Uncertainty and Shape-Aware Continual Test-Time Adaptation for Cross-Domain Segmentation of Medical Images

Jiayi Zhu[1,2](✉), Bart Bolsterlee[1,2], Brian V. Y. Chow[1,2], Yang Song[1], and Erik Meijering[1]

[1] University of New South Wales, Sydney, Australia
jiayi.zhu3@unsw.edu.au
[2] Neuroscience Research Australia (NeuRA), Randwick, Australia

Abstract. Continual test-time adaptation (CTTA) aims to continuously adapt a source-trained model to a target domain with minimal performance loss while assuming no access to the source data. Typically, source models are trained with empirical risk minimization (ERM) and assumed to perform reasonably on the target domain to allow for further adaptation. However, ERM-trained models often fail to perform adequately on a severely drifted target domain, resulting in unsatisfactory adaptation results. To tackle this issue, we propose a generalizable CTTA framework. First, we incorporate domain-invariant shape modeling into the model and train it using domain-generalization (DG) techniques, promoting target-domain adaptability regardless of the severity of the domain shift. Then, an uncertainty and shape-aware mean teacher network performs adaptation with uncertainty-weighted pseudo-labels and shape information. Lastly, small portions of the model's weights are stochastically reset to the initial domain-generalized state at each adaptation step, preventing the model from 'diving too deep' into any specific test samples. The proposed method demonstrates strong continual adaptability and outperforms its peers on three cross-domain segmentation tasks. Code is available at https://github.com/ThisGame42/CTTA.

Keywords: Continual Test-Time Adaptation · Segmentation · Convolutional Neural Networks

1 Introduction

Deep neural networks (DNN) have demonstrated state-of-the-art performance in medical image segmentation in recent years [1]. In practice, the discrepancies in the distributions between the target domain, where the test data come from, and the source domain that provides the training data, often lead to reduced test-time performance (Fig. 1). This phenomenon, known as the domain shift

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43898-1_63.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14222, pp. 659–669, 2023.
https://doi.org/10.1007/978-3-031-43898-1_63

[5], is common in medical imaging [2], thus necessitates model re-training across institutes, resulting in a waste of resources and precluding the use of DNNs in budget-challenged scenarios.

Many studies have attempted to address the domain shift. Earlier works adapt models to the target domain with access to the source domain [6–8], restricting their applications due to privacy concerns. In response, methods utilizing prior or anatomical information to remove the need for source data are proposed [9,10], yet their flexibility is limited. Lately, test-time domain adaptation (TTA), continual test-time adaptation (CTTA), and domain generalization (DG) methods have been gaining popularity [2,3,19]. TTA methods train a model on a labeled source domain and adapt it to an unlabeled target domain with access to target data only. Adaptation is usually performed via feature alignment through generative models [12], domain adversarial learning and paired consistency [13], image/feature translation via adaptor networks [14], and entropy minimization which fine-tunes the parameters of the batch normalization (BN) layers [15] on test data [2,16]. CTTA is an emerging approach aiming to improve the robustness of TTA methods during long-term continual adaptation, a scenario where TTA methods are susceptible to catastrophic forgetting and become overfitted to later test samples. Examples include stochastic parameter restoration [3] and normalization correction and data resampling [4]. DG methods aim to produce a more generalizable model from one or more source domains without updating parameters at test time. Popular methods involve data augmentations to enhance domain robustness [17] and learn domain-invariant features [18,19].

CTTA could be a useful technique to segment patient data acquired at different time points of longitudinal studies. However, we note that adaptation is possible only when the source model, typically trained with empirical risk minimization (ERM) [2–4,12,14–16], already demonstrates reasonable target-domain performance as the starting point. ERM models may struggle to provide adequate performance for further adaptation in severe domain shifts (see panels (g)–(l), Fig. 1). Domain knowledge can be utilized to design a preprocessing procedure that reduces the domain gap [20] and enables ERM models to perform adequately on the target domain. However, the effort to design preprocessing procedures significantly increases when a trained source model is shared with multiple end-users to account for different test-time data distributions.

