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 $f_1^c$, $f_2^c$, and $f_3^c$ are all 128).

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

$$f_{\text{fus}} = \frac{1}{\sqrt{d}} (R_1 f_c) \odot (R_2 f_{\text{ins}})$$  \hspace{1cm} (1)

where $f_c$ is the aggregated context vector and $f_{\text{ins}}$ is the instance-level feature. $R_1$ and $R_2$ are random matrices and each of their element follows the uniform distribution. We utilize the Eq. (1) for the approximation of $f_c \otimes f_{\text{ins}}$ (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.

2. Experiments

2.1. More Training Details

In all experiments, the weights of $L_{la}$, $L_{ma}$, $L_{ga}$, $L_{ins}$ are 1, 0.15, 1, and 0.5 respectively. We adopt focal loss for $L_{ga}$ and $L_{ins}$ to further up-weight the hard-to-distinguish samples. $L_{ma}$ 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 $\rightarrow$ 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

Figure 1: The t-SNE visualization of network activations on transfer task Cityscapes $\rightarrow$ 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).
