Cache-Driven Spatial Test-Time Adaptation for Cross-Modality Medical Image Segmentation

Xiang Li[1], Huihui Fang[3,4(✉)], Changmiao Wang[5], Mingsi Liu[6], Lixin Duan[1], and Yanwu Xu[2,4]

[1] Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen, China
[2] School of Future Technology, South China University of Technology, Guangzhou, China
[3] College of Computing and Data Science, Nanyang Technological University, Singapore.
[4] Pazhou Lab, Guangzhou, China
fanghuihuibit@163.com
[5] Medical Big Data Lab, Shenzhen Research Institute of Big Data, Shenzhen, China
[6] Assumption University of Thailand, Bangkok, Thailand

Abstract. Test-Time Adaptation (TTA) shows promise for addressing the domain gap between source and target modalities in medical image segmentation methods. Furthermore, TTA enables the model to quickly fine-tune itself during testing, enabling it to adapt to the continuously evolving data distribution in the medical clinical environment. Consequently, we introduce Spatial Test-Time Adaptation (STTA), for the first time considering the integration of inter-slice spatial information from 3D volumes with TTA. The continuously changing distribution of slice data in the target domain can lead to error accumulate on and catastrophic forgetting. To tackle these challenges, we first propose reducing error accumulation by using an ensemble of multi-head predictions based on data augmentation. Secondly, for pixels with unreliable pseudo-labels, regularization is applied through entropy minimization on the ensemble of predictions from multiple heads. Finally, to prevent catastrophic forgetting, we suggest using a cache mechanism during testing to restore neuron weights from the source pre-trained model, thus effectively preserving source knowledge. The proposed STTA has been bidirectionally validated across modalities in abdominal multi-organ and brain tumor datasets, achieving a relative increase of approximately 13% in the Dice value in the best-case scenario compared to SOTA methods. The code is available at: https://github.com/lixiang007666/STTA.

Keywords: Test-Time Adaptation · Cross-Modality · Medical Image Segmentation · Spatial Information · Cache Mechanism.

1 Introduction

Medical image segmentation plays an important role in clinical applications such as computer-aided diagnosis. In recent years, Deep Learning (DL) methods have

been widely used for medical image segmentation [17,3,12]. DL methods are data-driven and are usually based on the assumption that the training and test data follow the same distribution. However, if there is a data distribution inconsistency (called domain gap) between the training and test data, the deep model usually leads to a dramatic performance degradation. Domain gap usually occurs in real-world scenarios of medical image processing [4], when the training and test medical image data come from different locations, different scanners, and even different modalities, such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI).

Domain Adaptation (DA) holds the promise of addressing the domain gap issue between training and testing data [4,20,21,24]. However, one limitation of DA is that the model is fixed after training and cannot be adjusted during testing. Test-Time Adaptation (TTA) allows the model to quickly fine-tune and adapt during testing, enabling it to handle the evolving data distribution in real clinical environments, where the data distribution is constantly changing. Nado et al. [10] proposed a TTA method called PTBN, which modifies the statistical parameters in the Batch Normalization (BN) layers based on the data from the target domain. Wang et al. [16] introduced TENT, tuning BN layers by minimizing prediction entropy in the target domain. These methods, originally intended for natural images, assume that updates to the BN layers can adequately bridge domain gaps. However, this assumption has shown limited effectiveness in TTA for medical images [15]. Prabhu et al. [11] proposed URMA, using pseudo-labels and uncertainties from multiple branches to aid adaptation. However, due to potential inaccuracies in its pseudo-labels [13], it risks error accumulation. Wu et al. [22] proposed UPL-TTA, a method capable of adapting the source model to unlabeled target domains without knowledge of the source model's training strategy. Nevertheless, it still suffers from catastrophic forgetting.

