Bending Reality: Distortion-aware Transformers for Adapting to Panoramic Semantic Segmentation (Supplementary Materials)

1. Quantitative analysis

1.1. Analysis of hyper-parameters

As the spatial correspondence problem indicated in [3], if the deformable convolution is applied to the lower or middle layers, the spatial structures are susceptible to fluctuation [4]. To overcome this problem, we propose the regional restriction of learned offsets to stabilize the training of our early-stage and four-stage Deformable Patch Embedding (DPE) module. Table 1 shows that r=4 has a better result. Thus, the constraint r applied in the offset prediction module is set as 4 in our experiments.

To investigate the effect of various hyper-parameters in the proposed Trans4PASS framework, we analyze the weight α and the temperature $\mathcal{T}$ as shown in Fig. 1a and Fig. 1b. The weight α is used to combine the *Mutual Prototypical Adaptation (MPA)* loss and the source- and target segmentation losses. As α decreases from 0.1 to 0, we set the temperature $\mathcal{T}$=35 in the MPA loss and evaluate the mIoU(%) results on the target (DensePASS [9]) dataset. If α=0, the final loss is equivalent to that of the SSL-based method, *i.e.*, the MPA loss is excluded. When α=0.001 for combining both, MPA and SSL, Trans4PASS obtains a better performance.

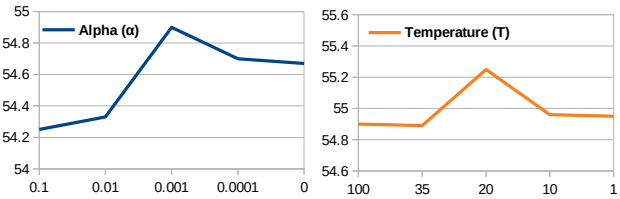
Apart from the combination weight α, we further investigate the effect of the temperature $\mathcal{T}$, which is used in the MPA loss. As shown in Fig. 1b, the performance is not sensitive to the distillation temperature, which illustrates the robustness of our MPA method. Nevertheless, we found that MPA performs better when the temperature is lower, so $\mathcal{T}$=20 is set as the default setting in our experiments.

1.2. Computational complexity

We reported the complexity of Deformable Patch Embedding (DPE) and Deformable MLP (DMLP) and compared with other methods on DensePASS in Table 2. The results indicate that our methods have significant improvement with the same order of complexity.

	None	r=1	r=2	r=4	r=8
mIoU(%)	45.74	44.51	45.59	**45.89**	45.57

Table 1. **Effect of regional restriction** (r) on DensePASS.



(a) mIoU(%) – α (b) mIoU(%) – $\mathcal{T}$

Figure 1: **Analysis of hyper-parameters**. The performance (mIoU) is evaluated in the outdoor target dataset (DensePASS).

	PE[79]	DPT[8]	DPE	CycleMLP[6]	ASMLP[43]	DMLP
GFLOPs	0.16	0.36	7.65	1.25	4.83	3.45
#Params(M)	0.01	0.02	2.90	0.45	1.04	0.79
mIoU(%)	45.14	**45.89**	36.50	40.16	42.05	**45.89**

Table 2. **Computational complexity** of DPE and DMLP. GFLOPs are calculated @512×512.

1.3. Detailed results in outdoor scenarios

Table 3 shows the per-class IoU results on DensePASS dataset. The first group of experiments is conducted to compare the performance of different backbones in P2PDA [14] method. Additionally, the adaptation process of the original FANet [7] and DANet [6] are shown in more detail, *i.e.*, the performance of the source-only model and that without using the SSL-based method are included. The experiments in the second and third groups are based on Trans4PASS-T and -S model, respectively. As shown in the third group, Trans4PASS-S obtains new state-off-the-art performance in mean IoU (56.38%). In addition, it achieves top scores on 7 out of 19 classes in per-class IoU, including *pole*, *traffic light*, *person*, *car*, *truck*, *motorcycle*, and *bicycle*.

