Supplementary Material: Harmonizing Transferability and Discriminability for Adapting Object Detectors

Chaoqi Chen[1], Zebiao Zheng[1], Xinghao Ding[1], Yue Huang[1]*, Qi Dou[2]
[1] Fujian Key Laboratory of Sensing and Computing for Smart City,
School of Informatics, Xiamen University, China
[2] Department of Computer Science and Engineering, The Chinese University of Hong Kong
cqchen94@stu.xmu.edu.cn, zbzheng@stu.xmu.edu.cn
dxh@xmu.edu.cn, huangyue05@gmail.com, qdou@cse.cuhk.edu.hk

1. Additional Details and Analysis

1.1. IWAT-I

Importance Weighted Adversarial Training with input Interpolation (IWAT-I) strategy aims to re-weight the input samples at image-level based on their transferability. The input space is interpolated by the generated source-like target images and target-like source images. The original source images and the generated target-like source images are utilized as the source input, while the original target images and the generated source-like target images are utilized as the target input. Note that the interpolation and the following adaptation paradigm are executed separately. The transferability is evaluated by the uncertainty of the domain discriminator with respect to an input sample without requiring any additional modules.

1.2. CILA

The goal of the proposed Context-aware Instance-Level Alignment (CILA) is to enforce the instance-level alignment between domain based on the fusion of the global context vector and the local instance-level feature. In experiments, the dimension of the aggregated context vector f_c is 384 (the dimension of f_c^1, f_c^2, and f_c^3 are all 128).

In order to tackle the dimension explosion issue, we propose to leverage the randomized methods as an unbiased estimator of the tensor product, *i.e.,*

$$\boldsymbol{f}_{fus} = \frac{1}{\sqrt{d}}(\boldsymbol{R}_1\boldsymbol{f}_c) \odot (\boldsymbol{R}_2\boldsymbol{f}_{ins}) \tag{1}$$

where $\boldsymbol{f}_c$ is the aggregated context vector and $\boldsymbol{f}_{ins}$ is the instance-level feature. $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ are random matrices and each of their element follows the uniform distribution. We utilize the Eq. 1 for the approximation of $\boldsymbol{f}_c \otimes \boldsymbol{f}_{ins}$ (Eq. (4) of our paper). In [2, 4, 1], theoretical findings reveal that adopting randomized multilinear map can accurately approximate the tensor product with bounded estimation variance. Please refer to [2] for the detailed proof.

*Corresponding author

2. Experiments

2.1. More Training Details

In all experiments, the weights of $\mathcal{L}_{la}$, $\mathcal{L}_{ma}$, $\mathcal{L}_{ga}$, $\mathcal{L}_{ins}$ are 1, 0.15, 1, and 0.5 respectively. We adopt focal loss for $\mathcal{L}_{ga}$ and $\mathcal{L}_{ins}$ to further up-weight the hard-to-distinguish samples. $\mathcal{L}_{ma}$ adopts the vanilla cross-entropy loss.

2.2. Feature Visualization

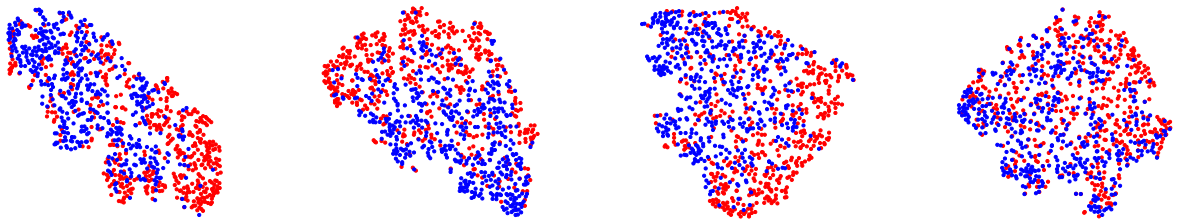
We go deeper into the feature transferability by visualizing in Figure 1(a)-1(d) the deep feature of the network activations on transfer task Cityscapes →Foggy-Cityscapes learned by SWDA [5] (mAP: 34.3), HTCN-w/o CILA (mAP: 36.6), HTCN-w/o Interpolation (mAP: 37.5) and HTCN (mAP: 39.8) respectively using t-SNE [3]. We can see that our feature embedding results (HTCN, HTCN-w/o Interpolation and HTCN-w/o CILA) are consistently much better than the state-of-the-art comparison method (SWDA). Moreover, when one of the proposed components is removed from HTCN, the performance drops.

References

[1] Purushottam Kar and Harish Karnick. Random feature maps for dot product kernels. In *Artificial Intelligence and Statistics*, pages 583–591, 2012.

[2] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *NIPS*, pages 1640–1650, 2018.

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



(a) SWDA (b) HTCN-w/o CILA (c) HTCN-w/o Interpolation (d) HTCN

Figure 1: The t-SNE visualization of network activations on transfer task Cityscapes →Foggy-Cityscapes generated by SWDA (mAP: 34.3), HTCN-w/o CILA (mAP: 36.6), HTCN-w/o Interpolation (mAP: 37.5), and HTCN (mAP: 39.8).

[4] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

[5] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, pages 6956–6965, 2019.