To avoid the issues of error accumulation and catastrophic forgetting mentioned above, while enabling the model to quickly fine-tune itself during testing to adapt to the constantly evolving distribution of medical data in clinical environments, we propose the Spatial Test-Time Adaptation (STTA) method for cross-modality medical image segmentation. The main contributions of this work are summarized as follows: 1) We present the STTA method, conceptualized for clinical environments, enabling off-the-shelf source pre-trained models to effectively adapt to continuously changing target medical data. It accomplishes this by assimilating inter-slice spatial information from 3D volumes into TTA; 2) We propose an innovative approach utilizing an ensemble of multi-head predictions based on data augmentation and applying entropy minimization to the ensemble results to reduce error accumulation; 3) We introduce a cache mechanism to efficiently preserve knowledge from the source model, mitigating the effects of catastrophic forgetting.

2 Proposed method

As shown in **Fig. 1**, STTA comprises three main components.

The first component of STTA is designed to mitigate error accumulation. We propose improving the quality of pseudo-labels within a self-training framework in two distinct ways. On one hand, inspired by the observation that predictions from a mean teacher model [13] often exceed the quality of those from a standard model, we utilize a mean teacher model as the foundation of STTA to provide more accurate predictions. On the other hand, for test data significantly influenced by domain gaps, we employ a data-augmented multi-head predictions ensemble to further enhance the quality of pseudo-labels. The second component addresses pixels with unreliable pseudo-labels by applying entropy minimization to the ensemble of predictions from multiple heads for regularization. Lastly, the third component aims to prevent forgetting. We suggest a cache mechanism to restore the weights of some operators from the source domain pre-trained model to the target domain student model.

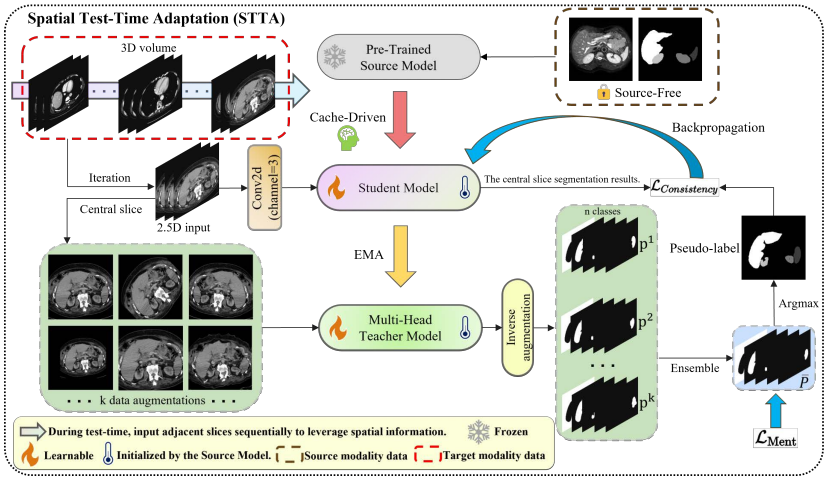

Fig. 1: Overview of our STTA, where p^k is the soft prediction of the k-th head. STTA requires inputting three adjacent slices of a 3D volume in sequence.

For the acquisition of the source model, let S represent the source domain with data distribution $\mu_S(x)$, and T represent the target domain with data distribution $\mu_T(x)$. Let $X_S = \{(x_i^s, y_i^s) \mid i = 1, \ldots, N_s\}$ denote the training images and their labels in the source domain, and $X_T = \{(x_j^t) \mid j = 1, \ldots, N_t\}$ represent unlabeled slices in the target domain for adaptation. Note that $\mu_S(x) \neq \mu_T(x)$. The pre-training stage in the source domain is represented as:

$$f_{\theta_0}(x) = \arg\min \frac{1}{N_s} \sum_{i=1}^{N_s} L_s \left(f_{\theta_0} \left(\boldsymbol{x}_i^s \right), y_i^s \right), \tag{1}$$

where $f_{\theta_0}(x)$ represents the source domain model to be transferred. L_s denotes the training loss in the source domain.