1.4. Detailed results in indoor scenarios

Apart from the detailed results in outdoor scenarios, per-class results on the outdoor Stanford2D3D-Panoramic dataset [1] are shown in Table 4. The experiments are conducted on the fold-1 dataset setting of Stanford2D3D [1]. Our proposed framework with the Trans4PASS-S backbone

and the MPA method obtains the best performance in the domain adaptation setting, reaching 52.15% in mean IoU. It also achieves best IoU scores on 7 out of 13 classes in the indoor scenario, especially on the *ceiling*, *column*, and *door* categories. In the supervised learning setting, Trans4PASS-S surpasses the CNN-based DANet by a large margin, achieving a score of 53.31% in mean IoU. Besides, its performance in per-class IoU is better than DANet in almost all categories, which lacks the capacity to learn long-range contexts and distortion-aware features in panoramas.

The comparison of segmentation performance with state-of-the-art methods on Stanford2D3D-Panoramic dataset is shown in Table 5. Since the results of these experiments are based on the average of all 3 data-splitting settings, we show the results of each individual split setting and its per-class IoU in detail (in gray). The small version of the Trans4PASS backbone is used in this experiment. Compared with the previous best fully-supervised method equipped with ResNet-101, Trans4PASS-S has much fewer parameters and is an order of magnitude smaller than ResNet-101. Still, our method obtains the new state-of-the-art performance on Stanford2D3D-Panoramic dataset, reaching 53.0% in mean IoU. Within all 13 classes, Trans4PASS obtains a total of 8 best per-class IoUs. In the setting of unsupervised domain adaptation (UDA), our proposed method achieves $+2.7\%$ in the average of three folds, and $+3.1\%$ when using multi-scale evaluation. It obtains best per-class scores on 9 out of 13 categories.

2. Qualitative analysis

2.1. More visualizations in indoor scenarios

Similar to the visualization in outdoor scenarios, more qualitative comparisons between the baseline and the proposed Trans4PASS are displayed in Fig. 2, which are from the evaluation set of Stanford2D3D-Panoramic [1] in the fold-1 setting. In Fig. 2(a), Trans4PASS can produce higher quality segmentation results in those categories highlighted by the black dashed rectangles, such as *column* and *bookcase* categories, while the baseline model can hardly identify these severely deformed objects. In Fig. 2(b), the *doors* are incorrectly segmented as part of the *wall* by the baseline model, and the correct segmentation results can be generated by our Trans4PASS model.

2.2. More visualizations in outdoor scenarios

To fully demonstrate the effect of Trans4PASS in dealing with image distortions and object deformations, more qualitative comparisons between the baseline and the proposed Trans4PASS are displayed in Fig. 3, which are generated from the evaluation set of DensePASS dataset [9]. Specifically, Trans4PASS can better classify and segment deformed foreground objects with accurate boundaries, such as the segmentation results of *cars* and *trucks* highlighted by the blue dashed rectangles in Fig. 3(a), while the baseline model without deformable PE and deformable MLP modules is likely to be confused or fail in these categories. Apart from the foreground object, the ultra-wide arranged background is particularly distorted and challenging. Thanks to the two distortion-aware modules, our Trans4PASS yields high-quality segmentation results in these categories, *e.g.*, *terrain*, *sidewalk*, and *wall* in Fig. 3(b).

