Supplementary for "Each Test Image Deserves A Specific Prompt: Continual Test-Time Adaptation for 2D Medical Image Segmentation"

Ziyang Chen[1] Yongsheng Pan[1†] Yiwen Ye[1] Mengkang Lu[1] Yong Xia[1,2,3†]
[1] School of Computer Science and Engineering, Northwestern Polytechnical University, China
[2] Research & Development Institute of Northwestern Polytechnical University in Shenzhen, China
[3] Ningbo Institute of Northwestern Polytechnical University, China
zychen@mail.nwpu.edu.cn, yspan@nwpu.edu.cn, ywye@mail.nwpu.edu.cn
lmk@mail.nwpu.edu.cn, yxia@nwpu.edu.cn

In the supplementary material, we provide more details about our implementation details of medical image segmentation tasks (Section A) and natural classification image task (Section B), and additional experiments of medical image segmentation tasks (Section C) and natural image classification task (Section D).

A. Other Implementation Details of Medical Image Segmentation Tasks

GPU device. We conducted all experiments on a single RTX-2080Ti GPU for two medical image segmentation benchmark tasks.

Training the source model for the OD/OC segmentation task. We utilized the SGD optimizer with a momentum of 0.99 and a weight decay of 0.0005, and the initial learning rate lr_0 was set to 0.001 and decayed according to $lr_t = lr_0 \times (1 - t/T)^{0.9}$, where t is the current epoch and the maximum epoch T is set to 200. The batch size was set to 8. Empirically, we chose the model trained after the last epoch as the testing model.

Training the source model for the polyp segmentation task. We utilized the publicly released code of PraNet [1] to train the source model. The Adam algorithm with a learning rate of 0.0001 was adopted as the optimizer and the maximum epoch was set to 20. The batch size was set to 16. Empirically, we chose the model trained after the last epoch as the testing model.

Implementation details of different test-time adaptation methods. Since these compared methods (*i.e.*, TENT-continual [9], CoTTA [10], DLTTA [12], DUA [5], SAR [7], and DomainAdaptor [13]) are all open-source, we adopted their source codes to conduct experiments. We set the batch size to 1 for all experiments following [12].

B. Other Implementation Details of Natural Image Classification Task

GPU device. We conducted all experiments on a single RTX-3090 GPU.

Datasets and Evaluation Metrics. PACS [3] dataset is commonly used in domain generalization and test-time adaptation, which comprises 9,991 images and 7 classes that are collected from 4 distinct domains: art, cartoons, photos, and sketches.

Training the source model. We trained the source model by empirical risk minimization (ERM) [8] algorithm with ResNet [2] backbone using the SGD optimizer with a learning rate of 5e-5. The batch size was set to 32 and the number of training iteration was set to 5k. We resized all images to 224×224 and used data augmentation during training, including random cropping, random flipping, color jittering, and intensity changing.

Implementation details of comparison experiments under the test-time adaptation setup. We compared our VPTTA with four methods (*i.e.*, BN [6], TENT [9], SHOT-IM [4], and TSD [11]) under the test-time adaptation setup following TSD (*i.e.*, given D domains, training on D-1 domains and testing on the left one) and used the publicly released code of TSD to conduct experiments for all methods.

[†] Yong Xia and Yongsheng Pan are the corresponding authors.

https://github.com/DengPingFan/PraNet
https://github.com/DequanWang/tent
https://github.com/qinenergy/cotta
https://github.com/med-air/DLTTA
https://github.com/jmiemirza/DUA
https://github.com/mr-eggplant/SAR
https://github.com/koncle/DomainAdaptor
https://github.com/SakurajimaMaiii/TSD

