Look at the Neighbor: Distortion-aware Unsupervised Domain Adaptation for Panoramic Semantic Segmentation – Supplementary Material –

Xu Zheng[1] Tianbo Pan[1] Yunhao Luo[3] Lin Wang[1,2*]
[1]AI Thrust, HKUST(GZ) [2]Dept. of CSE, HKUST [3]Brown University
zhengxu128@gmail.com, tpan695@connect.hkust-gz.edu.cn, devinluo27@gmail.com, 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 UDA framework. Sec. 2 provides the detailed settings of the proposed distortion-aware transformer. Sec. 3 gives the thorough derivation process of the distortion coefficient. Lastly, Sec. 4 presents additional quantitative and qualitative experimental results and more details of the ablation study.

1. Algorithm

As shown in Algorithm. 1, our framework is first trained with the source data and then performs unsupervised domain adaption with both source and target data.

2. Detailed settings of DATR

In our proposed DATR, the MLP decoder takes multi-scale features F_i from the encoder as inputs, and the channel dimensions are aligned. Then F_i are up-sampled to F_4's size and concatenated and fused together. Finally, the last MLP layer takes the fused features to predict the segmentation labels. The whole process can be formulated as:

$$\begin{aligned}
\hat{F^i} &= Linear(C_i, C)(F_i), \forall i; \\
\hat{F^i} &= Upsample(\frac{W}{4} \times \frac{W}{4})(\hat{F^i}), \forall i; \\
F &= Linear(4C, C)(Cat(Up(\hat{F^i}))), \forall i; \\
M &= Linear(C, N_{cls})(F),
\end{aligned} \tag{1}$$

where N_{cls} denotes the number of categories, Cat and Up denote the feature concatenation and the Up-sampling operation, respectively.

*Corresponding author.

Algorithm 1 Our Proposed UDA framework

Input: source input: x_s, y, target input: x_t, maximum iteration: T_s,T_t, model: $f(\Theta)$, Multi-scale pseudo labeling: PL($\cdot$);
Source Data Train:
for t $\leftarrow$ 1 to T_s **do**
  p_s, f_s = $f(x_s, \Theta)$;
  $\mathcal{L}_{SEG}$ = CE (p_s, y);
  Back propagation for $\mathcal{L}_{SEG}$;
  Update parameter set θ;
end for
Unsupervised Domain Adaptation:
for t $\leftarrow$ 1 to T_t **do**
  p_t, f_t = $f(x_t, \Theta)$, $\hat{y}^t$ = PL(p_t);
  $\mathcal{L}^t_{SS}$ = CE (p_t, $\hat{y}^t$)
  T^t = Proj (Mask (f_t, $\hat{y}^t$)), S^t = Proj(Mask(f_s, y))
  $C_s = (1 - \frac{1}{t})S^{t-1} + \frac{1}{t}S^t$, $C_t = (1 - \frac{1}{t})T^{t-1} + \frac{1}{t}T^t$;
  $\mathcal{L}_f$ = MSE (C_s, C_t)
  $\mathcal{L}_{all} = \mathcal{L}^t_{SS} + \mathcal{L}_f$;
  Back propagation for $\mathcal{L}_{all}$;
  Update parameter set θ;
end for
return θ^V
End.

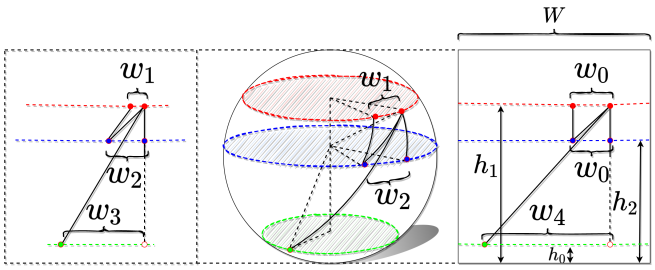
The detailed hyper-parameters of our proposed distortion-aware transformer (DATR) are given in Tab. 1. We scale up our DATR encoder by changing the hyper-parameters. The hyper-parameters are listed as follows:

- K_i: the patch size of overlapping patch merging in Stage i;
- S_i: the strides of the patch merging in Stage i;
- P_i: the padding size of the patch merging in Stage i;
- C_i: the channel number of the output of Stage i;
- L_i: the number of encoder layer in Stage i;

	Output Size	Layer Name	DATR		
			Mini	Tiny	Small
Stage1	$\frac{H}{4} \times \frac{W}{4}$	Overlapping Patch Embedding	$K_1 = 7; S_1 = 4; P_1 = 3$		
			C_1=32	C_1=64	
		Efficient Self-Attention	R_1=8 N_1=1 E_1=8 L_1=2	$R_1 = 8$ $N_1 = 1$ $E_1 = 8$ L_1=2	R_1=8 N_1=1 E_1=8 L_1=3
Stage2	$\frac{H}{8} \times \frac{W}{8}$	Overlapping Patch Embedding	K_2=3; S_2=2; P_2=1		
			C=64	C=128	
		Efficient Self-Attention	R_2=4 N_2=2 E_2=8 L=2	R_2=4 N_2=2 E_2=8 L_2=2	R_2=4 N_2=2 E_2=8 L_2=3
Stage3	$\frac{H}{16} \times \frac{W}{16}$	Overlapping Patch Embedding	K_3=3; S_3=2; P_3=1		
			C=160	C=320	
		Efficient Self-Attention	R_3=4 N_3=2 E_3=8 L_3=2	R_3=4 N_3=2 E_3=8 L_3=2	R_3=4 N_3=2 E_3=8 L=3
Stage4	$\frac{H}{32} \times \frac{W}{32}$	Overlapping Patch Embedding	K_4=3; S_4=2; P_4=1		
			C=256	C=512	
		Distortion-aware Attention	NS_4=13 L_4=2	NS_4=13 L_4=2	NS_4=13 L_4=3

Table 1: Structure settings of the proposed DATR. We follow the design principles of ResNet: the channel dimension increase when the spatial resolution shrink as the layer goes deeper.

- R_i: the reduction ratio of the Efficient Self-Attention in Stage i;
- N_i: the head number of the Efficient Self-Attention in Stage i;
- E_i: the expansion ratio of the feed-forward layer in Stage i;
- NS_i the neighboring size of our proposed Distortion-aware Attention in stage i.



(a) DLP (b) Sphere (c) ERP

Figure 1: Direct Linear Projection (DLP) (a) and ERP (c) are two projection formats of the same spherical data (b).

3. Theoretical Analysis of ERP Distortion

As shown in Fig. 1 (a), (b) and (c), the distance w_1, w_2 and w_0 between pixels in different formats are different. Intuitively, $w_0 = \frac{W}{n}$ according to the uniform distribution of pixels, where W is the length of ERP. For the red and blue pixels at different latitudes, the radius of the circumference of the circle at the corresponding latitude can be expressed

Method	Backbone	Cloudy	Foggy	Rainy	Sunny	ALL	
		val	val	val	val	val	test
PVT [3]	Tiny	39.92	34.99	34.01	39.84	36.83	32.37
	Small	40.75	36.14	34.29	40.14	37.47	32.68
Segformer [4]	MiT-B1	45.34	41.43	40.33	44.36	42.68	37.36
	MiT-B2	46.07	40.99	40.10	44.35	42.49	37.24
Trans4PASS [5]	Trans4PASS-T	46.90	41.97	41.61	45.52	42.49	37.24
	Trans4PASS-S	46.74	43.49	43.39	45.94	44.80	38.57
Trans4PASS+ [6]	Trans4PASS+-T	48.33	43.41	43.11	46.99	45.21	38.85
	Trans4PASS+-S	48.87	44.80	45.24	47.62	46.47	39.16
DATR	DATR-M	59.34	53.44	51.26	49.31	48.97	**37.10**
	DATR-T	50.94	50.90	52.69	50.63	50.67	**39.38**
	DATR-S	51.95	53.92	51.57	51.55	51.93	**41.55**