2.1 Spatial insights for TTA

Directly using 3D volumes incurs significant inference overheads, such as memory and time, making it unsuitable for TTA deployment scenarios. Conversely, 2D images do not fully exploit spatial information. The 2.5D strategy becomes an effective compromise. In previous work [25,8], their approach involved stacking slices from all 3D surfaces to implement the 2.5D. Inspired by these studies, STTA also adopts the 2.5D strategy, sequentially inputting three adjacent slices (one central slice and two adjacent slices). Furthermore, sequential input enhances TTA performance as the data distribution exhibits smoother transitions.

2.2 Implementation of the STTA algorithm

As illustrated in **Algorithm 1**, the fusion of enhanced pseudo-labels with the cache mechanism gives rise to our STTA method.

Algorithm 1 The proposed STTA.

1: **Initialization:** A source pre-trained model f_{θ_0} in Eq 1 (As the initial weights for the student model), teacher model $f_{\theta'_0}$ initialized from f_{θ_0}.
2: **Input:** For the j-th central slice (along the channel direction), current stream of data x_j^t.
3: **for** the j-th slice **do**
4: Input x_j^t and its augmentation into the teacher network $f_{\theta'_j}$ to obtain pseudo-labels that have undergone an ensemble of multi-head predictions and entropy minimization, based on Eq 2 and Eq 3.
5: Update student f_{θ_j} by consistent backpropagation in Eq 4.
6: Update teacher $f_{\theta'_j}$ by Exponential Moving Average (EMA) in Eq 5.
7: Restore the student model f_{θ_j} using caching through Eq 6.
8: **end for**
9: **Output:** Prediction $f_{\theta'_j}(x_j^t)$; Updated student model $f_{\theta_{j+1}}$; Updated teacher model $f_{\theta'_{j+1}}$.

STTA duplicates the decoder of the teacher model k times to implement a multi-head predictions ensemble. The data augmentations used include rotation, horizontal flipping, scaling, zooming, and elastic deformation [18]. The formulaic expression is as follows:

$$\begin{aligned} \tilde{y'}_j^t &= \frac{1}{k}\sum_{i=0}^{k-1} f_{\theta'_j}\left(\mathrm{aug}_k\left(x_j^t\right)\right), \\ {y'}_j^t &= \begin{cases} \hat{y'}_j^t, & \text{if } \mathrm{conf}\left(f_{\theta_0}\left(x_j^t\right)\right) \geq p_{th} \\ \tilde{y'}_j^t, & \text{otherwise,} \end{cases} \end{aligned} \tag{2}$$

where $\tilde{y'}_j^t$ is the data-augmented multi-head predictions ensemble from the teacher model, $\hat{y'}_j^t$ is the direct prediction from the teacher model, conf $\left(f_{\theta_0}\left(x_j^t\right)\right)$ represents the prediction confidence of the source pre-trained model on the slice x_j^t of the current 3D volume, and p_{th} is a confidence threshold. Our hypothesis is that lower confidence suggests a larger domain gap, while relatively high confidence levels indicate a smaller domain gap [19]. Consequently, when confidence is high and exceeds the threshold, we utilize $\hat{y'}_j^t$ directly as our pseudo-label, abstaining from any augmentation. In cases of low confidence, we employ additional k augmentations to enhance the quality of the pseudo-label.

Additionally, we applied Mean Prediction-Based Entropy Minimization [22] to the results after predictions ensemble, resulting in optimized pseudo-labels:

$$\mathcal{L}_{\text{Ment}} = -\frac{1}{HW}\sum_{n=1}^{HW}\sum_{c=1}^{C}\overline{\boldsymbol{p}}_{c,n}\log\left(\overline{\boldsymbol{p}}_{c,n}\right), \tag{3}$$

where $\overline{\boldsymbol{p}}$ represents the probability map obtained by averaging k predictions. H, W, and C represent the shape of the probability map, with $n = 1, 2, \ldots, HW$ being the pixel index, and $c = 1, 2, \ldots, C$, where C corresponds to the number of segmentation classes. Compared to individually minimizing the entropy of each augmented prediction head, minimizing the entropy of their mean prediction $\overline{\boldsymbol{p}}$ can not only reduce the uncertainty of a single augmented prediction head but also promote consensus among k heads for the same test sample, thereby enhancing the model's prediction robustness for unseen test samples.