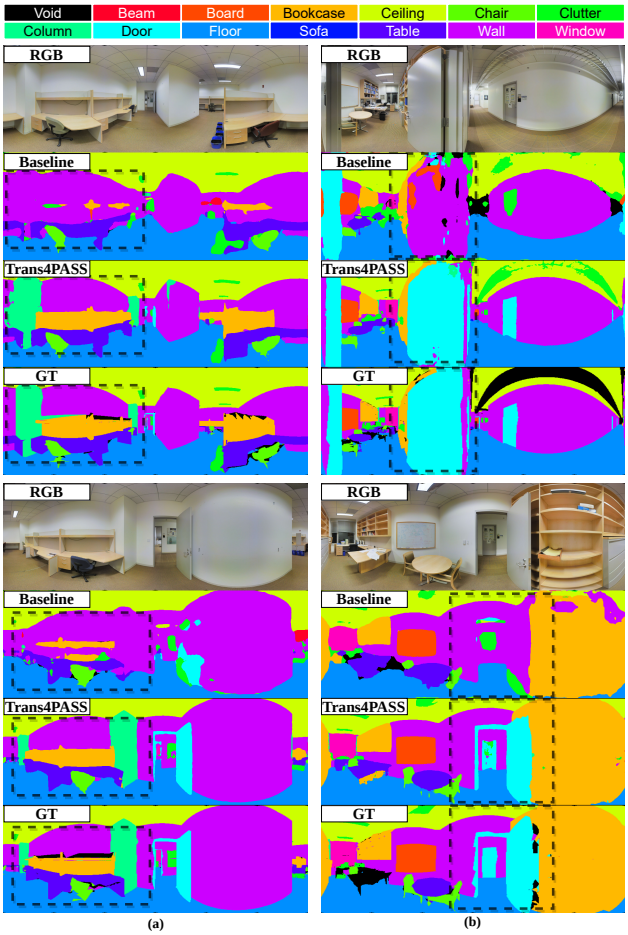


Figure 2: **Qualitative comparisons** in indoor scenarios.

3. Broader Impact.

This work promotes panoramic semantic segmentation of indoor and outdoor scenes, which benefits ultra-wide scene understanding. However, the proposed method has not been verified in practical applications such as those in intelligent vehicles and mobility assistive systems. As the experiments are conducted based on the referred datasets, there are still data biases in different test fields. If the learned model is directly applied to real scenarios, it may cause negative social impacts such as less reliable decision with less accurate segmentation, which should be considered in the downstream applications.

Network	Method	mIoU	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle
FANet	-	26.90	62.98	10.64	72.41	7.80	20.74	11.77	6.85	3.75	68.11	21.56	87.00	23.73	5.33	49.61	10.65	0.54	16.76	24.15	6.62
FANet	P2PDA	33.52	57.16	25.66	78.43	16.02	26.88	12.76	2.30	7.34	68.73	26.92	87.45	36.51	1.20	62.83	20.16	0.00	68.46	17.86	20.19
FANet	P2PDA + SSL	35.67	58.08	28.75	78.19	16.47	26.86	13.78	4.76	7.62	69.01	34.58	87.51	36.12	0.90	64.06	27.50	0.00	84.99	18.13	20.35
DANet	-	28.50	70.68	8.30	75.80	9.49	21.64	15.91	5.85	9.26	71.08	31.50	85.13	6.55	1.68	55.48	24.91	30.22	0.52	0.53	17.00
DANet	P2PDA	40.52	62.90	25.58	76.62	24.45	30.37	14.45	16.75	9.96	67.87	19.70	82.04	34.18	22.95	56.99	54.27	44.15	47.75	46.98	31.86
DANet	P2PDA + SSL	41.99	70.21	30.24	78.44	26.72	28.44	14.02	11.67	5.79	68.54	38.20	85.97	28.14	0.00	70.36	60.49	38.90	77.80	39.85	24.02
Trans4PASS-T	-	45.89	72.42	32.53	84.43	20.13	35.20	24.45	15.37	12.59	78.85	31.65	90.87	42.42	14.12	74.07	39.66	35.45	90.32	50.31	26.95
Trans4PASS-T	P2PDA	51.05	74.82	36.53	85.93	30.23	34.83	33.70	20.36	20.40	77.43	34.87	93.65	46.01	20.89	76.85	58.19	51.20	82.19	56.84	35.09
Trans4PASS-S	-	48.73	70.28	25.52	84.98	29.10	39.00	29.05	17.77	13.21	78.26	29.89	91.00	42.16	13.43	78.26	47.25	63.82	78.06	60.31	34.38
