SATTA: Semantic-Aware Test-Time Adaptation for Cross-Domain Medical Image Segmentation

Yuhan Zhang[1,2,3], Kun Huang[4], Cheng Chen[5](✉), Qiang Chen[4],
and Pheng-Ann Heng[1,2]

[1] Department of Computer Science and Engineering,
The Chinese University of Hong Kong, Hong Kong, China
zhangyuh@cse.cuhk.edu.hk

[2] Institute of Medical Intelligence and XR, The Chinese University of Hong Kong,
Hong Kong, China

[3] Shenzhen Research Institute, The Chinese University of Hong Kong,
Hong Kong, China

[4] Department of Computer Science and Engineering,
Nanjing University of Science and Technology, Nanjing, China

[5] Center for Advanced Medical Computing and Analysis,
Harvard Medical School and Massachusetts General Hospital, Boston, USA
cchen101@mgh.harvard.edu

Abstract. Cross-domain distribution shift is a common problem for medical image analysis because medical images from different devices usually own varied domain distributions. Test-time adaptation (TTA) is a promising solution by efficiently adapting source-domain distributions to target-domain distributions at test time with unsupervised manners, which has increasingly attracted important attention. Previous TTA methods applied to medical image segmentation tasks usually carry out a global domain adaptation for all semantic categories, but global domain adaptation would be sub-optimal as the influence of domain shift on different semantic categories may be different. To obtain improved domain adaptation results for different semantic categories, we propose Semantic-Aware Test-Time Adaptation (SATTA), which can individually update the model parameters to adapt to target-domain distributions for each semantic category. Specifically, SATTA deploys an uncertainty estimation module to measure the discrepancies of semantic categories in domain shift effectively. Then, a semantic adaptive learning rate is developed based on the estimated discrepancies to achieve a personalized degree of adaptation for each semantic category. Lastly, semantic proxy contrastive learning is proposed to individually adjust the model parameters with the semantic adaptive learning rate. Our SATTA is extensively validated on retinal fluid segmentation based on SD-OCT images. The experimental results demonstrate that SATTA consistently

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43895-0_14.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14221, pp. 148–158, 2023.
https://doi.org/10.1007/978-3-031-43895-0_14

improves domain adaptation performance on semantic categories over other state-of-the-art TTA methods.

Keywords: test-time adaptation · domain shift · medical image segmentation

1 Introduction

Deep learning has achieved remarkable success in medical image segmentation when the training and test data are independent and identically distributed (i.i.d) [4,13,21]. However, in many practical situations, training and test data are collected by different medical imaging devices, leading to the presence of distribution shifts. Therefore, the models trained on source-domain data perform poorly on target-domain data. An effective solution for this issue is to fine-tune the models with labeled target-domain data to adapt to the target-domain distributions [9], but it is impractical to label the target-domain data considering the high annotation cost. Existing unsupervised domain adaptation (UDA) methods [7,16,22] make full use of the labeled source-domain data and the unlabeled target-domain data in the model training, but even the target-domain data may not be available in the model training due to various practical problems.

Domain generalization (DG) methods exploit the diversity of source domains to improve the model generalization [8,14,24] when target-domain data is not available in the model training. However, it is also difficult and cost-consuming to collect multiple source-domain datasets with different domain distributions for DG. Another promising solution is test-time adaptation (TTA), which aims to gradually update the model parameters to adapt to target-domain distributions by learning from test data at test time. TTA shows greater flexibility than DG as the models could be pre-trained on single source-domain data. In TTA, a mainstream strategy is to adjust the affine parameters in BN layers for domain adaptation at test time by unsupervised loss, such as PTBN [12] and TENT [19]. Besides, auxiliary self-supervised tasks [10,17] and contrastive learning [3,20] are also considerable for TTA.