$$\mathcal{L}_{\text{Consistency}} = -\sum_{c} y'^t_{jc}\log\hat{y}^t_{jc}, \tag{4}$$

$$\theta'_{j+1} = \alpha\theta'_j + (1-\alpha)\theta_{j+1}, \tag{5}$$

where y'^t_{jc} represents the probability of class c in the teacher model's pseudo-label prediction, while $\hat{y}^t_{jc}$ denotes the prediction from the student model. The loss aims to enforce consistency between the teacher and student predictions. Upon updating the student model $f_{\theta_j} \rightarrow f_{\theta_{j+1}}$ using Eq 4, we further update the teacher model's weights using the EMA of the student model's weights through Eq 5, wherein α represents the smoothing factor. Our ultimate prediction (inference) for the input slice x_j^t is determined by identifying the class that exhibits the maximum probability within y'^t_j.

2.3 Cache mechanism

We designed a cache-driven method to efficiently recover knowledge from the source model, thus mitigating the effect of catastrophic forgetting. Segmentation networks are typically structured as encoder-decoder architectures, and we define "shallow layers" as those near the input end of the encoder and the output end of the decoder (a layer consists of Conv2d, BN, and ReLU). The threshold for dividing layers into "deep and shallow" can be adjusted to suit the dataset [9].

Accordingly, in STTA, we divide the weights of the source model into deep cache and shallow cache. The deep cache focuses on essential global features, while the shallow cache concentrates more on local features such as textures. During adaptation, the deep cache is accessed when the number of slices reaches a threshold, and the shallow cache is accessed for each remaining iteration. The use of deep cache leans towards stability and maintaining global information, whereas the frequent access of shallow cache allows for a rapid response to changes in local features. Let W_{j+1} represent the model weights in the student model after the gradient update at slice j. The implementation of cache mechanism is described as follows:

$$W_{j+1} = \begin{cases} \text{Merge}\left(W_{j+1}^{\text{S}}, W_{\text{cache}}^{\text{D}}\right), & \text{if } j = th \\ \text{Merge}\left(W_{j+1}^{\text{D}}, W_{j+1}^{\text{S}} \odot M + W_{\text{cache}}^{\text{S}} \odot (1 - M)\right), & \text{otherwise}, \end{cases} \tag{6}$$

where $M \sim \text{Bernoulli}(p)$ denotes a binary mask tensor, where $W_{\text{cache}}^{\text{D}}$ and $W_{\text{cache}}^{\text{S}}$ respectively represent the deep and shallow cache of the initial source weights. D refers to the network's deep layers, while S indicates the shallow layers. When j equals the threshold th, the deep cache is fully restored into W_{j+1}. Otherwise, the model's shallow cache is randomly restored into W_{j+1} according to M.

3 Experiments

3.1 Dataset and implementation

The Abdominal dataset. This dataset is commonly utilized in DA tasks [1,14]. It comprises two subsets of abdominal data: 20 MRI scans from the CHAOS challenge [6] and 30 CT scans from the Multi-Atlas Labeling Beyond the Cranial Vaulti-Workshop and Challenge [7]. The dataset includes labels for four organs: liver, right kidney (R.kidney), left kidney (L.kidney), and spleen. Each MRI scan has dimensions of $256 \times 256 \times L$ within a 3D volume, where L represents the length of the long axis and varies among subjects. Each CT scan is sized at $512 \times 512 \times L$, and we crop the images to $256 \times 256 \times L$. In the source domain, for each modality, we randomly split the dataset into training and test sets with a ratio of 4:1. In the target domain, as it is a TTA task, all data from the target modalities were used as the test set.