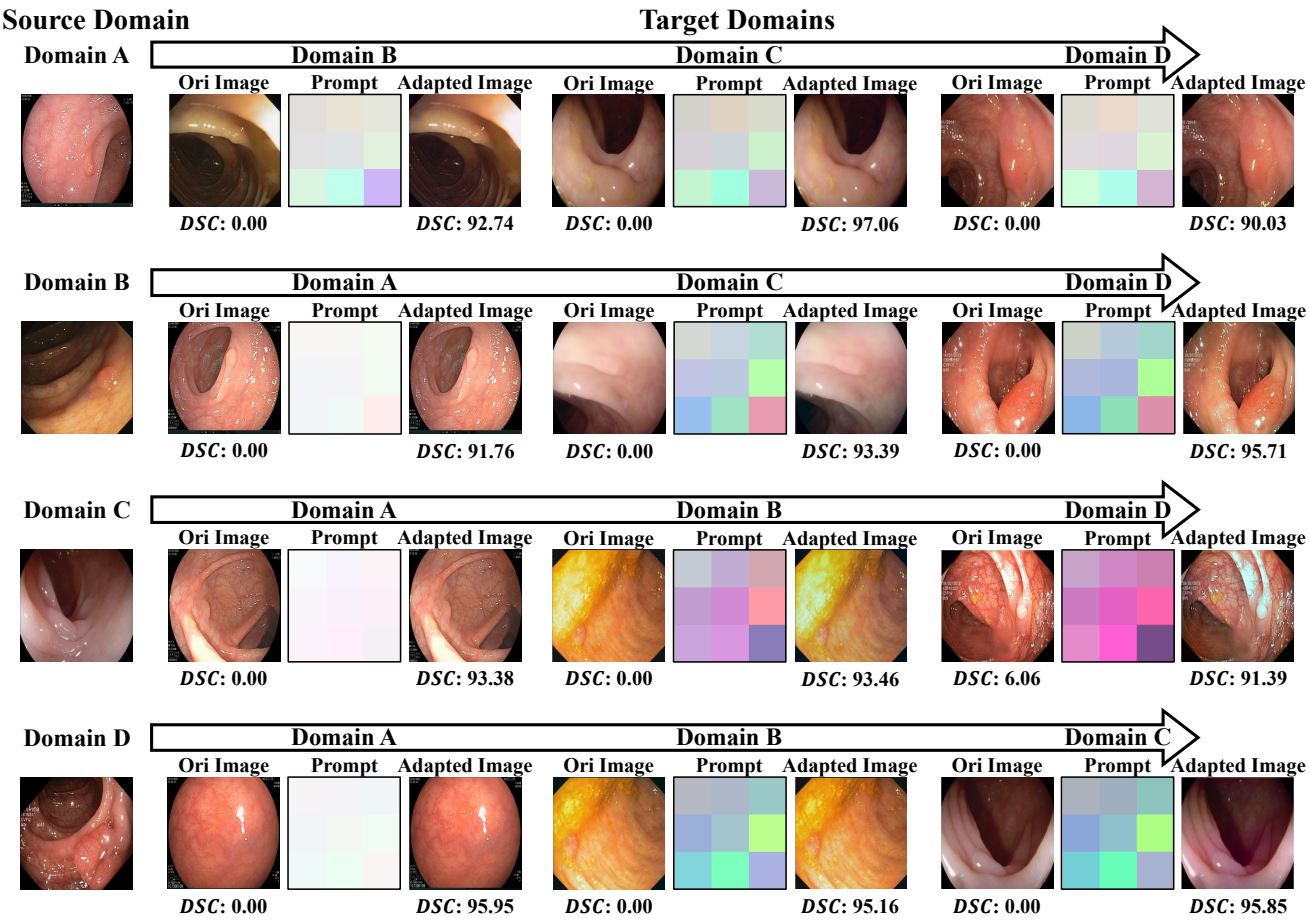

Figure 1. Visualization of the original images, estimated prompts, and adapted images on the polyp segmentation task. We normalize the prompts to [0, 1] for better visualization. The DSC of applying the frozen source model on the original and adapted images is displayed below each image. We also show an example of each source domain on the left side of this diagram. 'Ori': Abbreviation of 'Original'.

Table 1. Performance of our VPTTA, 'Source Only' baseline, and six competing methods on the OD/OC segmentation task. The best and second-best results in each column are highlighted in **bold** and underline, respectively.

Methods	Domain A	Domain B	Domain C	Domain D	Domain E	Average
	DSC	*DSC*	*DSC*	*DSC*	*DSC*	*DSC* ↑
Source Only (ResUNet-34)	64.53	76.06	71.18	52.67	64.87	65.86
TENT-continual (ICLR 2021) [9]	71.50	77.96	72.79	42.97	69.56	66.96
CoTTA (CVPR 2022) [10]	73.71	76.31	72.43	53.04	71.14	69.33
DLTTA (TMI 2022) [12]	**74.90**	78.73	**74.48**	50.99	69.25	69.67
DUA (CVPR 2022) [5]	73.06	75.74	70.82	57.04	70.31	69.39
SAR (ICLR 2023) [7]	74.48	77.49	70.78	**57.93**	73.05	70.75
DomainAdaptor (CVPR 2023) [13]	74.50	76.39	71.81	56.78	70.55	70.01
VPTTA (Ours)	74.24	**79.12**	74.05	55.84	**76.47**	**71.94**

We resized all test images to 224×224 and no data augmentation was used. For all experiments, we set the random seed to 0. To deploy our VPTTA, we utilized the Adam optimizer with a learning rate of 0.01. The hyperparameters α (size of prompt), S (size of memory bank), K (size of support set), and τ (temperature coefficient in warm-up) are set to 0.02, 64, 4, and 1. The code and the weights of pre-trained source models will be available.