TTA methods have also been recently applied in medical image applications. Ma et al. [11] innovated distribution calibration by dynamically aggregating multiple representative classifiers via TTA to deal with arbitrary label shifts. Hu et al. [6] designed regional nuclear-norm loss and contour regularization loss for TTA on medical image segmentation tasks. Bateson et al. [1] performed inference by minimizing the entropy of predictions and a class-ratio prior, and integrated shape priors through penalty constraints for guide adaptation. Varsavsky et al. [18] introduced domain adversarial learning and consistency training in TTA for sclerosis lesion segmentation. These TTA methods have a common limitation of using a fixed learning rate for all test samples. Since test samples arrive sequentially and the scale of domain shift would change frequently, a fixed learning rate would be sub-optimal for TTA. DLTTA [23] proposed a memory bank-based discrepancy measurement for dynamic learning rate adjustment of

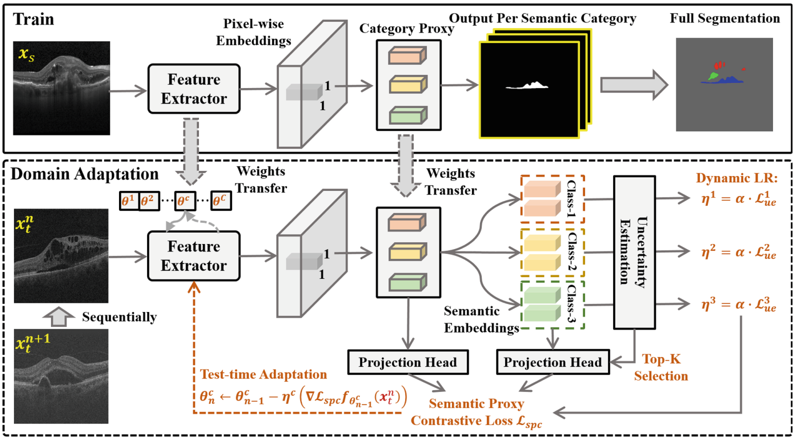
Train
x_s
Feature Extractor
Pixel-wise Embeddings
1
1
Category Proxy
Output Per Semantic Category
Full Segmentation
Domain Adaptation
Weights Transfer
Weights Transfer
θ^1 θ^2 ··· θ^c ··· θ^c
x_t^n
Feature Extractor
Sequentially
x_t^{n+1}
Class-1
Class-2
Class-3
Uncertainty Estimation
Dynamic LR:
$\eta^1 = \alpha \cdot \mathcal{L}_{ue}^1$
$\eta^2 = \alpha \cdot \mathcal{L}_{ue}^2$
$\eta^3 = \alpha \cdot \mathcal{L}_{ue}^3$
Semantic Embeddings
Projection Head
Projection Head
Top-K Selection
Test-time Adaptation
$\theta_n^c \leftarrow \theta_{n-1}^c - \eta^c\left(\nabla \mathcal{L}_{spc} f_{\theta_{n-1}^c}(x_t^n)\right)$
Semantic Proxy Contrastive Loss $\mathcal{L}_{spc}$

Fig. 1. Overview of SATTA for cross-domain medical image segmentation. Semantic adaptive learning rates are obtained by the uncertainty estimation module, and a semantic proxy contrastive loss is designed for individual semantic domain adaption.

TTA to effectively adapt the model to the varying domain shift. However, we find that the influence of domain shift on different semantic categories may also be different, DLTTA performed global domain adaptation for all semantic categories.

In this paper, we present **S**emantic-**A**ware **T**est-**T**ime **A**daptation (SATTA) for cross-domain medical image segmentation, aiming to perform individual domain adaptation for each semantic category at test time. SATTA first utilizes an uncertainty estimation module to effectively measure the discrepancies of different semantic categories in domain shift. Based on the estimated discrepancies, a semantic adaptive learning rate is then developed to achieve a personalized degree of adaptation for each semantic category. Lastly, a semantic proxy contrastive loss is proposed to individually adjust the model parameters with the semantic adaptive learning rate. Our SATTA is evaluated on retinal fluid segmentation based on spectral-domain optical coherence tomography (SD-OCT) images, and the experimental results show superior performance than other state-of-the-art TTA methods.