To address those issues, we propose a generalizable CTTA framework for the cross-domain segmentation task of medical images. We first incorporate shape-aware feature learning into existing models and train them on the source domain with DG techniques. This removes the need for carefully preprocessed target domain data and allows the source model to perform reasonably in most target domains regardless of the severity of the domain shift. Then, we use an uncertainty-weighted multi-task mean teacher network inspired by semi-supervised literature to perform adaptation, producing results with improved accuracy and refined contours. In addition, a small portion of the model weight is stochastically reset to its initial, domain-generalized state at each adaptation step to prevent the model from overfitting to later test samples. We show the

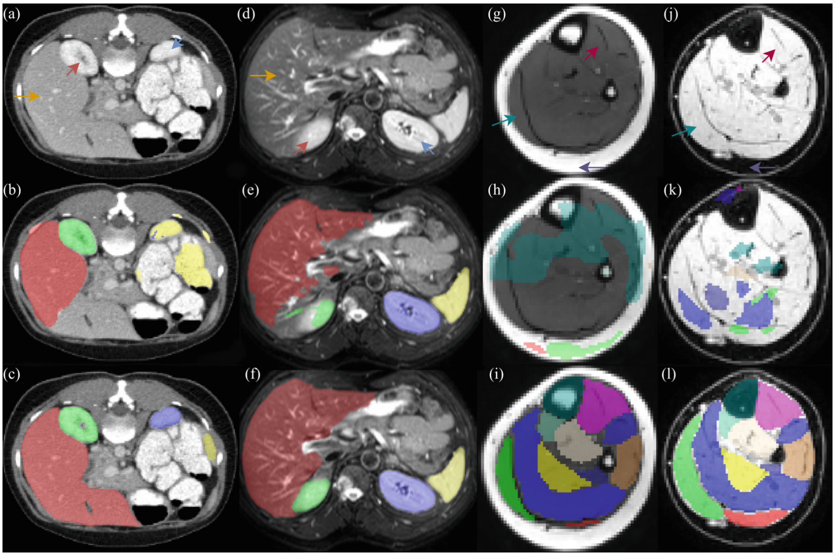

Fig. 1. Demonstrations of different severities of domain shifts. Panels (a) and (d) are preprocessed CT and MRI T_2 abdominal scans and (c) and (f) are their manual labels. (b) and (e) are cross-domain predictions for (a) and (d) by an ERM model trained on (d) and (a), respectively. Panels (g) and (j) are preprocessed MRI T_1 and mDixon muscle scans and (i) and (l) are their manual labels. (h) and (k) are labels predicted for (g) and (j) by an ERM model trained on (j) and (g), respectively. Arrows of the same color indicate the same anatomical structures across different domains.

proposed framework works with ERM and DG-trained source models and (1) outperforms several state-of-the-art methods on three challenging cross-domain segmentation tasks and (2) is better suited for CTTA than its peers in various scenarios.

2 Methodology

Overview. The proposed framework is a synergy of three components (Fig. 2): (1) shape-aware model training, (2) shape and uncertainty-aware mean teacher network for the model update, and (3) domain-generalized stochastic weight restoration for continual adaptation. Component (1) is used for model training in the source domain, while (2) and (3) are used simultaneously for CTTA. We describe each component in detail below.

Shape-Aware Model Training. Motivated by recent studies [18,19] suggesting that shape information enables generalizable performance due to their consistent and invariable presence across different domains, we propose integrating shape awareness into the model training in the source domain. We first model the shape information with the signed distance field (SDF [22], $\in [-1, 1]$) which measures the distance between any pixel to the nearest object boundary and the

