PIT: Position-Invariant Transform for Cross-FoV Domain Adaptation Supplementary Material

7. Additional Results

7.1. Class-wise Detection Results

In the paper, we follow the setting of DA-FRCN [2], SCL [9] and GPA [11] for fair comparison, which only have car results in the adaptation between Cityscapes → KITTI. In order to verify the generalization ability of our method, we conduct an experiment training with 4-classes label (4 overlapped classes in these two datasets). Results in Tab. 8 shows that PIT can work on multi-class training.

7.2. Class-wise Segmentation Results

Tab. 10 and Tab. 11 shows the IoU of each class in semantic segmentation experiments. The results demonstrate that PIT module tends to improve the performance of large objects, for the reason that their area spans a larger FoV and thus lead to a greater extent of intra-instance deformation in the original images.

7.3. Full-size FoV-decreasing Adaptation

Sec. 3.2 analyzes the different situation (*i.e.* whether the source image is sufficient) in FoV-decreasing case, and Sec. 4.3.2 gives the result of insufficient source images (a subset of source dataset). For reference, Tab. 9 shows the result of sufficient source images (the fullsize source dataset). In FoV-decreasing case, the PIT module works better when there are not enough source samples.

8. Computational Overhead

Datasets can be transformed and saved before training, and it takes little time to transform an image. For example, it takes 0.27s to process an image in Cityscapes (2048×1024 pixels), and 0.06s for one from KITTI (1242×375 pixels) with a Tesla V100 GPU.

Table 12 shows the time comparison of segmentation task Cityscapes→KITTI with and without PIT. Due to the fact the Self-Ensembling [3] is the repredentative method of the consistency regulaization [5, 6, 10, 12, 13], we use [3] as our backbone framework. It needs little additional time to train with PIT and reverse PIT modules. Using our reweighting strategy, training time for each iteration declines due to the smaller sizes of transformed images, and the performance remain similar (mIoU = 60.62% for reverse PIT and 61.00% for re-weighting). Adding the fixed time of PIT process, the average time rises little in few iterations, and even becomes less in a large number of iterations. In addition, the inference time per image in this task changes from 0.081s to 0.096s when adding our method, which only costs 10.9s extra time for the validation of 748 images.

Table 8: Multi-class detection results (%) of Cityscapes → KITTI.

Method	car	person	rider	truck	mAP
SWDA [8]	73.26	56.78	19.69	17.24	41.74
SWDA + PIT	**75.30**	**56.93**	**26.13**	**18.48**	**44.21**

Table 9: Detection results (carAP, %) of KITTI → Cityscapes and Virtual KITTI → Cityscapes.

Method	Datasets	
	K → C	VK → C
SWDA [8]	41.68	38.59
SWDA + PIT	**41.89**	**39.49**

9. Qualitative results

We visualize the qualitative results of task Cityscapes [4] $(50°, 25°)$ → KITTI [7] $(90°, 34°)$.

Fig. 7 shows the detection results using GPA [11] as the baseline. In results of the baseline (left column), the off-centered objects are likely to be recognized as several smaller objects or be detected partially due to their greater deformation extent. Our method (right column) solves these problem successfully by alleviating this kind of deformation, leading to clearer and more precise predicted bounding boxes.

Using Self-Ensembling [3] as the baseline, we get the qualitative segmentation results in Fig. 8. Our method provides more accurate predictions, especially in the off-centered pixels.

Table 10: Class-wise adaptive segmentation results (%) of Cityscapes → KITTI.