Table 2. Performance of our VPTTA, 'Source Only' baseline, and six competing methods on the polyp segmentation task. The best and second-best results in each column are highlighted in **bold** and underline, respectively.

Methods	Domain A			Domain B			Domain C			Domain D			Average		
	DSC	E_ϕ^{max}	S_α	DSC	E_ϕ^{max}	S_α	DSC	E_ϕ^{max}	S_α	DSC	E_ϕ^{max}	S_α	$DSC\uparrow$	$E_\phi^{max}\uparrow$	$S_\alpha\uparrow$
Source Only (PraNet)	79.90	87.97	84.66	66.33	78.51	76.72	73.89	84.64	81.28	82.95	90.84	88.08	75.77	85.49	82.69
TENT-continual (ICLR 2021) [9]	72.72	82.99	79.19	69.41	80.09	79.10	13.38	36.09	51.23	73.70	83.33	82.72	57.30	70.62	73.06
CoTTA (CVPR 2022) [10]	76.29	85.31	82.51	66.58	76.73	79.11	71.29	83.50	80.12	70.62	79.81	82.56	71.20	81.34	81.07
DLTTA (TMI 2022) [12]	75.52	84.69	81.88	66.66	77.21	79.34	63.75	78.79	75.55	70.79	81.14	83.32	69.18	80.46	80.02
DUA (CVPR 2022) [5]	78.79	87.14	83.93	69.13	80.62	79.03	74.66	84.96	82.07	86.63	93.62	90.06	77.30	86.58	83.77
SAR (ICLR 2023) [7]	76.48	85.89	81.49	66.45	77.35	78.05	71.46	83.23	79.40	70.41	80.11	81.07	71.20	81.65	80.00
DomainAdaptor (CVPR 2023) [13]	77.48	86.31	82.40	70.82	81.76	80.88	71.96	83.06	79.97	76.89	85.89	84.45	74.29	84.26	81.93
VPTTA (Ours)	**80.65**	**88.62**	**84.78**	**76.94**	**87.64**	**84.10**	**76.48**	**86.56**	**83.04**	**86.37**	**93.54**	**89.87**	**80.11**	**89.09**	**85.45**

C. More Experiments of Medical Image Segmentation Tasks

C.1. Visualization of prompts and adapted images on the polyp segmentation task

We also visualized the prompts and adapted images on the polyp segmentation task, as shown in Figure 1. We found that the prompts induce subtle alterations in the appearance of images, but still yield substantial performance gains, even on the hard samples which cannot be recognized by the model (*i.e.*, $DSC = 0.00$). Similar to the observation on the OD/OC segmentation task, the prompts of different target domains produced for the same source model exhibit high similarity.

C.2. Comparison experiments under mixed distribution shifts

Considering that test data may come arbitrarily in the complex real world, we conducted the experiments under the mixed distribution shifts, *i.e.*, training the model on a single source domain and testing it on a mixed domain composed of the left target domains. We used 2024 as the random seed to shuffle the data of left target domains for all methods, and the batch size was set to 1. The results of two medical segmentation benchmark tasks are displayed in Table 1 and Table 2. In Table 2, we observed similar phenomena that only DUA and our VPTTA outperform the baseline, but other methods fail due to the wrong gradients produced by the confident but terrible predictions. Meanwhile, the results in Table 1 and Table 2 reveal that our VPTTA still achieves the best overall performance across all domains on both two tasks, underscoring its superior applicability and robustness.

D. More Experiments of Natural Image Classification Task

We evaluated our VPTTA and four methods with different batch sizes (BS) on the PACS dataset. The results are shown in Table 3 and Table 4. We found that other compared methods, such as TSD, perform well with a large test batch size (BS=64) but fail with a small test batch size (BS=1), which means they heavily depend on batch size. The results show that our VPTTA surpasses other compared methods with a small test batch size and achieves competitive performance with a large test batch size. It demonstrates that our VPTTA is suitable to be deployed in scenarios with small batch sizes.