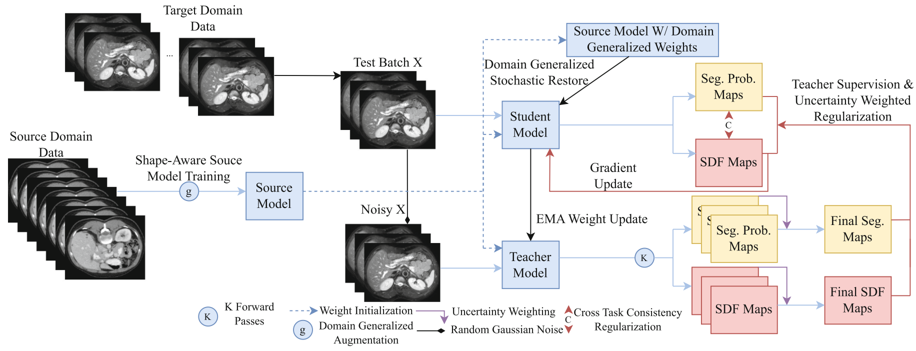

Fig. 2. Schematic of the proposed CTTA framework. The model (U-Net with EfficientNet-b2 backbone) is first trained on the source domain with shape-aware DG techniques for generalizable and adaptable baseline performance. Then, a multi-task uncertainty-weighted mean teacher setup performs target-domain adaptation. Small portions of the model are also reset to their initial shape-aware state at each step to counter catastrophic forgetting and improve the robustness of continual adaptation.

position of the pixel relative to the boundary: positive if outside, zero if on the boundary, and negative if inside. Then, an SDF head is appended to the source model to share features with the existing segmentation head to encode shape information into the model. Finally, the source training is performed with DG techniques to ensure reasonable performance even in extremely drifted target domains (such as T_1-weighted ↔ mDixon magnetic resonance images, Table 1), allowing for further adaptation to take place.

Specifically, we train the modified model on the source domain $(X^S, Y^S, Z^S) \in S$, where X^S denotes the input image, Y^S the manual annotations, and Z^S the ground-truth SDF calculated from Y^S using [22], by minimizing a multi-task loss $\frac{1}{N}\sum_{n=1}^{N} \ell_{\mathrm{seg}}(Y_n^S, \hat{Y}_n^S) + \ell_{\mathrm{sdf}}(Z_n^S, \hat{Z}_n^S)$. Here, ℓ_{seg} and ℓ_{sdf} represent loss functions used for optimizing the segmentation and SDF prediction tasks, respectively, N is the number of images in each batch, and $(\hat{Y}_n^S, \hat{Z}_n^S) = f(g(x))$ indicate the predicted segmentation probability and SDF maps produced by the source model f from the augmented input $g(x)$. We implement g with causality-inspired DG (CiDG) [19], a shallow randomly-weighted neural network that imposes domain-generalized shape-based feature learning through constant resampling of appearances of potentially correlated objects in the image.

Uncertainty and Shape-Aware Adaptation With Mean Teacher. The mean teacher network trains a student model and uses the exponential moving averages (EMA) of its weights to update an identical teacher model whose predictions further regularize the student model. Inspired by their rising popularity in semi-supervised studies [23], we use a mean teacher network to adapt *all parameters* of the trained shape-aware source model to the unlabeled target domain T. The overall architecture follows [21] except for the absence of the reconstruction task: both models predict SDF maps on top of segmentation labels, allowing for the utilization of shape information, and uncertainties are estimated from the teacher's outputs, avoiding misleading supervision during the adaptation phase.

Specifically, both models are initialized with the weights of the source model. Then, at each time step t, the student model first predicts segmentation probability maps $\tilde{Y}_t^T$ and SDF predictions $\tilde{Z}_t^T$ for the current test data x_t^T. Next, the teacher model performs K forward passes, producing K segmentation probability maps $\{\hat{Y}_{tk}^T\}_{k=1}^K$ and SDF predictions $\{\hat{Z}_{tk}^T\}_{k=1}^K$ from a set of noisy input images constructed by adding K random Gaussian noise vectors to x_t^T.

The final segmentation map of the teacher model at time step t is obtained by aggregating all K segmentation probability maps through their uncertainties. The pixel-wise uncertainty of each of the K segmentation probability maps is measured as the entropy $U_{tk} = -\sum_{c \in C} \hat{Y}_{tkc}^T \log_C \hat{Y}_{tkc}^T$, where the log function has a base of C, the number of segmentation classes. Next, the confidence map of kth probability map is calculated as $1 - U_{tk}$, as higher values in $U_{tk} \in [0, 1]$ denote areas with higher uncertainties. Then, all confidence maps are stacked in the first dimension where we apply the softmax function, i.e., $\{W_{tk}\}_{k=1}^K = \text{softmax}(\{1 - U_{tk}\}_{k=1}^K)$, to normalize the confidence value to $[0, 1]$. Lastly, the final segmentation probability map is constructed as a confidence-weighted combination of all K intermediate probability maps as $\hat{Y}_t^T = \sum_{k=1}^K W_{tk} \odot \hat{Y}_{tk}^T$. The entropy of the final segmentation represents its uncertainty $U_{\text{seg}} = -\sum_{c \in C} \hat{Y}_{tc}^T \log_C \hat{Y}_{tc}^T$.