Table 2: Performance of SOTA transformer-based panoramic semantic segmentation models. The SynPASS dataset is evaluated on full 22 classes and is divided into four conditions according to weather and two conditions according to light.

as:

$$r_{red} = \sqrt{r^2 - (h_1 - r)^2}; r_{blue} = \sqrt{r^2 - (h_2 - r)^2}, \quad (2)$$

where h_1 and h_2 are the latitudes of pixels and r is the radius of the sphere. Consequently, w_1 and w_2 can be formulated as:

$$w_1 = \frac{2\pi r_{red}}{n} = \frac{2\pi\sqrt{2h_1 r - h_1^2}}{n} = \frac{2\pi}{n}\sqrt{h_1(2r - h_1)}, \quad (3)$$

$$w_2 = \frac{2\pi r_{blue}}{n} = \frac{2\pi\sqrt{2h_2 r - h_2^2}}{n} = \frac{2\pi}{n}\sqrt{h_2(2r - h_2)}, \quad (4)$$

where n is the number of sampling pixels at each latitude. Finally, substituting h_1 and h_2 into the equation:

$$w_1 = \frac{2\pi}{n}\sqrt{h_1(\frac{W}{\pi} - h_1)}, \quad (5)$$

$$w_2 = \frac{2\pi}{n}\sqrt{h_2(\frac{W}{\pi} - h_2)}, \quad (6)$$

where W is the length of ERP. Suppose there are n' pixels in both w_3 and w_4, the w_4 in Fig. 1 (c) can be formulated by:

$$w_4 = \frac{n'}{n} \times W, \quad (7)$$

where n is the total pixel number in a latitude. Meanwhile, the w_3 in Fig. 1 (a) is similar with w_1 and w_2:

$$w_3 = \frac{2\pi n'}{n}\sqrt{h_0(\frac{W}{\pi} - h_0)} \quad (8)$$

Then, the distortion coefficient Dis is defined as the difference between w_4 and w_3:

$$Dis = w_4 - w_3 = \frac{n'}{n}(W - 2\pi\sqrt{h_0(\frac{W}{\pi} - h_o)}). \quad (9)$$

Obviously, Eq. 9 shows that Dis is an increasing function affected by n', indicating the smaller n', the simpler the distortion.

4. Experimental Results and Ablation

4.1. Experimental Results

We further provide more instance feature visualization in Fig. 2 to demonstrate the superiority of our proposed distortion-aware attention (DA) module.

To evaluate the performance of panoramic semantic segmentation of existing methods and our proposed DATR on the synthetic dataset, the SynPASS [6] benchmark with all 22 categories is established. As shown in Tab. 2, results of the Transformer-based methods [3, 4, 5, 6] on the SynPASS [6] dataset are reported. All the models are trained on the training set and their performance are reported in different weather and day/night conditions. Compared with the existing state-of-the-art semantic segmentation methods, our DATR-T surpasses Segformer-B1 [4] by +6.29% in mIoU on the validation set. **The largest improvement lies on the *Foggy* condition with a +12.01% gain**. All of the variants of our proposed DATR consistently outperform PVT [3] and Segformer [4] in all conditions, which showcases that our DATR achieves dramatical capability to capture the pixel-wise neighboring correlations on the synthetic