The BraTs 2018 dataset. This dataset is a comprehensive dataset that includes multimodal 3D brain MRIs along with their corresponding ground truth segmentations. This dataset encompasses four MRI modalities for each case, namely T1, T1c, T2, and FLAIR. In our research, we focus on two specific MRI modalities for low-grade glioma cases, FLAIR and T2 [20]. The method of dataset partitioning is consistent with that of the Abdominal Dataset. Additionally, we resized each axial slice to dimensions of 192×168.

Implementation details. We implemented STTA on a device equipped with a 6-core 42Gi GeForce RTX 3090 using torch-1.8.1-cu11.1-cudnn8. In the pre-training phase within the source domain, we used Dice loss and the Adam optimizer, training the model for 200 epochs with an initial learning rate of

0.001 that decays by 10% every 20 epochs. During the adaptation periods in the target domain, model parameters were updated over 20 epochs using the Adam optimizer with a fixed learning rate of 0.0001. STTA employs a 2.5D strategy by sequentially inputting three stacked slices of a 3D volume to generate segmentation results for the central slice of each stack. In Eq 6, the th value is set to the total number of slices in a 3D volume. For other hyperparameters, the default EMA factor (α) was set to 0.999. To filter images and avoid augmentations on high-confidence images, we used a threshold (p_{th}) in Eq 2, defined as $p_{th} = \text{conf}^S - \delta$. Here, conf^S indicates the 5% quantile of softmax confidence from the source model f_{θ_0}, and δ, set to 0.05, serves as a minor tolerance, enabling STTA to establish a threshold without test data reliance [5]. Additionally, in the current dataset, we found that setting one shallow cache layer is optimal.

3.2 Results

The Abdominal results. We conducted bidirectional cross-modality TTA experiments between MRI and CT scans in the Abdominal dataset, refer to **Table 1**. "Source-only" refers to the baseline approach of applying the source model directly to the target data without any adaptation. In the "Target supervised" approach, the model was exclusively trained using annotated images from the target domain. To ensure a fair comparison, all the compared methods employed the same backbone (DeepLabV3 [2]). In the $CT \rightarrow MRI$ setting, STTA achieved up to a 13% higher Dice score than other SOTA methods (UPL-TTA). Furthermore, it also performs better on average symmetric surface distance (ASSD) [23]. In **Fig. 2**, we also conducted a visual qualitative comparison that includes "Source only," "PTBN," "TENT," "URMA," "UPL-TTA," and "STTA". STTA is closer to the ground truth, with fewer overfitting or underfitting pixels.

The BraTS 2018 results. In this section, to highlight STTA's generalization, we performed brain tumor segmentation tasks, including bidirectional cross-modality TTA experiments with FLAIR and T2 modalities. As shown in **Table 2**, STTA still achieves the best Dice and ASSD values. Additionally, in **Fig. 3**, a qualitative comparison for brain tumor segmentation shows that our method is closer to the ground truth.

Ablation study. We performed an in-depth analysis to assess the impact of STTA's components. The baseline method involved solely utilizing the pretrained model's predictions as pseudo-labels for adaptation, while the introduced components consisted of: 1) Spatial. Sequentially inputting three slices, in contrast to the traditional method of randomly inputting a single slice into the model; 2) Ensemble. Ensemble of multi-head predictions based on data augmentation; 3) Cache. Utilizing a cache mechanism to preserve source knowledge; 4) L_{Ment}. Corresponding to Eq 3. As these components are gradually introduced, the performance of STTA improves incrementally, as shown in **Table 3**. Additionally, we set k=6 (the number of teacher model heads), with the input including the five augmentations mentioned in section 2.2 and the original image. Performance improves with increasing k, plateauing at k=7. Considering performance and memory trade-offs, we opted for k=6.

Table 1: Quantitative comparison with other methods on the Abdominal dataset.