Entropy cannot be calculated on real-valued outputs such as SDF maps. As such, the final SDF prediction is obtained by averaging all K SDF maps, i.e., $\hat{Z}_t^T = \frac{1}{K} \sum_{k=1}^K \hat{Z}_{tk}^T$, and we follow [24] to estimate the uncertainty using the variance $U_{\text{sdf}} = \sum_{k=1}^K (\hat{Z}_{tk}^T - \hat{Z}_t^T)^2$.

The student model is therefore guided by the teacher model by minimizing four loss terms:

$$\ell_t = \frac{1}{N} \sum_{n=1}^{N} \ell_{\text{seg}} \left(\tilde{Y}_n^T, \bar{Y}_n^T \right) + \ell_{\text{sdf}} \left(\tilde{Z}_n^T, \hat{Z}_n^T \right) + \ell_{\text{seg}}^{\text{con}} \left(\tilde{Y}_n^T, \bar{Y}_n^T \right) + \ell_{\text{sdf}}^{\text{con}} \left(\tilde{Z}_n^T, \hat{Z}_n^T \right) \quad (1)$$

where ℓ_{sdf} and ℓ_{seg} are the MSE and the Dice loss [26], $\bar{Y}_n^T$ is the one-hot encoded pseudo-labels calculated from $\hat{Y}_n^T$ with the argmax function, and N denotes the number of images in each test batch. $\ell_{\text{seg}}^{\text{con}} = \exp(-U_{\text{seg}}) \odot \|\tilde{Y}_n^T - \bar{Y}_n^T\|^2$ and $\ell_{\text{sdf}}^{\text{con}} = \exp(-U_{\text{sdf}}) \odot \|\tilde{Z}_n^T - \hat{Z}_n^T\|^2$ also penalize inconsistencies between the student and teacher models, but are weighted by the calculated uncertainty maps to encourage learning of confident predictions from the teacher model. The student model also performs self-regularization comprising two loss terms:

$$\ell_s = \frac{1}{N} \sum_{n=1}^{N} \left\| \tilde{Y}_n^T - \sigma \left(\kappa \cdot \tilde{Z}_n^T \right) \right\|^2 + \ell_e \left(\tilde{Y}_n^T \right) \quad (2)$$

where σ is the sigmoid function and κ is a multiplying factor approximating the inverse transformation from segmentation labels to SDF maps. The first loss term converts SDF maps into approximations of their corresponding segmentation labels and enforces a cross-task consistency [25], and the second term $\ell_e = -\sum_c \tilde{Y}_c^T \log \tilde{Y}_c^T$ reduces the entropy in the predicted segmentation maps. The final objective function is therefore formulated as a weighted sum as $\ell = \ell_t + \alpha \ell_s$.

Domain Generalized Stochastic Restore. Continual and unsupervised model adaptation to T would likely result in performance degradation due to accumulations of errors, leading to catastrophic forgetting of earlier samples. Therefore, we combine DG source training and a stochastic weight restoration mechanism [3] to reset small portions of the model to its initial domain-generalized weights, stopping the model from 'diving too deep' into specific target data while providing a decent baseline performance for the model to roll back.

Let W_{t+1} denote the weights of a trainable conv layer after the gradient update at time step t. A small portion of W_{t+1} is reset to its initial weights as $W_{t+1} = M \odot W_0 + (1 - M) \odot W_{t+1}$, where $M \sim \text{Bernoulli}(p)$ is a binary mask tensor, and W_0 denotes the initial domain-generalized weights of the conv layer.

