Both Style and Distortion Matter: Dual-Path Unsupervised Domain Adaptation for Panoramic Semantic Segmentation – Supplementary Material–

Xu Zheng[2] Jinjing Zhu[1] Yexin Lie[1] Zidong Cao[1] Chong Fu[2,4] Lin Wang[1,3*]
[1]AI Thrust, HKUST(GZ) [2]Northeastern University [3]Dept. of CSE, HKUST
[4]Key Laboratory of Intelligent Computing in Medical Image, Ministry of Education, NEU, China

zhengxu128@gmail.com, zhujinjing.hkust@gmail.com, yliu292@connect.hkust-gz.edu.cn
caozidong1996@gmail.com, fuchong@mail.neu.edu.cn, linwang@ust.hk

Abstract

Due to the lack of space in the main paper, we provide more details of the proposed method and experimental results in the supplementary material. Sec. 1 adds the algorithm of the proposed DPPASS framework. Sec. 2 provides the implementation details of the proposed DPPASS method. Sec. 3 presents additional quantitative and qualitative experimental results and more details in ablation study.

1. Algorithm

The algorithm of the proposed method DPPASS is shown in Algorithm 1. Given the target domain data consisting a set of n unlabeled ERP $P = E_p^1, ..., E_p^n$ and a set of annotated pinhole images in the source domain are transformed to the m pseudo ERP $S = (I_s^1, Y_s^1), ..., (I_s^m, Y_s^m)$. The tangent images $T = E_t^1, ..., E_t^{18n}$ are projected by function $f_{E2T}(\cdot)$ from the ERP image set P, and the pseudo tangent image set $T^* = I_{t^*}^1, ..., I_{t^*}^{18m}$ are projected by the overlap patch merging from the pseudo ERP set S. The $f_{E2T}(\cdot)$ means the transformation from ERP to tangent images.

2. Implementation details

Our framework is implemented with Pytorch and trained on multiple NVIDIA GPUs. Both models in our framework are based on the efficient Segformer [1]. **Ours-T** and **Ours-S** are two implementations of our framework, which are based on **SeformerB1** and **SegormerB2**, respectively. We initialize the encoder with ImageNet-1K pre-trained weight and randomly initialize the segmentation head (decoder).

*Corresponding author.

Algorithm 1 The DPPASS framework

Input: ERP images $P = E_p^1, ..., E_p^n$;
Pinhole images $S = (I_s^1, Y_s^1), ..., (I_s^m, Y_s^m)$;
Max iterations: T
model: $f(Input, \theta_e)$, $f(Input, \theta_t)$, $f(Input, \theta_{d1})$, $f(Input, \theta_{d2})$;
Initialization: Set θ^e and θ^t with pre-trained MiT;
for t $\leftarrow$ 1 to T do **do**
  Attain the segmentation prediction maps and feature representations for both models:
  $P_e^p, F_e^p = f(E_p^i, \theta_e), P_e^s, F_e^s = f(I_s^i, \theta_e)$,
  $P_t^t, F_t^t = f(E_t^i, \theta_t), P_t^s, F_t^s = f(I_{t_*}^i, \theta_t)$;
  Forward process of the discriminators:
  $D_u = f(F_e^p, \theta_{d1}), D_s = f(F_e^s, \theta_{d1})$ (ERP Path),
  $D_u^t = f(F_t^t, \theta_{d2}), D_s^t = f(F_t^s, \theta_{d2})$ (Tangent Path);
  Compute the IGAN loss:
  $\mathcal{L}_d$ = BCE(D_u, D_s) + BCE(D_u^t, D_s^t);
  Compute the prediction consistency training loss:
  $\mathcal{L}_pc$ = KL($f_{E2T}(P_e^p)$, P_t^t);
  Compute the TFCT loss:
  $\mathcal{L}_fc$ = InfoNCE($f_{E2T}(F_e^p)$, F_t^t);
  Compute the supervised loss:
  $\mathcal{L}_s$ = CE(P_e^s, Y) + CE(P_t^s, $f_{E2T}(Y)$);
  $\mathcal{L}_{all} = \mathcal{L}_s up + \alpha * \mathcal{L}_d + \beta * \mathcal{L}_p c + \mathcal{L}_f c$;
  Back propagation for $\mathcal{L}_{all}$;
  Update parameter set θ_e, θ_t. θ_{d1}, θ_{d2};
end for
return θ^e
End.

For both models in the framework, we use a batch size of 4 and the models are trained with AdamW optimizer for 50k iterations. We set the initialized learning rate as 0.00006