2 Methods

2.1 Test-Time Adaptation Review

Given a labeled source-domain dataset $\mathcal{S} = \{(\boldsymbol{x}_n^s, \boldsymbol{y}_n^s)\}_{n=1}^{N^s}$, model parameters θ are pre-trained on $\mathcal{S}$ by supervised risk minimization:

$$\theta^s = \arg\min_{\theta} \frac{1}{N^s} \sum_{n=1}^{N^s} \mathcal{L}_{sup}(\mathcal{F}_\theta(\boldsymbol{x}_n^s), \boldsymbol{y}_n^s) \tag{1}$$

where $\mathcal{L}_{sup}$ is the supervised loss for model optimization, such as the cross-entropy loss. However, for an unlabeled target-domain dataset $\mathcal{T} = \{(\boldsymbol{x}_n^t)\}_{n=1}^{N^t}$ that has different domain distributions with $\mathcal{S}$, the model $\mathcal{F}_{\theta^s}$ may have an obvious performance degeneration. To make the model $\mathcal{F}_{\theta^s}$ adapt to the target-domain distributions, an unsupervised TTA loss $\mathcal{L}_{tta}$ (such as rotation prediction loss [17], entropy minimization loss [19], contrastive loss [3], etc.) is designed to fine-tune model based on target-domain samples at test time:

$$\theta_n^t \leftarrow \theta_{n-1}^t - \eta(\bigtriangledown \mathcal{L}_{tta}(\mathcal{F}_{\theta_{n-1}^t}(\boldsymbol{x}_n^t))),\ n \in [1, N^t] \tag{2}$$

where η is learning rate, θ_0^t is initialized with θ^s. The final prediction on $\boldsymbol{x}_n^t$ can be given by $\boldsymbol{y}_n^t \sim \hat{\boldsymbol{y}}_n^t = \mathcal{F}_{\theta_n^t}(\boldsymbol{x}_n^t)$.

2.2 Semantic Adaptive Learning Rate

Pseudo-labeling for Semantic Aggregation. The semantic segmentation model contains a feature extractor and category predictor. For pixels $\{\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_{hw}\}$ in image $\boldsymbol{x}$, feature extractor encodes them into high-dimensional pixel embeddings $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_{hw}\} \in \mathbb{R}^d$, and category predictor outputs the categories of the pixel embeddings. Category predictor is a projection matrix $\boldsymbol{W} \in \mathbb{R}^{d \times C} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_C]$ consisted of C category proxies, where each category proxy $\boldsymbol{w}_c$ can be regarded as the high-dimensional representative of the category. Category predictor measures the similarities between pixel embeddings and all category proxies. We represent the classification process of pixel $\boldsymbol{p}_i$ as:

$$\boldsymbol{y}_i \sim \hat{\boldsymbol{y}}_i \in \mathbb{R}^C = \boldsymbol{W} \cdot (f(\boldsymbol{p}_i)),\ i \in [1, hw] \tag{3}$$

where $f(\cdot)$ is the feature extractor, C is the category number, h and w are the height and width of images. Since pixel-wise labels are not available for target-domain samples at test time, we are hard to obtain the semantic information of all pixel embeddings directly. To address this problem, we assign pseudo labels to all pixel embeddings by passing them through the category predictor and then aggregate all pixel embeddings into C semantic clusters as $[\Omega_1, \Omega_2, \cdots, \Omega_C]$ according to their pseudo labels.

Semantic Uncertainty Estimation. After performing semantic aggregation by pseudo-labeling, we need to estimate the varying discrepancies of domain shift on categories. Here, we employ Monte Carlo Dropout [5] for semantic uncertainty estimation. We enable dropout at test time and perform L stochastic forward passes through the model to obtain a set of predictive outputs for pixel embedding $\boldsymbol{e}_i$:

$$u_i^l = \mathcal{M}(\boldsymbol{e}_i),\ l \in [1, L],\ i \in [1, hw] \tag{4}$$

where $\mathcal{M}(\cdot)$ is a mapping network that maps the pixel embedding $\boldsymbol{e}_i$ into an additional probability output. Then we estimate the standard deviation of the L outputs as the uncertainty score of $\boldsymbol{e}_i$:

$$s_i = std(u_i^1, u_i^2, \cdots, u_i^L),\ i \in [1, hw] \tag{5}$$

Here we take a single pixel as an example to show the uncertainty score computation, it should be noted that all pixels are performed parallel computation in the semantic segmentation model. Therefore, the computation cost does not increase with the number of pixels. For category c, its semantic uncertainty score can be calculated by:

$$\mathcal{U}^c = \frac{1}{N^{\Omega_c}} \sum_{\boldsymbol{e}_i \in \Omega_c} s_i,\ c \in [1, C],\ i \in [1, hw] \tag{6}$$

where N^{Ω_c} is the number of pixel embeddings in Ω_c. $\mathcal{U}^c$ captures the unique semantic domain discrepancy over category c. Later, semantic adaptive learning rate η^c of category c for TTA is obtained directly based on the semantic domain discrepancy $\mathcal{U}^c$:

$$\eta^c = \alpha \cdot \mathcal{U}^c \tag{7}$$

where α is a scale factor. In this work, α could be set as the learning rate used for the model pre-training with the source-domain dataset. Each semantic category has its own individual learning rate in each iteration.