Method	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motocycle	bike	**mIoU**$_{19}$
Self-Ensembling [3]	85.36	48.30	80.50	37.98	39.95	45.64	58.95	53.08	87.86	52.70	93.40	58.09	47.01	87.21	52.20	68.08	37.41	49.24	48.28	59.54
Self-Ensembling + PIT	**88.86**	**49.00**	**81.18**	**43.04**	**40.90**	38.10	57.57	**53.46**	87.43	**57.15**	**93.62**	52.93	**47.75**	**90.17**	**60.52**	**69.80**	**65.04**	32.99	**49.41**	**61.00**
CowMix [6]	85.26	49.16	80.64	38.63	41.93	43.16	60.47	56.97	86.81	47.65	93.08	58.41	42.23	87.15	49.16	66.58	40.66	46.80	49.09	59.15
CowMix + PIT	**90.44**	48.31	**80.71**	**41.46**	38.16	38.10	57.68	53.06	**87.76**	**60.81**	**93.39**	50.27	**50.87**	**90.41**	**58.05**	**76.99**	36.95	42.48	**51.13**	**60.37**
CutMix [5]	85.61	47.56	77.89	37.71	39.71	45.35	59.85	55.50	86.91	48.31	92.77	54.71	53.46	86.70	45.06	72.74	31.23	49.05	46.73	58.78
CutMix + PIT	**90.81**	**48.40**	**80.48**	**48.99**	37.21	39.79	58.11	52.57	**87.28**	**61.70**	92.73	50.99	51.65	**89.86**	**52.67**	70.63	**37.00**	42.56	**48.20**	**60.09**
DACS [10]	84.90	46.10	80.04	34.10	37.84	43.21	56.13	55.12	88.91	58.58	93.05	57.15	45.80	86.98	46.67	76.35	39.48	47.53	46.73	59.19
DACS + PIT	**90.30**	**51.18**	78.63	**41.22**	**41.08**	41.35	**59.29**	**55.23**	86.75	**59.15**	90.78	54.14	**49.94**	**88.63**	**56.62**	66.15	**56.67**	37.10	**51.42**	**60.82**

Table 11: Class-wise adaptive segmentation results (%) of Virtual KITTI → KITTI.

Method	road	building	pole	light	sign	vegetation	terrain	sky	car	truck	**mIoU**$_{10}$
GIO-Ada*(CVPR'19) [1]	81.4	71.2	11.3	26.6	23.6	82.8	56.5	88.4	80.1	12.7	53.5
Self-Ensembling [3]	84.63	71.42	10.14	28.51	40.09	46.58	89.33	84.85	16.19	82.80	55.45
Self-Ensembling + PIT	**86.67**	**72.83**	7.14	**29.77**	**40.70**	**56.95**	**90.13**	**85.98**	**16.95**	**85.10**	**57.22**
CowMix [6]	83.89	68.96	12.58	30.30	39.02	50.77	89.05	84.02	18.00	84.14	56.07
CowMix + PIT	**87.16**	67.54	8.43	**30.33**	**42.53**	**54.23**	**91.10**	**86.69**	**20.85**	83.58	**57.24**
CutMix [5]	84.05	71.95	12.09	34.04	36.95	51.47	87.64	83.89	10.77	82.99	55.58
CutMix + PIT	**87.53**	66.08	**12.54**	30.15	**41.35**	**55.61**	**90.83**	**86.56**	**12.76**	**83.79**	**56.72**
DACS [10]	87.35	66.81	10.49	30.24	41.94	54.92	90.97	86.63	16.81	83.69	56.98
DACS + PIT	85.82	**72.21**	5.08	26.64	**42.07**	**52.87**	90.43	**86.07**	**20.50**	**84.03**	**56.57**

* the reported performance from its original paper.

Table 12: Time comparison of segmentation task Cityscapes → KITTI on a Tesla V100 GPU. RPIT refers to directly using reversed PIT in loss calculation, while re-weighting means using our proposed re-weighting strategy.

Iteration	Method	T_{PIT}(s)	T_{train} (s)	T_{total}(s)	$T_{average}$(s)
10k	Self-Ensembling [3]	0	9,454.8	9,454.8	0.95
	Self-Ensembling + PIT (RPIT)	1,207.2	9,849.1	11,056.3	1.11
	Self-Ensembling + PIT (re-weighting)	1,207.2	9,150.3	10,357.5	1.04
100k	Self-Ensembling [3]	0	94,548.0	94,548.0	0.95
	Self-Ensembling + PIT (RPIT)	1,207.2	98,491.0	99,698.2	1.00
	Self-Ensembling + PIT (re-weighting)	1,207.2	91,503.0	92,710.2	0.93

(a) GPA **(b) Ours (GPA)**

Figure 7: Qualitative detection results of task Cityscapes → KITTI, all the predictions are car class.

