADT follows Dab-Deformable-Detr [2], which consists of 6 encoder layers and 6 decoder layers. DABA and SABA represent the "Deformable Attention Bi-Alignment" and "Self Attention Bi-Alignment" illustrated in the main paper.



Domain head $F_{\mathcal{I}}$ Domain head $F_{\mathcal{D}}$

Figure 2: Architecture of domain heads $F_{\mathcal{I}}$ and $F_{\mathcal{D}}$.

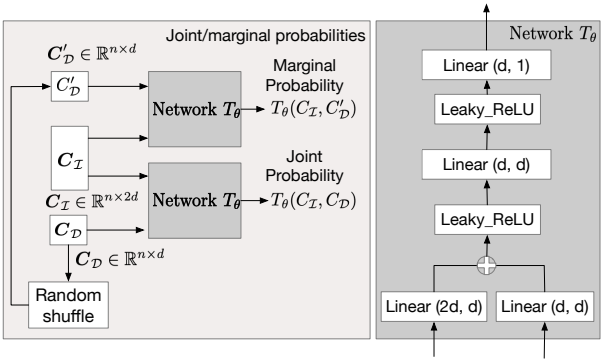


Figure 3: *Left*: Joint and marginal probabilities used in MI (mutual information) [4] loss in Eq.16 in our paper. Two networks T_θ share weights. *Right*: Detailed architecture of Network T_θ. d is set to 128 in the implementation.

Row	loss term	margin	mAP
1	BCE	-	48.2
2	Hinge	1	48.4
3	Hinge	0.5	47.9
4	DyHinge	dynamic	48.7

Table 2: Evaluation of different losses on the alignment of the domain-invariant features.

λ_{MI}	0	5e-6	1e-5	5e-5	1e-4
mAP	48.7	47.3	48.8	**49.4**	46.7

Table 3: Evaluation of different values of the weight λ_{MI}.

may not be a good choice. In contrast, when using the proposed dynamic hinge loss (DyHinge), the mAP increases to 48.7. Therefore, in Figure 2 in our main paper, we say that DyHinge uses the derived "domain shift" from the bottom branch to dynamically determine a suitable level for the alignment of the domain-invariant features in the top branch. The dynamic hinge loss is applied on different tokens adaptively, and thus mitigates the negative transfers on domain-invariant features with weaker domain characteristics.

4. Sensitivity analysis of MI loss

Fig. 3 shows the calculation of the joint and marginal probabilities used in the MI (mutual information) loss adopted from [4]. The left of Fig. 3 is the detailed architecture of Network $T_\theta()$. In contrast to [4], our domain-invariant features have doubled the channel size to that of the domain-specific features. As shown in the right of Fig. 3, we apply two different linear projections to resize them to the same channel dimension.

The sensitivity analysis to the value of λ_{MI} is provided in

Figure 4: Visualization of the predicted domain masks in the source and target domains by BiADT. The predicted domain masks are corresponding to the domain-specific region, *i.e.*, the fog region.

Tab. 3. From this table, there is an empirically optimal λ_{MI}.

5. Qualitative results visualization

Fig. 4 shows additional examples of our results beyond those presented in Figure 5 in the main paper. The first column and the third column show the source domain and the target domain images, the second column and the fourth column show the predicted domain masks by BiADT. As we can see, the predicted domain masks identify reasonably well the domain-specific image regions, *i.e.*, the fog region.

Fig. 5 illustrates the detections made by the following approaches: SFA [3], AQT with DAB-Deformable-Detr backbone [1], and our BiADT. The ground truth bounding boxes are shown in the rightmost column. As we can see, in general, our BiADT misses fewer objects and hence gives a higher true-positive detection rate than the other baseline models.

Fig. 6 shows some corner cases of our BiADT. In the top row, BiADT misses to predict the white trucks, probably because the whiteness of the trucks appears very similar to the fog – *i.e.*, the automatically identified domain characteristic. This is consistent to our results presented in Table 1 in the main paper, where our BiADT has slightly lower accuracy on the category of "truck". In the middle and bottom rows, BiADT fails to predict the train and the bus. One reason might be that the categories of these objects contain significantly fewer instances in the training dataset than the other object categories.