2.3 Semantic Proxy Contrastive Learning

General contrastive losses focus on exploring rich sample-to-sample relations, but they are hard to learn specific semantic information from samples. Proxy contrastive loss can model semantic relations by category proxies, as category proxies are more robust intuitively to noise samples [24]. Therefore, a proxy contrastive loss is more suitable for unsupervised TTA optimization with our proposed semantic adaptive learning rate.

Projection Heads. We regard each category proxy as the anchor and consider all proxy-to-sample relations. Since proxy-based methods converge very easily, we consider applying projection heads to map both pixel embeddings and category proxies to a new feature space where proxy contrastive loss is applied. Given semantic clusters $[\Omega_1, \Omega_2, \cdots, \Omega_C]$ and category proxy weights $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_C]$, We use a three-layer MLP $\mathcal{H}_1(\cdot)$ for projecting pixel embeddings and one-layer MLP $\mathcal{H}_2(\cdot)$ for projecting category proxy weights. The new pixel embedding and category proxy weight can be given by $\boldsymbol{z}_i = \mathcal{H}_1(\boldsymbol{e}_i)$ and $\boldsymbol{v}_c = \mathcal{H}_2(\boldsymbol{w}_c)$.

Top-K Selection. Pixel embeddings with high uncertainty scores contribute little to semantic proxy contrastive learning. Besides, the computation cost is huge for all pixel embeddings. To address this problem, we select K pixel embeddings with the highest confidence from each semantic cluster Ω_c. Specifically, for each semantic cluster Ω_c, we first order all pixel embeddings in it from smallest to largest according to their uncertainty scores. Then we select the first K pixel embeddings as the new semantic cluster Ω'_c for the next proxy contrastive loss by $\Omega'_c = TopK(Order(\Omega_c))$.

Semantic Proxy Contrastive Loss. For an anchor category cluster $\mathit{\Omega}'_c$, we associate all pixel embeddings in it with category proxy weight $\boldsymbol{v}_c$ to form the positive pairs. We ignore the sample-to-sample positive pairs and only consider the sample-to-sample negative pairs. The semantic proxy contrastive loss for category c can be given by:

$$\begin{aligned} &\mathcal{L}_{spc}(\boldsymbol{x}, \boldsymbol{W}, c) = -\frac{1}{K} \sum_{\boldsymbol{z}_i \in \mathit{\Omega}'_c} \log \frac{\exp(\boldsymbol{v}_c^\top \boldsymbol{z}_i \cdot \tau)}{\mathcal{Z}} \\ &s.t.\ \mathcal{Z} = \exp(\boldsymbol{v}_c^\top \boldsymbol{z}_i \cdot \tau) + \sum_{r_0=1}^{C-1} \exp(\boldsymbol{v}_{r_0}^\top \boldsymbol{z}_i \cdot \tau) + \frac{1}{C'} \sum_{r_1=1}^{C'} \sum_{\boldsymbol{z}_j \in \mathit{\Omega}'_{r_1}} \exp(\boldsymbol{z}_i^\top \boldsymbol{z}_j \cdot \tau) \end{aligned} \tag{8}$$

where $\{\boldsymbol{z}_i\}_{i=1}^K$ are obtained by $\boldsymbol{x}$ and C' is the number of categories appearing in samples.