Trans4PASS-S	P2PDA	52.91	76.29	41.02	86.86	**31.96**	42.15	35.15	20.98	19.49	79.44	29.26	93.64	49.62	17.47	78.77	62.80	66.38	77.98	59.23	36.73
Trans4PASS-T	-	45.89	72.42	32.53	84.43	20.13	35.20	24.45	15.37	12.59	78.85	31.65	90.87	42.42	14.12	74.07	39.66	35.45	**90.32**	50.31	26.95
Trans4PASS-T	Warm-up	50.56	76.54	38.94	84.99	27.1	33.61	30.75	18.75	16.73	79.15	41.43	92.19	43.1	18.49	78.42	59.0	51.09	79.9	58.88	31.54
Trans4PASS-T	SSL	51.86	78.24	41.16	85.82	27.86	36.01	30.92	21.26	17.70	79.11	**46.44**	93.47	44.72	17.66	79.44	63.69	48.14	81.56	59.09	32.96
Trans4PASS-T	MPA	51.93	77.27	45.61	85.66	23.57	37.10	31.22	20.13	15.35	**79.91**	43.81	93.95	46.37	21.63	79.34	62.09	56.05	78.43	56.31	32.89
Trans4PASS-T	MPA + SSL	53.26	78.14	41.24	85.99	30.21	37.28	32.60	21.71	19.05	79.05	45.70	93.87	48.71	18.15	79.63	64.69	54.71	84.57	59.26	37.31
Trans4PASS-T	MPA + SSL + MS	**54.72**	78.42	42.26	85.88	30.97	38.10	33.83	21.57	**20.92**	78.26	44.90	93.57	48.43	22.53	79.90	66.00	66.32	85.10	60.54	42.09
Trans4PASS-S	-	48.73	70.28	25.52	84.98	29.10	39.00	29.05	17.77	13.21	78.26	29.89	91.00	42.16	13.43	78.26	47.25	63.82	78.06	60.31	34.38
Trans4PASS-S	Warm-up	52.59	75.28	37.08	86.21	31.34	38.84	34.6	20.92	17.13	79.18	34.86	93.81	49.15	24.12	80.01	55.38	62.2	77.8	61.14	40.2
Trans4PASS-S	SSL	54.67	79.72	44.34	85.28	28.88	43.46	34.08	22.63	17.21	78.93	43.98	92.84	49.58	**26.28**	81.04	65.92	67.37	76.96	59.90	40.25
Trans4PASS-S	MPA	54.77	**80.55**	**51.12**	**87.12**	25.87	**45.55**	34.64	23.44	14.45	79.60	31.77	**93.98**	49.55	22.98	78.97	66.73	66.28	88.65	61.09	38.25
Trans4PASS-S	MPA + SSL	55.25	78.39	41.62	86.47	31.56	45.47	34.02	22.98	18.33	79.63	41.35	93.80	49.02	22.99	81.05	67.43	**69.64**	86.04	60.85	39.20
Trans4PASS-S	MPA + SSL + MS	**56.38**	79.91	42.68	86.26	30.68	42.32	**36.61**	**24.81**	19.64	78.80	44.73	93.84	**50.71**	24.39	**81.72**	**68.86**	66.18	88.62	**63.87**	**46.62**

