Supplementary Material: Continual Test-time Domain Adaptation via Dynamic Sample Selection

1. Experiments

1.1. Imagenet-R experiment

ImageNet-R [2] encompasses a diverse array of shifts of ImageNet classes. These shifts include cartoons, deviant art, graffiti, embroidery, graphics, origami, paintings, patterns, plastic objects, plush objects, sculptures, sketches, tattoos, toys, and video games. The dataset comprises 200 classes and a total of 30,000 images. Here, we show the CTDA performance result of our DSS method and other baseline approaches, and experiments are conducted using the standard ResNet-50 model, pretrained on ImageNet through cross-entropy loss. In general, all baseline methods show a certain performance improvement compared to direct testing using the source model. The performance of Tent, Conjugate PL, and CoTTA methods showcases a degree of similarity, while the BN method slightly lags behind. Notably, our proposed DSS method achieves the lowest error rate of 56%.

Method	Error
Source	63.8
TENT-cont [6]	57.3
BN Adapt [3]	60.3
Conjugate PL [1]	57.3
CoTTA [7]	57.4
DSS (Ours)	**56.0**

Table 1. Classification error rate (%) on ImageNet-R [2]. The best numbers are in bold.

1.2. Modelnet40-C experiment

ModelNet40-C [5] is a benchmark for assessing the robustness of our proposed method on 3D point cloud data. In this setting, 15 different forms of corruption are introduced to the original test dataset of ModelNet40 [8]. For 3D experiments, random rotation and translation are used in the augmentation module to generate augmentation-weighted pseudo-labels. As shown in Table 2, all methods reduce error by a certain amount and DSS has the lowest error rate in average.

References

[1] Sachin Goyal, Mingjie Sun, Aditi Raghunathan, and J Zico Kolter. Test time adaptation via conjugate pseudo-labels. *Advances in Neural Information Processing Systems*, 35:6204–6218, 2022. 1

[2] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 1

[3] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779*, 2016. 1

[4] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 2

[5] Jiachen Sun, Qingzhao Zhang, Bhavya Kailkhura, Zhiding Yu, Chaowei Xiao, and Z Morley Mao. Benchmarking robustness of 3d point cloud recognition against common corruptions. *arXiv preprint arXiv:2201.12296*, 2022. 1

[6] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. 1, 2

[7] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022. 1, 2

[8] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 1

Method	uniform	gaussian	background	impulse	upsampling	rbf	rbf-inv	den-dec	dens-inc	shear	rot	cut	distort	oclsion	lidar	Mean
Source	14.7	18.8	95.3	33.3	15.0	29.5	27.6	**12.9**	**10.5**	42.7	72.8	**14.9**	34.8	56.3	59.0	35.9
TENT-cont [6]	15.3	**15.6**	92.1	26.6	17.5	**26.5**	**25.1**	16.0	13.0	37.7	**58.7**	17.1	32.6	**54.1**	56.9	33.7
CoTTA [7]	14.3	17.4	90.9	25.5	14.4	27.1	26.1	13.4	12.2	38.4	63.7	15.2	32.5	56.1	**56.6**	33.6
DSS	**14.2**	17.7	**89.5**	**25.0**	**13.9**	26.7	25.4	13.7	12.5	**37.4**	63.6	15.4	**32.3**	54.7	58.1	**33.2**

Table 2. Classification error rate (%) on ModelNet40-C. PointNet [4] is adopted as the backbone. The best numbers are in bold.