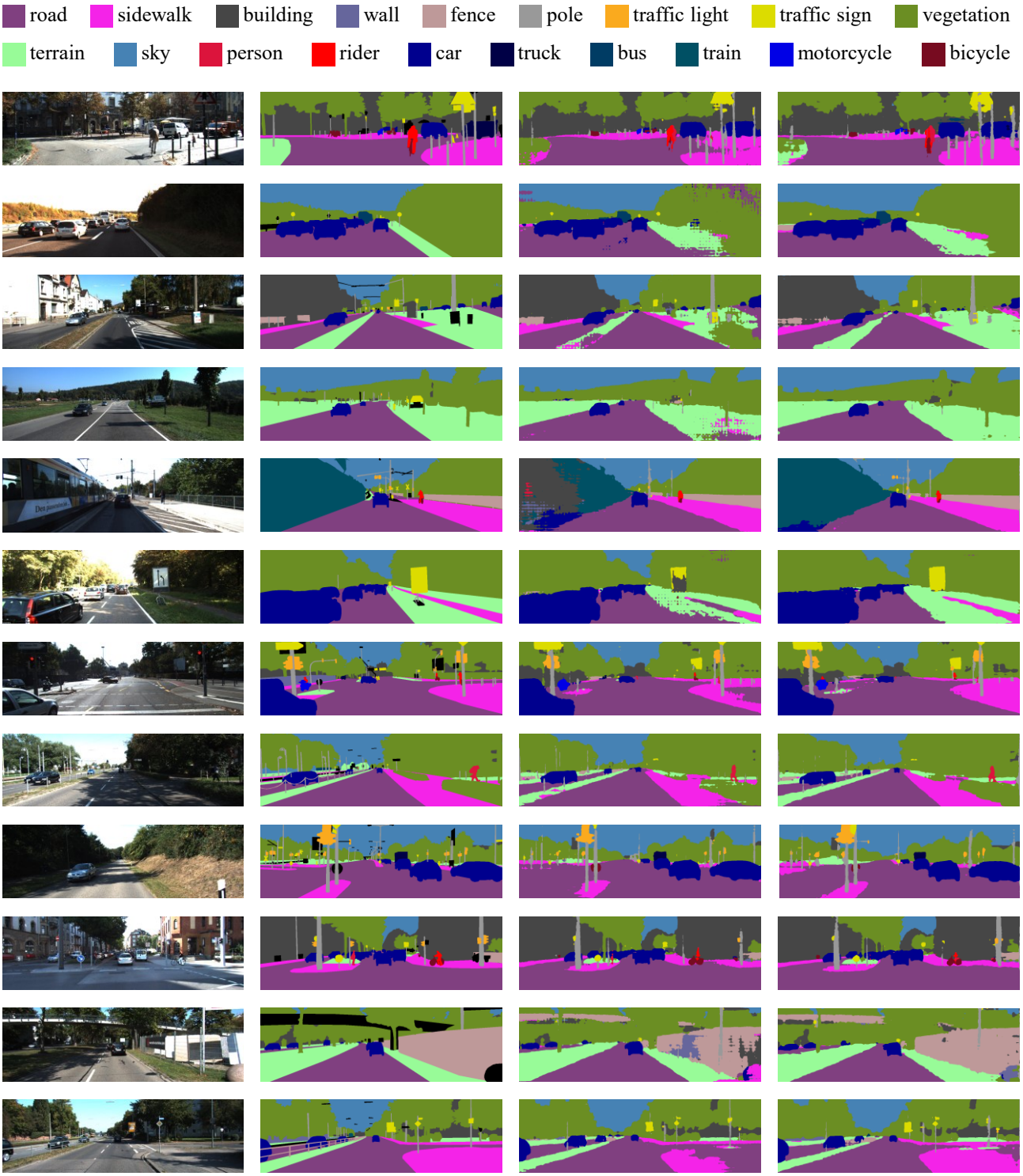

Figure 8: Qualitative segmentation results of task Cityscapes → KITTI.

References

[1] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *CVPR*, 2019.

[2] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, 2018.

[3] Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *ICCV*, 2019.

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[5] Geoff French, Samuli Laine, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *BMVC*, 2020.

[6] Geoff French, Avital Oliver, and Tim Salimans. Milking cowmask for semi-supervised image classification. *arXiv preprint arXiv:2003.12022*, 2020.

[7] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *IJR*, 32(11):1231–1237, 2013.

[8] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, 2019.

[9] Zhiqiang Shen, Harsh Maheshwari, Weichen Yao, and Marios Savvides. Scl: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses. *arXiv preprint arXiv:1911.02559*, 2019.

[10] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. *arXiv preprint arXiv:2007.08702*, 2020.

[11] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *CVPR*, 2020.

[12] Qianyu Zhou, Zhengyang Feng, Guangliang Cheng, Xin Tan, Jianping Shi, and Lizhuang Ma. Uncertainty-aware consistency regularization for cross-domain semantic segmentation. *arXiv preprint arXiv:2004.08878*, 2020.

[13] Qianyu Zhou, Zhengyang Feng, Qiqi Gu, Jiangmiao Pang, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Context-aware mixup for domain adaptive semantic segmentation. *arXiv preprint arXiv:2108.03557*, 2021.