Table 3. **Per-class results on DensePASS dataset.** 'SSL' represents the self-supervised learning with pseudo-labels. '-' means no adaptation. 'MS' denotes multi-scale evaluation.

Network	Method	mIoU	beam	board	bookcase	ceiling	chair	clutter	column	door	floor	sofa	table	wall	window
DANet	-	40.28	0.00	56.07	52.09	72.05	35.72	20.54	5.81	19.43	72.84	31.76	41.80	68.43	47.13
DANet	P2PDA	42.26	0.22	57.49	50.92	73.09	44.63	21.72	9.09	24.02	83.18	30.94	41.36	65.43	47.24
PVT-Tiny	-	24.45	0.06	28.05	32.99	58.97	13.68	12.97	3.03	2.46	76.56	0.00	28.65	51.20	9.23
PVT-Tiny	P2PDA	39.66	0.38	60.55	54.08	75.14	33.99	26.20	7.23	12.66	82.58	9.14	42.74	65.75	45.12
PVT-Small	-	23.11	0.42	29.82	26.20	58.65	5.89	12.62	3.57	1.80	77.11	0.00	28.49	48.24	7.58
PVT-Small	P2PDA	43.10	0.00	66.24	55.31	76.92	40.95	28.99	5.60	13.62	88.35	14.53	52.08	68.26	49.50
Trans4PASS-T	-	46.08	0.28	65.21	60.07	76.36	50.30	33.09	11.89	20.72	86.87	26.14	50.84	68.64	48.56
Trans4PASS-T	MPA	47.48	0.16	66.8	60.54	76.06	52.50	31.50	14.55	20.73	86.53	36.09	52.10	69.73	50.01
Trans4PASS-S	-	48.34	**2.41**	**70.15**	60.22	77.97	**62.10**	**35.37**	13.68	16.15	89.44	31.78	**62.03**	67.63	**54.40**
Trans4PASS-S	MPA	**52.15**	1.03	68.02	**61.38**	**82.23**	58.74	35.18	**17.39**	**36.36**	**90.26**	**46.15**	56.79	**73.46**	50.91
DANet	supervised	44.15	0.27	55.13	53.40	73.92	54.03	**34.60**	5.27	12.45	90.05	30.57	50.25	66.63	47.44
Trans4PASS-S	supervised	**53.31**	**0.43**	**69.45**	**62.24**	**82.77**	**58.52**	34.26	**21.86**	**44.87**	**91.19**	**40.78**	**57.69**	**74.80**	**54.20**