Method ($CT \rightarrow MRI$)		Dice ↑					ASSD ↓				
		Liver	R.kidney	L.kidney	Spleen	Average	Liver	R.kidney	L.kidney	Spleen	Average
	Source only	0.582	0.789	0.686	0.009	0.517	4.661	2.678	1.545	14.518	5.850
	Target supervised	0.727	0.941	0.941	0.713	0.831	3.250	0.484	0.272	6.120	2.532
TTA	PTBN [10]	0.587	0.766	0.752	0.495	0.650	4.659	2.323	2.410	7.478	4.217
	TENT [16]	0.599	0.771	0.747	0.504	0.655	4.655	2.369	2.406	7.512	4.236
	URMA [11]	0.546	0.762	0.726	0.466	0.625	4.739	2.359	2.665	8.704	4.617
	UPL-TTA [22]	0.622	0.759	0.722	0.507	0.653	4.489	**2.188**	2.280	7.372	4.082
	STTA (Our)	**0.645**	**0.813**	**0.817**	**0.679**	**0.739**	**4.458**	2.212	**1.449**	**6.260**	**3.595**
Method ($MRI \rightarrow CT$)		Dice ↑					ASSD ↓				
		Liver	R.kidney	L.kidney	Spleen	Average	Liver	R.kidney	L.kidney	Spleen	Average
	Source only	0.842	0.607	0.622	0.520	0.648	6.628	5.141	5.067	6.914	5.938
	Target supervised	0.961	0.917	0.915	0.945	0.934	1.071	1.052	1.363	0.644	1.033
TTA	PTBN [10]	0.769	0.565	0.646	0.616	0.649	4.170	3.421	3.860	3.193	3.661
	TENT [16]	0.818	0.591	0.680	0.616	0.676	4.280	3.255	3.743	3.012	3.573
	URMA [11]	0.762	0.546	0.651	0.609	0.642	4.795	4.974	4.087	4.442	4.575
	UPL-TTA [22]	0.839	0.619	0.726	0.664	0.712	4.595	4.710	3.972	4.208	4.371
	STTA (Our)	**0.889**	**0.645**	**0.750**	**0.708**	**0.748**	**3.773**	**2.684**	**3.275**	**2.524**	**3.064**

Table 2: Comparison of different SOTA methods on the BRATS 2018 dataset.

Method		($FLAIR \rightarrow T2$)		($T2 \rightarrow FLAIR$)	
		Dice ↑	ASSD ↓	Dice ↑	ASSD ↓
	Source only	0.656	5.520	0.770	3.718
	Target supervised	0.859	2.393	0.886	1.483
TTA	PTBN [10]	0.672	5.468	0.786	3.604
	TENT [16]	0.703	5.428	0.791	3.563
	URMA [11]	0.665	5.512	0.787	3.617
	UPL-TTA [22]	0.714	4.402	0.798	3.411
	STTA (Our)	**0.752**	**3.630**	**0.838**	**3.306**

Table 3: Ablation study of the proposed method on the BRATS 2018 dataset. FLAIR and T2 were used as the source and target domains, respectively.

Components				Dice ↑	ASSD ↓
Spatial	Ensemble	Cache	L_{Ment}		
				0.671	5.376
✓				0.708	4.801
✓	✓			0.730	4.033
✓	✓	✓		0.749	3.796
✓	✓	✓	✓	**0.752**	**3.630**

4 Conclusions

We propose a novel method termed STTA that improves long-term adaptation in constantly changing medical clinical environments, addressing domain gaps between the source and target domains. STTA represents a pioneering effort to integrate inter-slice spatial information into TTA for medical image segmentation. To curb error accumulation, STTA employs a multi-head prediction ensemble derived from data-augmented inputs and enforces consistency by minimizing the entropy of the ensemble's aggregated outputs. Additionally, STTA introduces a caching mechanism that is employed during iterative processes to reinstate the source model weights, thereby preventing the occurrence of catastrophic forgetting. Empirical evaluations on both the Abdominal and BraTS 2018 datasets have yielded evidence of STTA's capability to substantially enhance segmentation performance, thereby affirming its practical efficacy.