2.4 Training and Adaptation Procedure

The overview of our SATTA is shown in Fig. 1. Given the source-domain dataset $\mathcal{S} = \{(\boldsymbol{x}_n^s, \boldsymbol{y}_n^s)\}_{n=1}^{N^s}$, the model parameters θ are pre-trained by the combination of supervised cross-entropy loss and semantic proxy contrastive loss:

$$\mathcal{L}_{total} = \mathcal{L}_{ce}(\boldsymbol{x}_n^s, \boldsymbol{y}_n^s) + \lambda \cdot \frac{1}{C} \sum_{c=1}^{C} \mathcal{L}_{spc}(\boldsymbol{x}_n^s, \boldsymbol{W}, c) \tag{9}$$

At test time, for a target-domain sample at time step n, we perform a forward pass to obtain semantic clusters and uncertainty scores and calculate the semantic adaptive learning rate of each category to serve for semantic proxy contrastive loss. For category c, the model parameters are updated to achieve desired adaptation by:

$$\theta_n^c \leftarrow \theta_{n-1}^c - \eta^c(\bigtriangledown \mathcal{L}_{spc}(f_{\theta_{n-1}^c}(\boldsymbol{x}_n^t))),\ n \in [1, N^t] \tag{10}$$

The updated model parameters are stored in a memory bank and will be loaded for the next domain adaptation of category c. If category c does not appear in the test sample by pseudo-labeling, we ignore the update of model parameters θ_{n-1}^c. We only update the parameters of the feature extractor and freeze the parameters of the category predictor.

3 Experiments

3.1 Materials

Our SATTA was evaluated on retinal fluid segmentation based on RETOUCH challenge [2], which is a representative benchmark for segmenting all of the three fluid types in SD-OCT images, including intraretinal fluid (IRF), subretinal fluid

Table 1. Quantitative comparison of different TTA methods on RETOUCH challenge using DSC metric. (Note: CD denotes cross domain.)

Methods	Domain-1			Domain-2			Domain-3		
	IRF	SRF	PED	IRF	SRF	PED	IRF	SRF	PED
U-Net [15] w/o CD	0.716	0.794	0.802	0.738	0.904	0.786	0.711	0.664	0.768
U-Net [15] w/ CD	0.637	0.751	0.667	0.587	0.733	0.624	0.536	0.512	0.565
TENT [19]	0.648	0.772	0.733	0.605	0.828	0.703	0.589	0.603	0.637
FTTA [6]	0.663	0.769	0.746	0.632	0.831	0.729	0.611	0.618	0.669
TTAS [1]	0.672	0.774	0.762	0.674	0.865	0.762	0.663	0.641	0.724
CoTTA [20]	0.668	0.776	0.755	0.683	0.852	0.754	0.670	0.634	0.716
SATTA (Ours)	**0.704**	**0.788**	**0.786**	**0.722**	**0.883**	**0.781**	**0.702**	**0.658**	**0.743**

(SRF) and pigment epithelial detachment (PED). SD-OCT images were acquired by three different vendors: Cirrus, Spectralis, and Topcon. The training set consists of 3072 (Cirrus), 1176 (Spectralis), and 3072 (Topcon) SD-OCT images, and the test set consists of 1792 (Cirrus), 686 (Spectralis) and 1792 (Topcon) SD-OCT images. We regard the SD-OCT images from three different vendors as three different domains, namely Domain-1 (Cirrus), Domain-2 (Spectralis), and Domain-3 (Topcon). We employ the dice similarity coefficient (DSC) as the quantitative segmentation metric and a higher DSC indicates a better segmentation performance.

3.2 Comparison with State-of-the-Arts

We compare our SATTA with four state-of-the-art TTA methods, including TENT [19], FTTA [6], TTAS [1] and CoTTA [20]. The four comparative methods have been reviewed in Sect. 1. To verify the TTA performance on cross-domain retinal fluid segmentation based on the RETOUCH challenge with three different domains, we train the segmentation models on two source domains and run TTA methods on the remaining target domain at test time. We carry out three times until all of three domains are tested as unseen target domains. For fair comparisons, all of TTA methods use U-Net [15] as a feature extractor and share the same experimental setting, such as initial learning rate, batch size, etc.