Table 3. Performance of our VPTTA, ERM baseline, and four competing methods on the PACS dataset with ResNet-18 backbone. The best and second-best results in each column are highlighted in bold and underline, respectively.

ResNet-18		A	C	P	S	Average
ERM [8]		78.37	77.05	95.57	65.69	79.17
BS=64	BN (NeurIPS 2020) [6]	81.05	80.63	95.03	72.26	82.24
	Tent (ICLR 2021) [9]	81.35	80.89	95.27	73.45	82.74
	SHOT-IM (ICML 2020) [4]	82.71	76.75	94.97	67.32	80.44
	TSD (CVPR 2023) [11]	**87.89**	**86.90**	**96.53**	71.77	**85.77**
	VPTTA (Ours)	81.15	80.67	96.23	**77.40**	83.86
BS=1	BN (NeurIPS 2020) [6]	13.09	16.60	11.32	5.17	11.54
	Tent (ICLR 2021) [9]	10.06	16.60	11.32	4.07	10.51
	SHOT-IM (ICML 2020) [4]	13.77	16.60	11.32	5.83	11.88
	TSD (CVPR 2023) [11]	12.99	15.53	11.80	16.75	14.27
	VPTTA (Ours)	**79.35**	**77.52**	**95.63**	**69.56**	**80.51**

Table 4. Performance of our VPTTA, ERM baseline, and four competing methods on the PACS dataset with ResNet-50 backbone. The best and second-best results in each column are highlighted in bold and underline, respectively.

ResNet-50		A	C	P	S	Average
ERM [8]		83.35	79.82	96.77	81.85	85.45
BS=64	BN (NeurIPS 2020) [6]	84.96	82.81	96.83	75.46	85.01
	Tent (ICLR 2021) [9]	85.45	83.36	96.77	77.30	85.72
	SHOT-IM (ICML 2020) [4]	83.30	82.00	93.35	63.99	80.66
	TSD (CVPR 2023) [11]	**88.87**	**89.12**	**97.43**	82.13	**89.39**
	VPTTA (Ours)	86.47	83.53	97.37	**84.78**	88.04
BS=1	BN (NeurIPS 2020) [6]	29.30	16.60	11.32	4.10	15.33
	Tent (ICLR 2021) [9]	19.19	16.60	11.32	4.07	12.79
	SHOT-IM (ICML 2020) [4]	31.74	16.60	11.32	4.20	15.96
	TSD (CVPR 2023) [11]	12.21	14.16	12.57	17.03	13.99
	VPTTA (Ours)	**84.57**	**81.66**	**97.25**	**82.62**	**86.52**

References

[1] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, pages 263–273. Springer, 2020. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 1

[3] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Int. Conf. Comput. Vis.*, pages 5542–5550, 2017. 1

[4] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Int. Conf. Mach. Learn.*, pages 6028–6039. PMLR, 2020. 1, 3

[5] M Jehanzeb Mirza, Jakub Micorek, Horst Possegger, and Horst Bischof. The norm must go on: Dynamic unsupervised domain adaptation by normalization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 14765–14775, 2022. 1, 2, 3

[6] Zachary Nado, Shreyas Padhy, D Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020. 1, 3

[7] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *Int. Conf. Learn. Represent.*, 2023. 1, 2, 3

[8] Vladimir Vapnik. Statistical learning theory. *Wiley*, 1998. 1, 3

[9] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *Int. Conf. Learn. Represent.*, 2021. 1, 2, 3

[10] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7201–7211, 2022. 1, 2, 3

[11] Shuai Wang, Daoan Zhang, Zipei Yan, Jianguo Zhang, and Rui Li. Feature alignment and uniformity for test time adaptation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 20050–20060, 2023. 1, 3

[12] Hongzheng Yang, Cheng Chen, Meirui Jiang, Quande Liu, Jianfeng Cao, Pheng Ann Heng, and Qi Dou. Dltta: Dynamic learning rate for test-time adaptation on cross-domain medical images. *IEEE Trans. Med. Imaging*, 41(12):3575–3586, 2022. 1, 2, 3

[13] Jian Zhang, Lei Qi, Yinghuan Shi, and Yang Gao. Domainadaptor: A novel approach to test-time adaptation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18971–18981, 2023. 1, 2, 3