Multi-Target Domain Adaptation with Class-Wise Attribute Transfer in Semantic Segmentation (Supplementary Material)

Changjae Kim†[1]
kimcj5434@gmail.com
Seunghun Lee[2]
lsh5688@dgist.ac.kr
Sunghoon Im*[2]
sunghoonim@dgist.ac.kr

[1] LG Electronics, Korea
[2] DGIST, Korea

1 Training Details

1.1 Training Loss for Translation Network

We train the image translation network θ and multi-head discriminator D by minimizing the total loss L_{trans} as follows:

$$L_{trans} = \lambda_{rec} L_{rec} + \lambda_{adv} L_{adv} + \lambda_{dom} (L^{D}_{dom} + L^{\theta}_{dom}), \tag{1}$$

where $L_{rec}, L_{adv}, L^{D}_{dom}$ and L^{θ}_{dom} are the reconstruction, adversarial, and domain discrimination losses, respectively. We set all the weight terms λ as 1.0.

The reconstruction loss is the L1 distance between the original input images and the reconstructed images $\dot{I}_{\mathcal{S}}, \dot{I}_{\mathcal{T}_k}$ generated by an image decoder g^I as follows:

$$L_{rec} = L_1(\dot{I}_x, I_x). \tag{2}$$

We also impose the adversarial learning between the image translation network and multi-head discriminator [1] to generate target domain images. It consists of an encoder, domain head D'_{enc}, D'_{dom}, and adversarial head D'_{adv}. The adversarial layers and domain classification layers are shorten as follows:

$$D_{adv}(x) = D'_{adv}(D'_{enc}(x)),\ D_{dom}(x) = D'_{dom}(D'_{enc}(x)). \tag{3}$$

The vanilla GAN loss [4] is adopted as the adversarial loss. The translation network tries to maximize and adversarial layers of the discriminator to minimize adversarial loss.

$$L_{adv} = \sum_{k=1}^{N} \Big(E\left[log(D_{adv}(I_{\mathcal{T}_k}))\right] \quad + E\left[1 - log(D_{adv}(I_{\mathcal{S}\to\mathcal{T}_k}))\right] \Big). \tag{4}$$

† Work done in DGIST.
* Corresponding author.

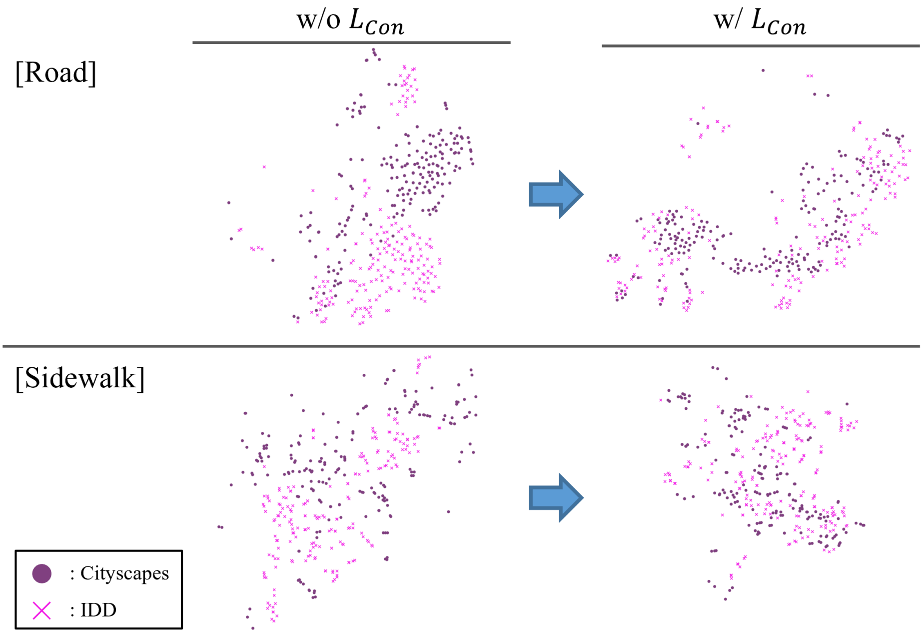

Figure 1: Qualitative ablation study of cross-domain feature consistency. We use a model transferred from GTA5 to Cityscapes, IDD, and Mapillary. We show the results for both Cityscapes and IDD for a clear illustration.

We train the model to generate images with different characteristics for each domain by imposing domain classification loss. The domain classification layers and image translation networks are trained with the domain classification losses L^{D}_{dom} and L^{θ}_{dom} as follows:

$$L^{D}_{dom} = -\sum_{k=1}^{N} t_k log(D_{dom}(I_{\mathcal{T}_k})), \quad L^{\theta}_{dom} = -\sum_{k=1}^{N} t_k log(D_{dom}(I_{\mathcal{S}\to\mathcal{T}_k})), \tag{5}$$

where t_k is the one-hot encoded class label of the target domain $\mathcal{T}_k$. We use the typical cross-entropy loss for the domain classification losses.

1.2 Implementation Details

Following the typical domain adaptive segmentation settings [9], we use the DeepLab-v2 [2] model with ResNet-101 [5] as a backbone pre-trained on ImageNet [3]. We also adopt the same network architecture as [7] for the discriminators. All networks for image translation are trained from scratch. We set the interpolation ratio α as 0.6 for the aggregation of attribute features. We optimize the segmentation model with the SGD optimizer [1] where the weight decay and momentum are set to 0.9 and 5×10^{-4}, respectively. The learning rate is set to 2.5×10^{-4}. The image translation network and discriminator are optimized with the Adam optimizer [6] with momenta of 0.9 and 0.99 and a learning rate set to 10^{-4}. The training procedure is performed with a single RTX A6000 GPU.

2 Additional Study

2.1 Cross-Domain Feature Consistency

In this section, we conduct an additional ablation study on cross-domain feature consistency loss L_{con}. We visualize the learned feature using T-SNE [8] to intuitively show how feature consistency works. We randomly sample features of each class from 1000 images and show the result for road and sidewalk classes of two domains, Cityscapes, and IDD, for clear visibility in Fig. 1. The result without feature consistency indicates the features of each domain are closely distributed, but they are not mixed and maintain separated clusters. The proposed method with the consistency term L_{con} projects the images into the feature space where the two domains are to be mixed regardless of the domain. This shows that the feature consistency term helps to map images into the domain-invariant and class-specific space.

References

[1] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2017.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Ieee, 2009.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[7] Seunghun Lee, Wonhyeok Choi, Changjae Kim, Minwoo Choi, and Sunghoon Im. Adas: A direct adaptation strategy for multi-target domain adaptive semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[8] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[9] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.