Method	mIoU	Road	S.W.	Build.	Wall	Fence	Pole	Tr.L.	Tr.S.	Veget.	Terr.	Sky	Persin	Rider	Car	Truck	Bus	Train	M.C.	B.C.
DAFormer [2]	54.67	73.75	27.34	86.35	35.88	45.56	36.28	25.53	10.65	79.87	41.64	94.74	49.69	25.15	77.70	63.06	65.61	86.68	65.12	48.13
Trans4PASS-T [5]	53.18	78.13	41.19	85.93	29.88	37.02	32.54	21.59	18.94	78.67	45.20	93.88	48.54	16.91	79.58	65.33	55.76	84.63	59.05	37.61
Trans4PASS-S [5]	55.22	78.38	41.58	86.48	31.54	45.54	33.92	22.96	18.27	79.40	41.07	93.82	48.85	23.36	81.02	67.31	69.53	86.13	60.85	39.09
DATR-M (SS)	45.71	75.26	45.22	83.06	24.51	29.64	29.06	16.71	11.40	77.46	25.89	92.37	40.64	10.68	74.62	42.57	49.42	75.85	41.18	23.02
DATR-T (SS)	48.27	77.04	45.93	84.53	30.87	32.41	31.12	19.73	14.63	76.79	30.11	91.61	44.91	17.98	76.93	58.11	50.81	51.18	54.23	28.29
DATR-S (SS)	54.96	**80.92**	**52.69**	87.39	43.77	44.02	36.47	25.30	18.61	79.32	33.45	93.73	52.53	26.39	80.32	63.52	48.70	66.01	69.56	41.51
DATR-M (SS + CFA)	52.90	78.71	48.43	86.92	34.92	43.90	33.43	22.39	17.15	78.55	28.38	93.72	52.08	13.24	77.92	56.73	59.53	**93.98**	51.12	34.06
DATR-T (SS + CFA)	54.60	79.43	49.70	87.39	37.91	44.85	35.06	25.16	19.33	78.73	25.75	93.60	53.52	20.20	78.07	60.43	55.82	91.11	67.03	34.32
DATR-S (SS + CFA)	**56.81**	80.63	51.77	**87.80**	**44.94**	43.73	**37.23**	**25.66**	**21.00**	78.61	26.68	93.77	**54.62**	**29.50**	80.03	**67.35**	63.75	87.67	**67.57**	37.10

Table 3: Per-class results of the SOTA panoramic image semantic segmentation methods on DensePASS test set. (SS: self-supervised (SS) training)

Resolutions:	400 × 512	400 × 1024	400 × 2048	200 × 2048	100 × 2048
PVT-S [3]	26.07 (-12.67)	34.28 (-4.46)	38.74	37.50 (-1.24)	28.91 (-9.83)
Trans4PASS+-S [6]	36.67 (-8.62)	42.36 (-2.93)	45.29	44.29 (-1.00)	39.69 (-5.60)
DATR-T	51.68 (-1.55)	52.45 (-0.787)	53.23	50.46 (-2.77)	47.58 (-5.65)

Table 4: Input resolution size vs. performance.

dataset even considering different weather and day/night scenarios.

Comparing our DART and the state-of-the-art UDA method for panoramic semantic segmentation, **DATR-S performs more accurately in all conditions and clearly elevates the overall mIoU scores on both testing and validation set**. As shown in Tab. 2, our DATR-S consistently outperforms the Trans4PASS [6] by +3.08%, 9.12%, 6.33%, 3.93%, 5.46% and 2.39% mIoU increments in *Cloudy, Foggy, Rainy, Sunny, and validation and testing scenarios*, respectively.

As shown in Tab. 3, we provide the per-class mIoU results of our DATR with different combinations of UDA modules. Also, additional qualitative results on DensePASS dataset with 19 and 13 categories are shown in Fig. 4 and Fig. 5. Fig. 4 and Tab. 3 present more qualitative and quantitative results, and our framework significantly achieves better segmentation performance. Meanwhile, as discussed in the main paper, all of the variants of our method (DATR) also consistently outperform previous SOTA methods in the SynPASS-to-DensePASS scenario. Specifically, **our DATR-S achieves dramatical mIoU increment on the most challenging categories, including *Fence* (+23.04↑), *Pole* (+11.75↑), *Tr.Light* (+18.29↑), *Tr.Sign* (+16.75↑), *Person* (+8.52↑), and *Car* (+6.75↑)**. This is also demonstrated in Fig. 5, these aforementioned categories are better segmented by our framework than the existing state-of-the-art method [5].