Acknowledgments. This research is supported by Pazhou Laboratory's basic cloud computing platform.

Disclosure of Interests. The authors declare no competing interests.

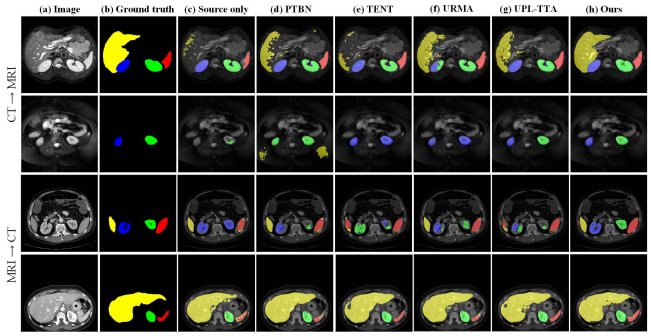

Fig. 2: Visualization of the segmentation results on the Abdominal dataset. The structure of the Liver, R.kidney, L.kidney, and Spleen are shown in yellow, green, blue, and red colors, respectively.

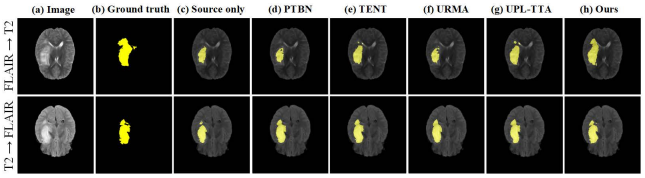

Fig. 3: Visualization of the BraTS 2018 dataset segmentation results.

References

1. Chen, C., Dou, Q., Chen, H., Qin, J., Heng, P.A.: Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. Proceedings of the AAAI Conference on Artificial Intelligence **33**(01), 865–872 (Jul 2019). https://doi.org/10.1609/aaai.v33i01.3301865, https://ojs.aaai.org/index.php/AAAI/article/view/3874
2. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
3. Fang, H., Li, F., Fu, H., Wu, J., Zhang, X., Xu, Y.: Dataset and evaluation algorithm design for goals challenge. In: Antony, B., Fu, H., Lee, C.S., MacGillivray, T., Xu, Y., Zheng, Y. (eds.) Ophthalmic Medical Image Analysis. pp. 135–142. Springer International Publishing, Cham (2022)
4. Guan, H., Liu, M.: Domain adaptation for medical image analysis: A survey. IEEE Transactions on Biomedical Engineering **69**(3), 1173–1185 (2022). https://doi.org/10.1109/TBME.2021.3117407
5. Guillory, D., Shankar, V., Ebrahimi, S., Darrell, T., Schmidt, L.: Predicting with confidence on unseen distributions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1134–1144 (October 2021)
6. Kavur, A.E., Gezer, N.S., et al., M.B.: Chaos challenge - combined (ct-mr) healthy abdominal organ segmentation. Medical Image Analysis **69**, 101950 (2021). https://doi.org/https://doi.org/10.1016/j.media.2020.101950, https://www.sciencedirect.com/science/article/pii/S1361841520303145