References

[1] 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 *International Joint Conference on Artificial Intelligence (IJCAI)*, 2022. 3, 4

[2] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *International Conference on Learning Representations*, 2022. 2

[3] Wen Wang, Cao Yang, Zhang Jing, He Fengxiang, Zha Zhengjun, Wen Yonggang, and Tao Dacheng. Exploring sequence feature alignment for domain adaptive detection transformers. In *ACM MultiMedia*, 2021. 3, 4

[4] Aming Wu, Yahong Han, Linchao Zhu, and Yi Yang. Instance-invariant domain adaptive object detection via progressive disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

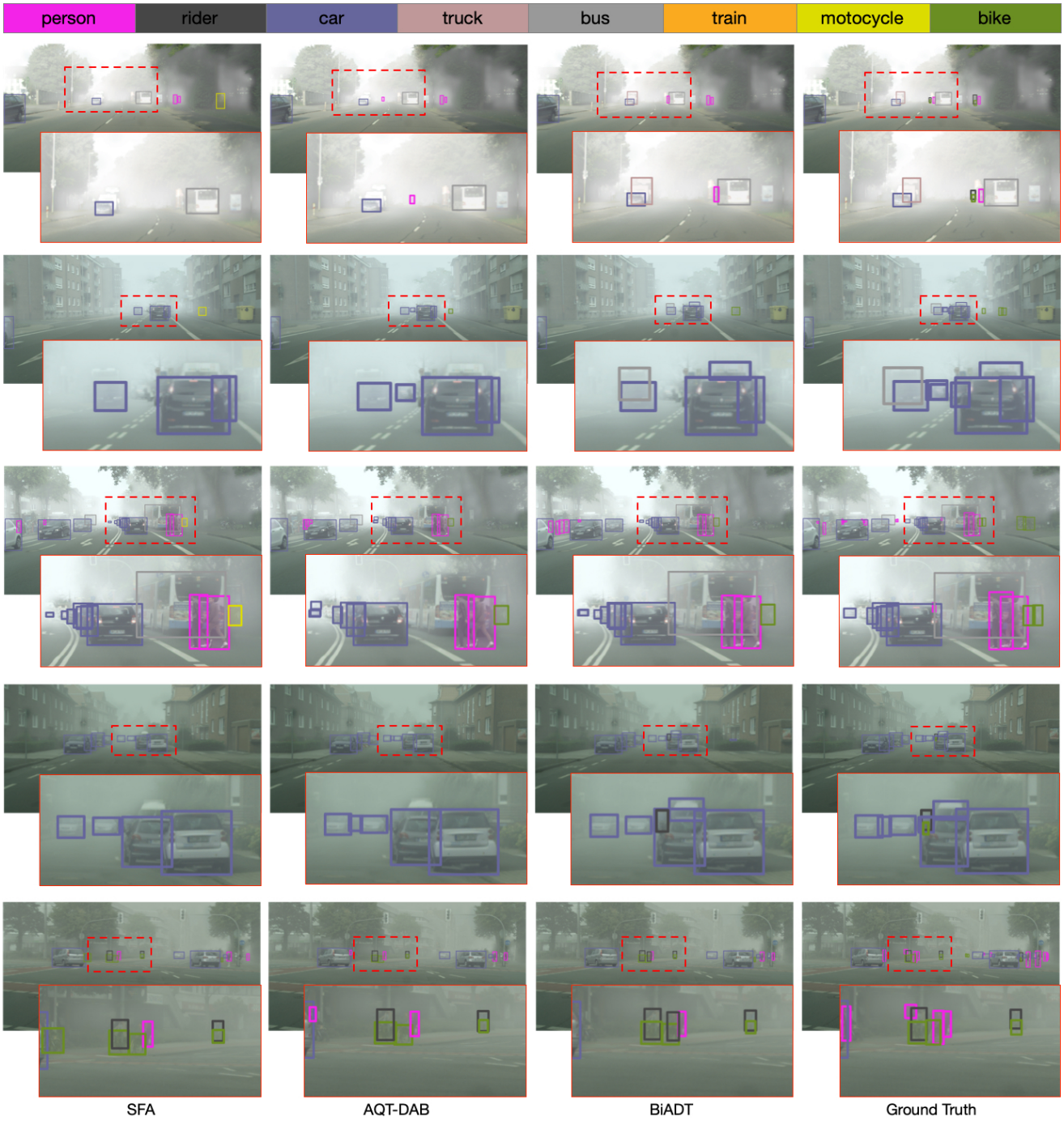


Figure 5: Detection results visualization comparison for (1) SFA [3], (2) AQT with DAB-Deformable-Detr backbone [1], (3) our BiADT and (4) ground truth.

Figure 6: Qualitative visualized results for failure cases of BiADT.