The quantitative comparison results are presented in Table 1. We also include “U-Net w/ CD” as the lower bound and “U-Net w/o CD" as the upper bound, where “CD” denotes cross domain. “U-Net w/o CD” denotes that the segmentation model is trained and tested on the samples from the same domain. “U-Net w/ CD” denotes that the segmentation model is trained and tested on the samples from different domains, without any domain adaptation. We observe that different TTA methods consistently improve the segmentation performance over “U-Net w/ CD”. Our SATTA achieves the highest DSC than other methods. The qualitative comparison results are shown in Fig. 2. These visual results confirm

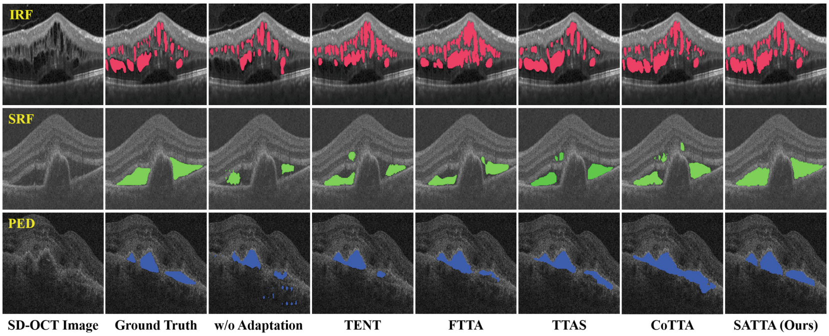

Fig. 2. Qualitative comparison of cross-domain segmentation of different TTA methods on RETOUCH challenge. The first row shows the IRF segmentation on Spectralis SD-OCT images, the second row shows the SRF segmentation on Topcon SD-OCT images, and the third row shows the PED segmentation on Cirrus SD-OCT images.

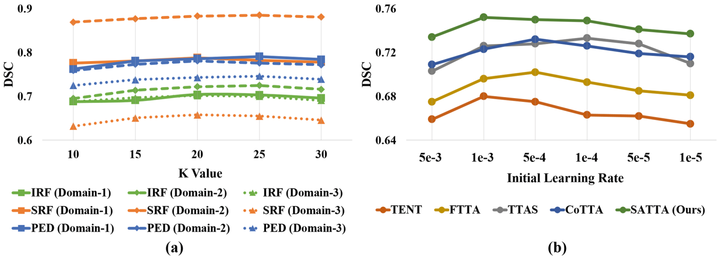

Fig. 3. (a) Ablation study with different K values in SATTA. (b) Ablation study with different initial learning rates in all TTA methods.

that a segmentation model trained only on source-domain distributions performs poorly on target-domain distributions without domain adaptation.

3.3 Ablation Study

We conduct ablation studies to analyze the key factors regarding our SATTA. We first explore the effect of K value in the Top-K selection strategy. The Top-K selection strategy aims to select pixel embeddings with high confidence scores to improve the semantic proxy contrastive learning and reduce the computation cost significantly. Figure 3(a) shows the effect of different K values on three domains for IRF, SRF, and PED. The DSC values consistently increase when rising the K value from 10 to 20, generally, peak when the K value is between 20 and 25, and consistently decrease when further rising K value. This affirms that the pixel embeddings with high confidence scores are conducive to semantic proxy contrastive learning while the pixel embeddings with low confidence scores weaken semantic proxy contrastive learning.

We also investigate the effect of the initial learning rate. We select different initial learning rates from the set {5e-3, 1e-3, 5e-4, 1e-4, 5e-5, 1e-5} for TTA, and Fig. 3(b) shows the total average DSC values of all TTA methods. Our SATTA consistently performs better than other state-of-the-art TTA methods. We also find that different initial learning rates actually affect the domain adaptation ability. Therefore, a proper initial learning rate is essential for TTA methods.

4 Conclusion

In this paper, we present the SATTA method for cross-domain medical image segmentation. Aiming at the problem that the domain shift has different effects on the semantic categories, our SATTA provides a semantic adaptive parameter optimization scheme at test time. Although our SATTA shows superior cross-domain segmentation performance than other state-of-the-art methods, it still has a limitation. Since SATTA adjusts the model for each semantic category, it is not quite suitable for the samples with too many semantic categories due to high computation costs.

Acknowledgements. This work was supported in part by the Shenzhen Portion of Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone under HZQB-KCZYB-20200089, and partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Number: T45-401/22-N) and by a grant from the Hong Kong Innovation and Technology Fund (Project Number: GHP/080/20SZ). This work was also supported by the National Natural Science Foundation of China under Grants (62202408, 62172223).