7. Landman, B., Xu, Z., Iglesias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge, 2015 . https://doi.org/10.7303/syn3193805
8. Lv, P., Wang, J., Wang, H.: 2.5d lightweight riu-net for automatic liver and tumor segmentation from ct. Biomedical Signal Processing and Control **75**, 103567 (2022). https://doi.org/https://doi.org/10.1016/j.bspc.2022.103567, https://www.sciencedirect.com/science/article/pii/S1746809422000891
9. Ma, X., Fang, G., Wang, X.: Deepcache: Accelerating diffusion models for free. arXiv **abs/2312.00858** (2023)
10. Nado, Z., Padhy, S., Sculley, D., D'Amour, A., Lakshminarayanan, B., Snoek, J.: Evaluating prediction-time batch normalization for robustness under covariate shift. arXiv **abs/2006.10963** (2020)
11. S, P.T., Fleuret, F.: Uncertainty reduction for model adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9613–9623 (June 2021)
12. Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J.N., Wu, Z., Ding, X.: Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. Medical Image Analysis **63**, 101693 (2020). https://doi.org/https://doi.org/10.1016/j.media.2020.101693, https://www.sciencedirect.com/science/article/pii/S136184152030058X
13. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/68053af2923e00204c3ca7c6a3150cf7-Paper.pdf
14. Tomar, D., Lortkipanidze, M., Vray, G., Bozorgtabar, B., Thiran, J.P.: Self-attentive spatial adaptive normalization for cross-modality domain adaptation. IEEE Transactions on Medical Imaging **40**(10), 2926–2938 (2021). https://doi.org/10.1109/TMI.2021.3059265
15. Tomar, D., Vray, G.M.G., Bozorgtabar, B., Thiran, J.P.: Opttta: Learnable test-time augmentation for source-free medical image segmentation under domain shift. pp. 1192–1217. PMLR (2022), http://infoscience.epfl.ch/record/303995
16. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=uXl3bZLkr3c
17. Wang, J., Li, X., Cheng, Y.: Towards an extended efficientnet-based u-net framework for joint optic disc and cup segmentation in the fundus image. Biomedical Signal Processing and Control **85**, 104906 (2023). https://doi.org/https://doi.org/10.1016/j.bspc.2023.104906, https://www.sciencedirect.com/science/article/pii/S1746809423003397
18. Wang, J., Lv, P., Wang, H., Shi, C.: Sar-u-net: Squeeze-and-excitation block and atrous spatial pyramid pooling based residual u-net for automatic liver segmentation in computed tomography. Computer Methods and Programs in Biomedicine **208**, 106268 (2021). https://doi.org/https://doi.org/10.1016/j.cmpb.2021.106268, https://www.sciencedirect.com/science/article/pii/S0169260721003424
19. 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 (CVPR). pp. 7201–7211 (June 2022)

20. Wang, Y., Cheng, J., Chen, Y., Shao, S., Zhu, L., Wu, Z., Liu, T., Zhu, H.: Fvp: Fourier visual prompting for source-free unsupervised domain adaptation of medical image segmentation. IEEE Transactions on Medical Imaging **42**(12), 3738–3751 (2023). https://doi.org/10.1109/TMI.2023.3306105
21. Wu, J., Gu, R., Dong, G., Wang, G., Zhang, S.: Fpl-uda: Filtered pseudo label-based unsupervised cross-modality adaptation for vestibular schwannoma segmentation. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). pp. 1–5 (2022). https://doi.org/10.1109/ISBI52829.2022.9761706
22. Wu, J., Gu, R., Lu, T., Zhang, S., Wang, G.: Upl-tta: Uncertainty-aware pseudo label guided fully test time adaptation for fetal brain segmentation. In: Frangi, A., de Bruijne, M., Wassermann, D., Navab, N. (eds.) Information Processing in Medical Imaging. pp. 237–249. Springer Nature Switzerland, Cham (2023)
23. Yeghiazaryan, V., Voiculescu, I.D.: Family of boundary overlap metrics for the evaluation of medical image segmentation. Journal of Medical Imaging **5**(1), 015006 (2018). https://doi.org/10.1117/1.JMI.5.1.015006, https://doi.org/10.1117/1.JMI.5.1.015006
24. You, F., Li, J., Zhu, L., Chen, Z., Huang, Z.: Domain adaptive semantic segmentation without source data. In: Proceedings of the 29th ACM International Conference on Multimedia. p. 3293–3302. MM '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3474085.3475482, https://doi.org/10.1145/3474085.3475482
25. Zhang, H., Valcarcel, A.M., Bakshi, R., Chu, R., Bagnato, F., Shinohara, R.T., Hett, K., Oguz, I.: Multiple sclerosis lesion segmentation with tiramisu and 2.5d stacked slices. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. pp. 338–346. Springer International Publishing, Cham (2019)