4.2. Ablation and Discussion

Difference between DA and Deformable MLP In [6] DMLP is introduced as a decoder structure that combines feature patches at multiple scales, the motivation is to address the distortion by mixing patches across the channel dimension, resulting in a large receptive field. By contrast, our proposed DA module is an attention mechanism incorporated into the feature extractor (encoder). DA aims to reduce the receptive field of the feature extractor, enabling better distortion awareness. Hence our proposed DA module has significant difference with the DMLP. Compared with the Deformable Patch Embedding (DPE) that operates on the overall scene in ERP images, our RPE only operates in the neighboring region and provides local neighboring positional information to alleviate the distortion problem, which is more targeted, light-weight and efficient.

We also conduct experiments in Syn-to-DP, using Trans4PASS-S (DMLP) with our proposed CFA achieves 48.33 mIoU, while our DATR-T with CFA reaches 52.11 mIoU in Table. 3 of the main paper. Meanwhile, Table. 3 also provides the comparison between Trans4PASS+-S and our DATR. These results confirm that DATR is a better backbone model for UDA in panoramic segmentation.

Difference between MPA and CFA. Mutual prototypical adaptation (MPA) aims to align the feature embeddings with the prototypes obtained in the source and target domains, inspired by the knowledge distillation loss in [1]. By contrast, our proposed class-wise feature aggregation (CFA) module focuses on aligning the class centers (prototypes) of both domains, which is computationally cheaper and achieves better results.

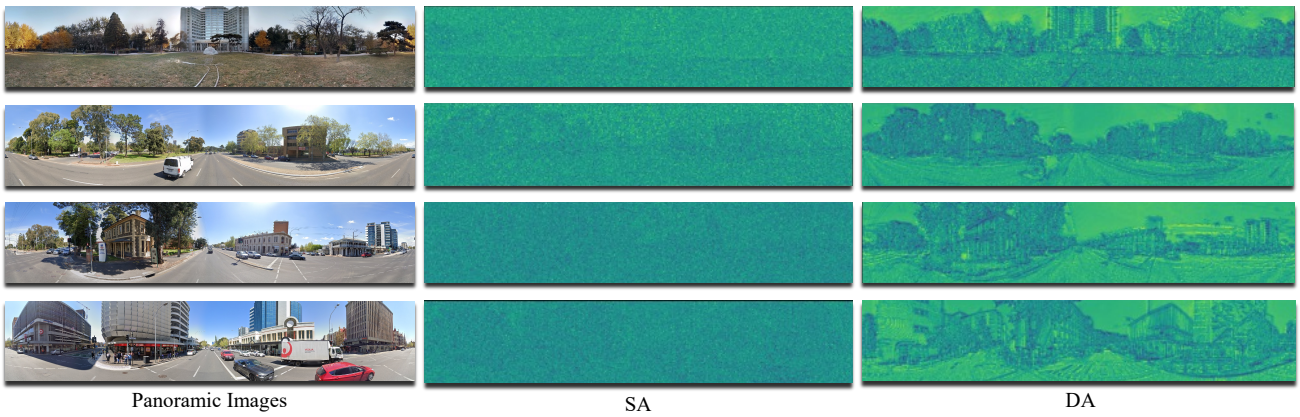


Figure 2: Visualization of the extracted features from panoramic images by Self-Attention (SA) and Distortion-aware Attention (DA).