References

1. Bateson, M., Lombaert, H., Ben Ayed, I.: Test-time adaptation with shape moments for image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 736–745. Springer (2022). https://doi.org/10.1007/978-3-031-16440-8_70
2. Bogunović, H., et al.: Retouch: the retinal oct fluid detection and segmentation benchmark and challenge. IEEE Trans. Med. Imaging **38**(8), 1858–1874 (2019)
3. Chen, D., Wang, D., Darrell, T., Ebrahimi, S.: Contrastive test-time adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 295–305 (2022)
4. Farshad, A., Yeganeh, Y., Gehlbach, P., Navab, N.: Y-net: a spatiospectral dual-encoder network for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 582–592. Springer (2022). https://doi.org/10.1007/978-3-031-16434-7_56
5. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: representing model uncertainty in deep learning. In: International Conference on Machine Learning, pp. 1050–1059. PMLR (2016)
6. Hu, M., et al.: Fully Test-Time Adaptation for Image Segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 251–260. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_24

7. Hu, S., Liao, Z., Xia, Y.: Domain specific convolution and high frequency reconstruction based unsupervised domain adaptation for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 650–659. Springer (2022). https://doi.org/10.1007/978-3-031-16449-1_62
8. Lee, S., Seong, H., Lee, S., Kim, E.: Wildnet: learning domain generalized semantic segmentation from the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9936–9946 (2022)
9. Li, J., Li, X., He, D., Qu, Y.: A domain adaptation model for early gear pitting fault diagnosis based on deep transfer learning network. Proc. Instit. Mech. Eng. Part O: J. Risk Reliabi. **234**(1), 168–182 (2020)
10. Liu, Y., Kothari, P., van Delft, B., Bellot-Gurlet, B., Mordan, T., Alahi, A.: Ttt++: when does self-supervised test-time training fail or thrive? Adv. Neural. Inf. Process. Syst. **34**, 21808–21820 (2021)
11. Ma, W., Chen, C., Zheng, S., Qin, J., Zhang, H., Dou, Q.: Test-time adaptation with calibration of medical image classification nets for label distribution shift. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 313–323. Springer (2022). https://doi.org/10.1007/978-3-031-16437-8_30
12. Nado, Z., Padhy, S., Sculley, D., D'Amour, A., Lakshminarayanan, B., Snoek, J.: Evaluating prediction-time batch normalization for robustness under covariate shift. arXiv preprint arXiv:2006.10963 (2020)
13. Peiris, H., Hayat, M., Chen, Z., Egan, G., Harandi, M.: A robust volumetric transformer for accurate 3d tumor segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 162–172. Springer (2022). https://doi.org/10.1007/978-3-031-16443-9_16
14. Peng, D., Lei, Y., Hayat, M., Guo, Y., Li, W.: Semantic-aware domain generalized segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2594–2605 (2022)
15. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
16. Sun, X., Liu, Z., Zheng, S., Lin, C., Zhu, Z., Zhao, Y.: Attention-enhanced disentangled representation learning for unsupervised domain adaptation in cardiac segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 745–754. Springer (2022). https://doi.org/10.1007/978-3-031-16449-1_71
17. Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., Hardt, M.: Test-time training with self-supervision for generalization under distribution shifts. In: International Conference on Machine Learning. pp. 9229–9248. PMLR (2020)
18. Varsavsky, T., Orbes-Arteaga, M., Sudre, C.H., Graham, M.S., Nachev, P., Cardoso, M.J.: Test-time unsupervised domain adaptation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 428–436. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_42
19. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: fully test-time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726 (2020)
20. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7201–7211 (2022)

21. Xing, Z., Yu, L., Wan, L., Han, T., Zhu, L.: Nestedformer: nested modality-aware transformer for brain tumor segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 140–150. Springer (2022). https://doi.org/10.1007/978-3-031-16443-9_14
22. Xu, Z., et al.: Denoising for relaxing: unsupervised domain adaptive fundus image segmentation without source data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 214–224. Springer (2022). https://doi.org/10.1007/978-3-031-16443-9_21
23. Yang, H., et al.: Dltta: Dynamic learning rate for test-time adaptation on cross-domain medical images. arXiv preprint arXiv:2205.13723 (2022)
24. Yao, X., et al.: Pcl: proxy-based contrastive learning for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7097–7107 (2022)