Table 4. **Per-class results on Stanford2D3D-Panoramic dataset** according to the fold-1 data setting [1].

	Method	Input	mIoU	beam	board	bookcase	ceiling	chair	clutter	column	door	floor	sofa	table	wall	window
Supervised	StdConv [12]	RGB	32.6	0	46.6	44.9	60.8	32.4	18.8	0	13.0	78.0	0	32.6	54.8	40.1
	CubeMap [12]	RGB	33.8	0.2	48.3	48.5	61.3	33.4	23.4	0	15.4	72.7	0	33.8	61.7	36.9
	DistConv [12]	RGB	34.6	0.3	50.8	47.1	61.5	35.4	19.5	0	13.8	83.4	0	34.5	57.1	42.6
	UNet [10]	RGB-D	35.9	8.5	27.2	30.7	78.6	35.3	28.8	4.9	33.8	89.1	8.2	38.5	58.8	23.9
	GaugeNet [2]	RGB-D	39.4	–	–	–	–	–	–	–	–	–	–	–	–	–
	UGSCNN [8]	RGB-D	38.3	8.7	32.7	33.4	82.2	42.0	25.6	10.1	41.6	87.0	7.6	41.7	61.7	23.5
	HexRUNet [13]	RGB-D	43.3	10.9	39.7	37.2	**84.8**	50.5	29.2	11.5	45.3	**92.9**	19.1	49.1	63.8	29.4
	Tangent (ResNet-101) [5]	RGB	45.6	–	–	–	–	–	–	–	–	–	–	–	–	–
	HoHoNet (ResNet-101) [11]	RGB	52.0	–	–	–	–	–	–	–	–	–	–	–	–	–
	Trans4PASS (F-1)	RGB	53.3	0.4	69.5	62.2	82.8	58.5	34.3	21.9	44.9	91.2	40.8	57.7	74.8	54.2
	Trans4PASS (F-2)	RGB	45.7	12.5	46.9	32.6	82.3	64.7	37.5	20.1	42.7	86.6	17.7	45.2	70.3	35.1
	Trans4PASS (F-3)	RGB	57.2	21.4	65.4	58.3	80.2	55.8	41.9	28.6	76.3	88.6	45.4	58.8	59.3	63.6
	Trans4PASS (Avg)	RGB	52.1	**11.4**	60.6	51.1	81.8	59.7	37.9	23.5	**54.6**	88.8	**34.6**	53.9	68.1	51.0
	Trans4PASS (F-1, MS)	RGB	54.2	0.7	72.1	64.1	83.4	61.3	35.5	22.4	42.2	92.0	41.6	59.4	75.3	54.4
	Trans4PASS (F-2, MS)	RGB	46.4	3.1	48.2	32.1	82.9	66.4	37.8	20.3	42.7	87.2	16.8	45.9	71.3	38.0
	Trans4PASS (F-3, MS)	RGB	58.4	1.7	67.1	60.1	81.3	56.8	42.6	29.8	77.6	89.5	45.3	59.9	60.1	67.3
	Trans4PASS (Avg, MS)	RGB	**53.0**	1.8	**62.5**	**52.1**	82.6	**61.5**	**38.6**	**24.2**	54.2	89.5	34.5	**55.1**	**68.9**	**53.2**
UDA	Trans4PASS (F-1)	RGB	48.6	0.1	65.8	58.3	80.5	54.2	29.1	17.4	23.7	89.0	34.3	54.9	73.2	51.6
	Trans4PASS (F-2)	RGB	40.6	10.2	38.3	28.9	77.8	54.6	32.5	15.7	32.9	83.2	13.7	38.0	67.9	33.6
	Trans4PASS (F-3)	RGB	55.2	17.4	64.7	60.2	76.4	58.3	41.4	5.0	76.6	84.5	47.2	57.3	63.8	64.5
	Trans4PASS (Avg)	RGB	48.1	9.2	56.3	49.1	78.2	55.7	34.3	**12.7**	44.4	85.6	31.8	50.1	68.3	49.9
	Trans4PASS (F-1, MPA)	RGB	52.2	1.0	68.0	61.4	82.2	58.7	35.2	17.4	36.4	90.3	46.2	56.8	73.5	50.9
	Trans4PASS (F-2, MPA)	RGB	41.8	11.0	35.1	30.9	78.6	59.3	32.7	14.3	45.6	80.1	22.9	37.0	66.2	29.6
	Trans4PASS (F-3, MPA)	RGB	58.5	24.5	70.4	59.0	81.3	58.5	43.3	4.6	76.1	89.6	53.3	62.0	65.7	72.0
	Trans4PASS (Avg, MPA)	RGB	50.8	**12.2**	57.8	50.4	80.7	58.8	37.1	12.1	**52.7**	**86.7**	40.8	51.9	**68.4**	50.8
	Trans4PASS (F-1, MPA, MS)	RGB	52.6	0.8	70.7	63.3	82.2	60.8	36.2	16.4	33.4	90.5	45.9	58.4	73.1	51.5
	Trans4PASS (F-2, MPA, MS)	RGB	42.6	11.7	35.5	31.6	79.2	60.8	33.2	15.6	46.5	78.8	24.1	38.0	66.2	32.5
	Trans4PASS (F-3, MPA, MS)	RGB	58.3	22.6	70.6	59.4	81.5	58.8	43.9	4.2	76.7	89.5	52.8	62.0	66.0	70.7
	Trans4PASS (Avg, MPA, MS)	RGB	**51.2**	11.7	**58.9**	**51.4**	**81.0**	**60.1**	**37.7**	12.0	52.2	86.2	**40.9**	**52.8**	**68.4**	**51.6**