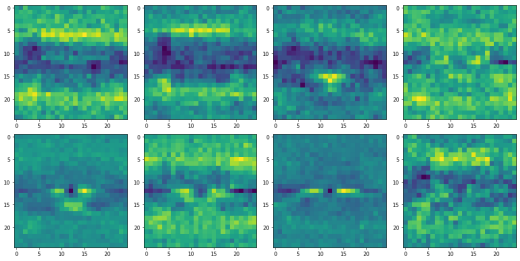
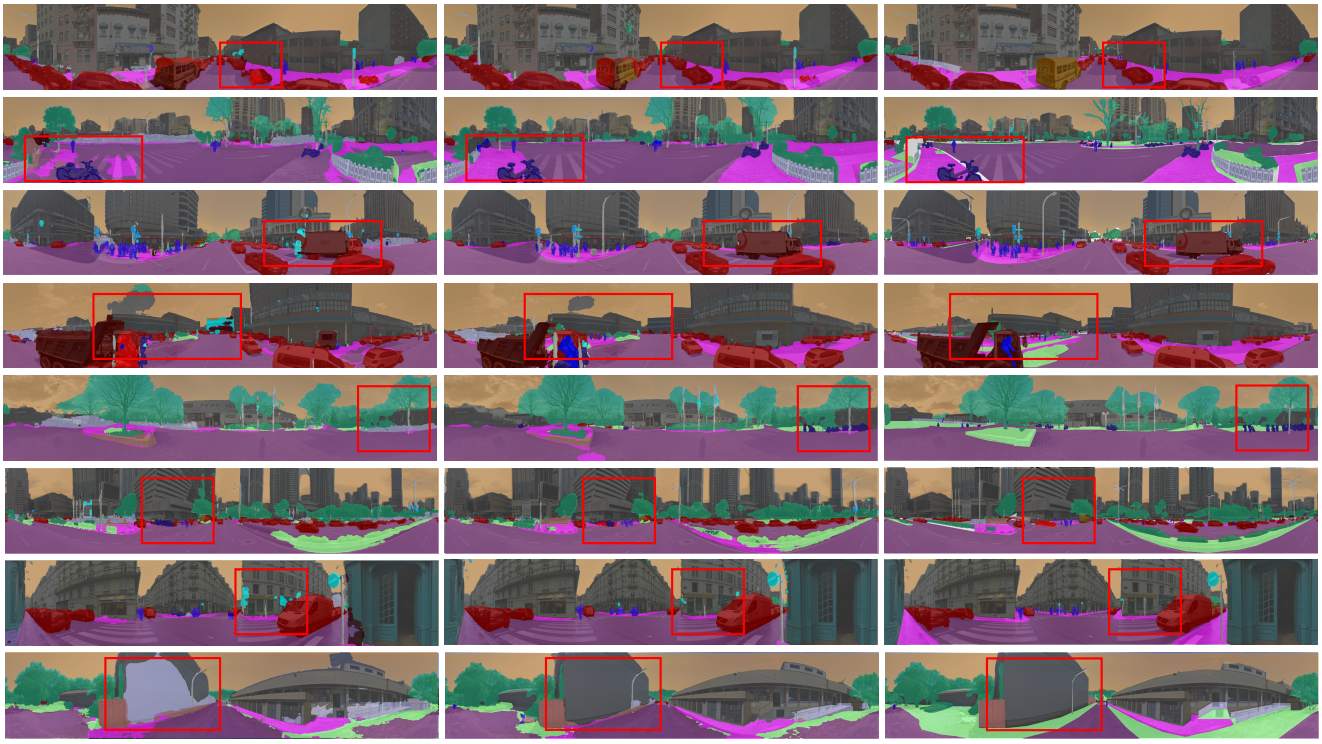
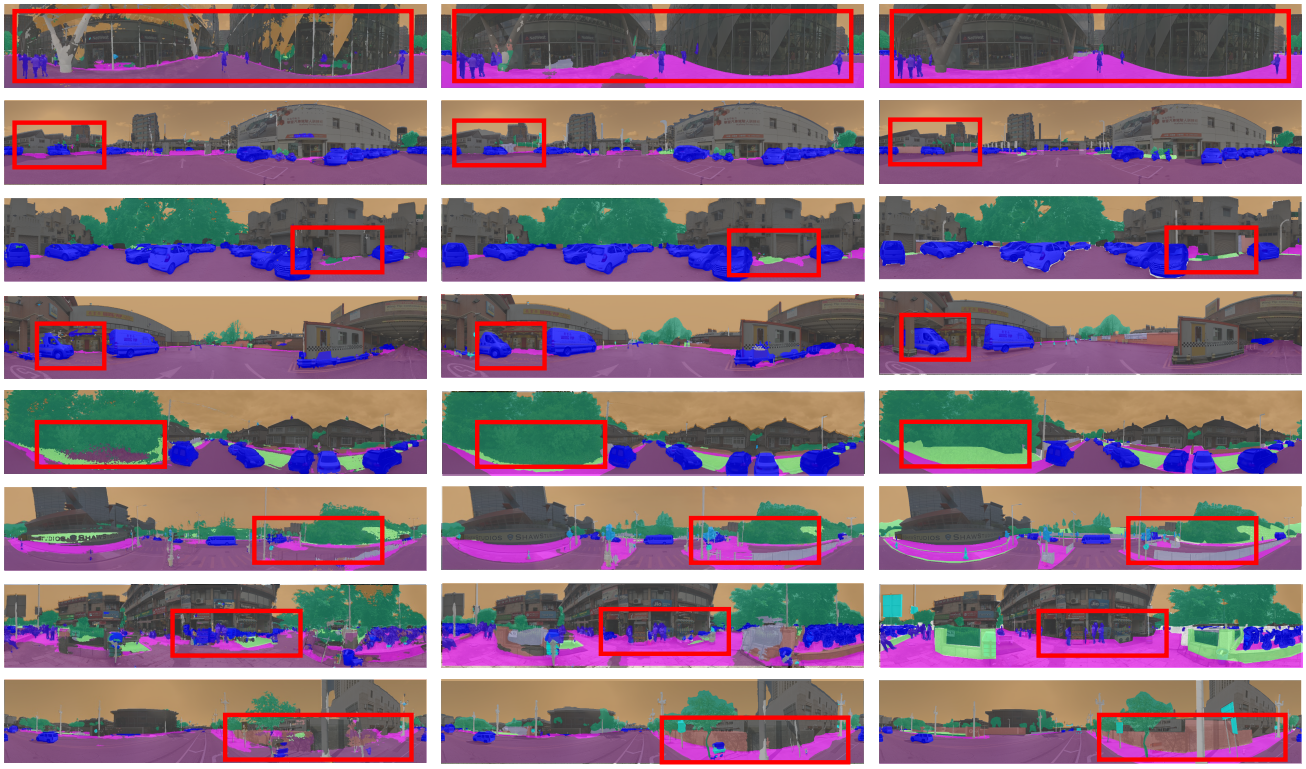


Figure 3: Visualization examples of our proposed RPE.



(a) Trans4PASS (b) Ours (c) Ground Truth

Figure 4: Qualitative results on Cityscapes-to-DensePASS (Pinhole-to-Panoramic) dataset. (a)Trans4PASS [5], (b)ours with DATR-S and (c) Ground Truth.



(a) Trans4PASS (b) Ours (c) Ground Truth

Figure 5: Qualitative results on SynPASS-to-DensePASS (Synthetic-to-Real) dataset. (a)Trans4PASS [5], (b)ours with DATR-S and (c) Ground Truth.

References

[1] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020.

[2] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 9914–9925. IEEE, 2022.

[3] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021.

[4] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

[5] 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.

[6] Jiaming Zhang, Kailun Yang, Hao Shi, Simon Reiß, Kunyu Peng, Chaoxiang Ma, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Behind every domain there is a shift: Adapting distortion-aware vision transformers for panoramic semantic segmentation. *arXiv preprint arXiv:2207.11860*, 2022.