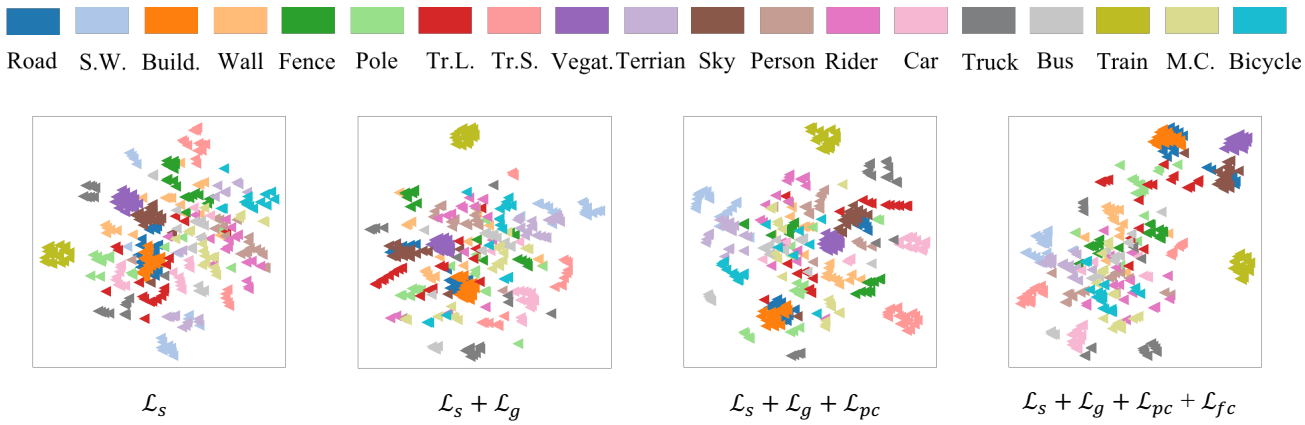

Figure 1. TSNE visualization of the features extracted by our DPPASS trained with different loss combinations, including the supervised loss $\mathcal{L}_s$, the intra-projection loss $\mathcal{L}_g$, the prediction consistency training loss $\mathcal{L}_{pc}$ and the TFCT loss $\mathcal{L}_{fc}$.

and use a poly learning rate schedule with a power factor of 1.0. The input panoramic image size of the ERP path model is 400 $\times$ 2048 and 224 $\times$ 224 for the tangent path model. For the source domain data, we randomly crop the images in Cityscapes into ERP (400 $\times$ 2048) size. Meanwhile, we perform overlap patch merging on the Cityscapes images to get tangent size patches, making it possible for the cross-model prediction consistency training and the tangent-wise feature contrastive training.

3. Experimental Results and Ablation Study

Additional Experimental Results Fig. 2 shows more qualitative results on DensePASS dataset. The red boxes show that our DPPASS achieves better segmentation performance than the prior SoTA Trans4PASS [2]. Specifically, our DPPASS-S achieves dramatical mIoU increment on the most challenging categories, includ- ing: Sidewalk (+6.56↑), Traffic Sign (+4.42↑), Person (+7.14↑) and Motorcycle (+9.57↑). This also demonstrated in Fig 2, these aforementioned categories are better segmented by ours framework than the existing state-of-the-art method [2].

Additional Ablation Study The qualitative results of the loss functions ablation study are shown in Fig. 1. To demonstrate the effectiveness of our proposed intra-projection and cross-projection modules, we visualize the features extracted by our DPPASS that are trained with different loss combinations. Significantly, all of our proposed loss functions make positive contributions in clarifying the class-wise features. This and the quantitative results reported in the main paper prove the effectiveness of all the proposed modules in our DPPASS.

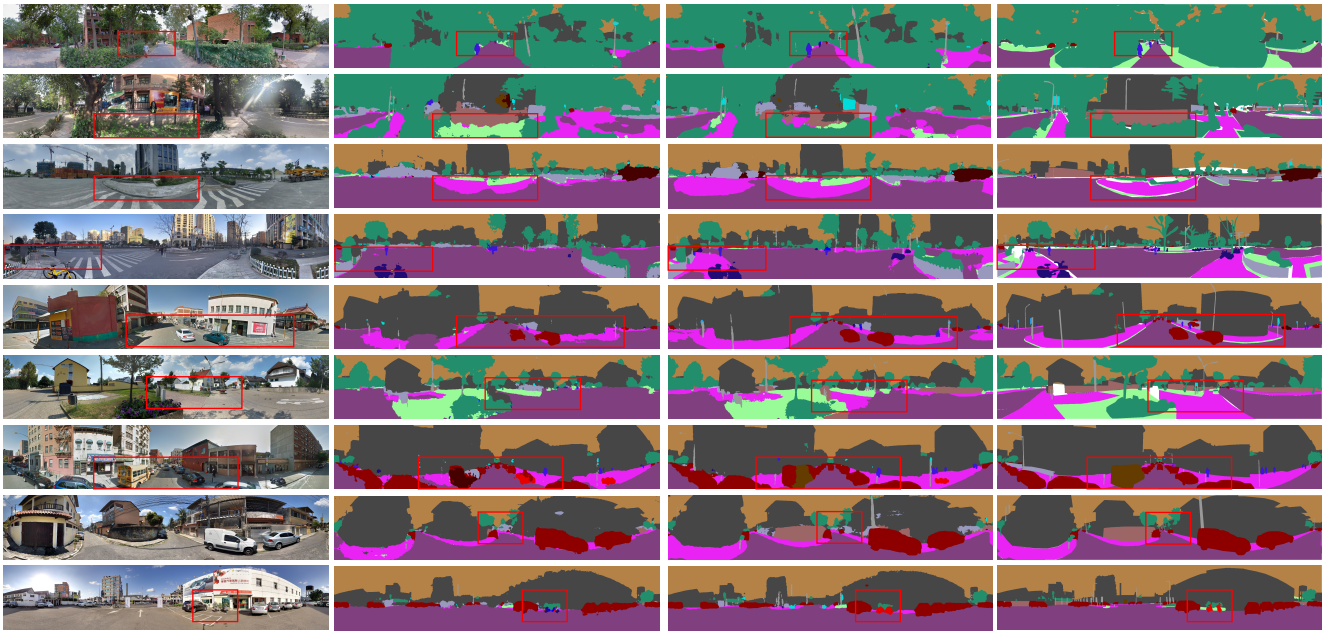

Figure 2. Example visualization results from DensePASS test set. (a) Input, (b) Trans4PASS-T [2], (c) DPPASS-T, and (d) Ground truth.

References

[1] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 12077–12090, 2021. 1

[2] Jiaming Zhang, Kailun Yang, Chaoxiang Ma, Simon Reiß, Kunyu Peng, and Rainer Stiefelhagen. Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16917–16927, 2022. 2, 3