Table 5. **Comparison on Stanford2D3D-Panoramic dataset.** 'F-i' is the result of the fold-i (in gray) setting of Stanford2D3D [1]. 'Avg' is the averaged result of all 3 folds. 'MS' is multi-scale evaluation. 'UDA' is short for unsupervised domain adaptation.

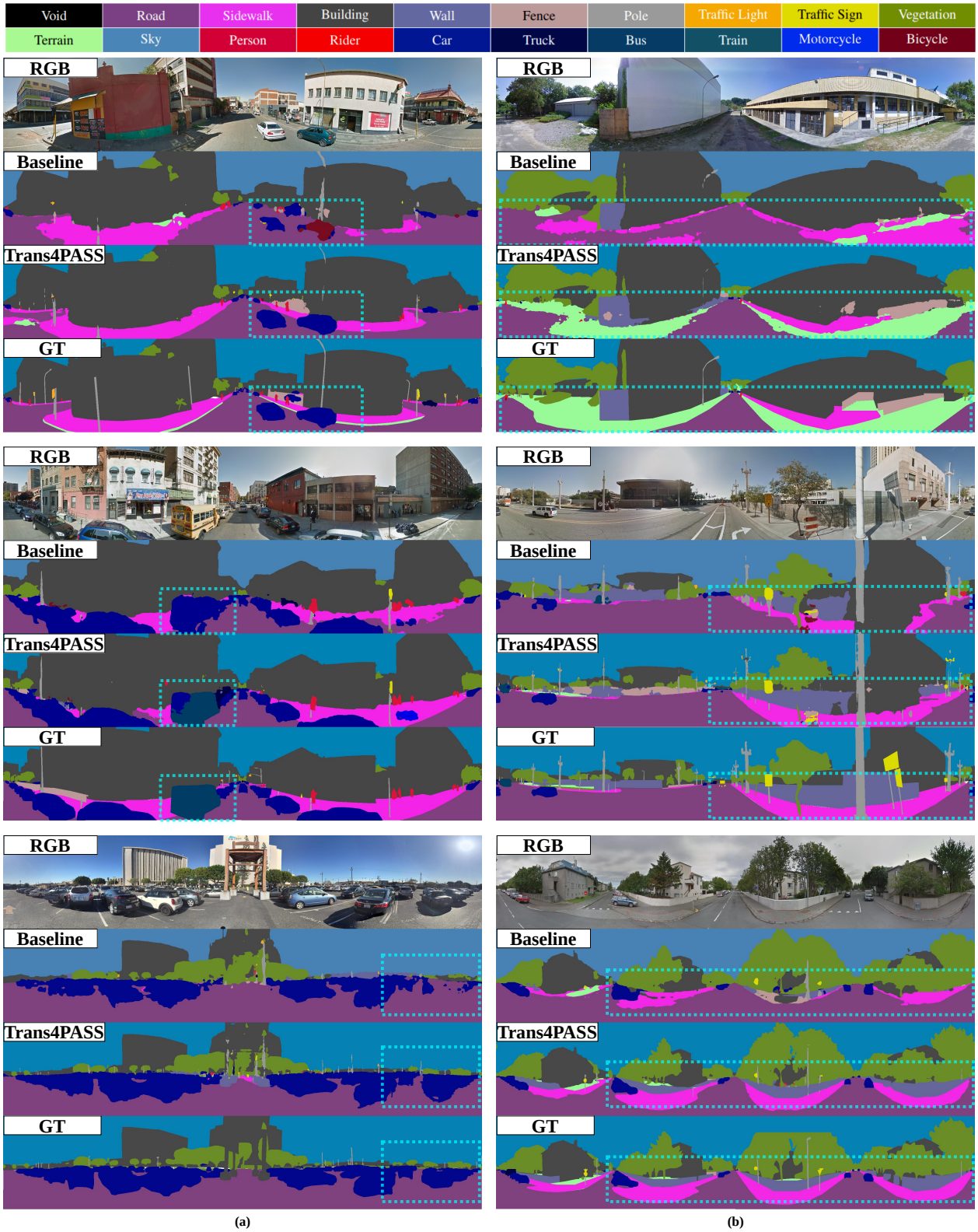


Figure 3: **Qualitative comparisons** in outdoor scenarios.

References

[1] Iro Armeni, Sasha Sax, Amir R. Zamir, and Silvio Savarese. Joint 2D-3D-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.

[2] Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral CNN. In *ICML*, 2019.

[3] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017.

[4] Liuyuan Deng, Ming Yang, Hao Li, Tianyi Li, Bing Hu, and Chunxiang Wang. Restricted deformable convolution-based road scene semantic segmentation using surround view cameras. *T-ITS*, 2020.

[5] Marc Eder, Mykhailo Shvets, John Lim, and Jan-Michael Frahm. Tangent images for mitigating spherical distortion. In *CVPR*, 2020.

[6] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019.

[7] Ping Hu, Federico Perazzi, Fabian Caba Heilbron, Oliver Wang, Zhe Lin, Kate Saenko, and Stan Sclaroff. Real-time semantic segmentation with fast attention. *RA-L*, 2021.

[8] Chiyu Max Jiang, Jingwei Huang, Karthik Kashinath, Prabhat, Philip Marcus, and Matthias Nießner. Spherical CNNs on unstructured grids. In *ICLR*, 2019.

[9] Chaoxiang Ma, Jiaming Zhang, Kailun Yang, Alina Roitberg, and Rainer Stiefelhagen. DensePASS: Dense panoramic semantic segmentation via unsupervised domain adaptation with attention-augmented context exchange. In *ITSC*, 2021.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[11] Cheng Sun, Min Sun, and Hwann-Tzong Chen. HoHoNet: 360 indoor holistic understanding with latent horizontal features. In *CVPR*, 2021.

[12] Keisuke Tateno, Nassir Navab, and Federico Tombari. Distortion-aware convolutional filters for dense prediction in panoramic images. In *ECCV*, 2018.

[13] Chao Zhang, Stephan Liwicki, William Smith, and Roberto Cipolla. Orientation-aware semantic segmentation on icosahedron spheres. In *ICCV*, 2019.

[14] Jiaming Zhang, Chaoxiang Ma, Kailun Yang, Alina Roitberg, Kunyu Peng, and Rainer Stiefelhagen. Transfer beyond the field of view: Dense panoramic semantic segmentation via unsupervised domain adaptation. *T-ITS*, 2021.