3 Experiments

Setup. We implemented our method with PyTorch 1.10.0 and trained it on one Nvidia Tesla V100 GPU. We evaluated our method and other benchmarking methods on three cross-domain datasets with varying degrees of domain shifts: (1) cross-site binary prostate segmentation from T_2-weighted MRI scans collected from six different sites (12–30 scans/site) [29–31], (2) cross-site and cross-modality multi-class (liver, left and right kidneys, and spleen) abdominal segmentation between 30 CT and 20 MRI T_2-SPIR scans [32,33], and (3) same-site cross-modality muscle segmentation of 13 lower-leg muscles and bones between 30 MRI T_1 and 30 mDixon scans [34]. All scans were collected from healthy and diseased individuals and normalized to zero mean and unit variance before being reformatted to 2D. The prostate and abdominal scans were resized to 192×192 pixels while the muscle scans were spatially resized to 128×128 pixels. Lastly, a window of $[-275, 125]$ in Houndsfield units was applied to CT scans and the top 0.5% of the histogram of MRI scans were clipped as per [3].

We treated each site as the source domain and adapted to all other sites in the first experiment. For other experiments, we first performed adaptation from modality A to B, then from B to A. All experiments were performed in an online manner: each test scan arrived randomly and was broken down into multiple batches if needed. The model adapted itself to each batch before making a prediction. U-Net with an EfficientNet-b2 backbone was used as the source model for all our experiments. The Adam optimizer [35] was used with a learning rate of 0.001 and a batch size of 32. α was set to 1, κ to -1500, and p to 0.01. The model was empirically updated for two steps per test batch for prostate and muscle segmentation and 10 steps for abdominal segmentation. In addition, we calculated the final performance of each model by using each model to re-predict the segmentation labels of all test samples *after* the adaptation was completed. We then compared the final performance of each model against their running

Table 1. Quantitative evaluation of all methods w/ CiDG-trained source model. Results are shown as Dice/ASSD. The second row shows source/target domains. Source, general, medical, and (our) ablated methods are placed into their respective groups. † denotes statistical significance with our method ($p < 0.05$ w/Wilcoxon signed-rank test). Running performance shown. Best results in bold.

Prostate							Abdomen		Muscles	
	A/Rest	B/Rest	C/Rest	D/Rest	E/Rest	F/Rest	CT/T_2	T_2/CT	T_1/mDixon	mDixon/T_1
ERM	62/4.7 †	77/3.2 †	67/5.3 †	67/4.4 †	52/7.4 †	54/7.4 †	75/6.2 †	70/6.6 †	4/NaN †	7/NaN †
CiDG (also source)	72/3.9 †	**81**/2.8	76/3.6 †	73/3.6	68/4.8 †	74/4.4 †	83/5.1 †	81/5.6 †	72/2.1 †	73/3.0 †
BN Stats	73/4.3 †	76/6.0 †	70/7.4 †	71/4.3 †	65/5.9 †	72/6.0 †	79/5.7 †	78/7.9 †	62/6.4 †	71/6.4 †
Tent	73/3.8 †	80/5.8 †	74/7.6 †	71/3.8 †	68/5.0 †	73/5.8 †	80/7.0 †	78/11.5 †	69/6.5 †	75/6.2 †
CoTTA	74/3.7 †	**81**/4.1 †	77/6.8 †	72/4.3 †	69/4.8 †	74/5.2 †	84/7.8 †	80/6.1 †	72/4.8 †	77/6.4 †
DLTTA + Tent	73/4.1 †	81/5.4 †	74/8.0 †	72/4.1 †	69/5.2 †	73/5.5 †	84/7.8 †	80/7.6 †	72/8.6 †	72/8.9 †
DLTTA + CoTTA	72/4.0 †	80/5.5 †	75/6.9 †	72/4.3 †	67/5.8 †	73/5.8 †	84/8.2 †	81/7.8 †	70/6.3 †	77/6.2 †
ATTA	71/4.9 †	79/5.4 †	72/7.2 †	69/7.1 †	67/5.7 †	70/6.1 †	77/4.7 †	74/6.9 †	65/7.1 †	70/7.5 †
DLTTA + ATTA	71/4.4 †	78/5.1 †	74/7.0 †	70/7.3 †	67/5.2 †	72/6.3 †	77/4.4 †	75/5.6 †	69/6.7 †	71/7.1 †
Ours (w/o SDF)	78/3.5 †	80/3.9 †	78/6.7 †	72/4.1 †	70/4.4 †	75/4.9 †	85/7.1 †	82/6.3 †	76/3.9 †	79/3.1 †
Ours (w/o Uncertainties)	78/2.6	80/3.1 †	78/3.0 †	73/3.9 †	73/4.0	77/3.7 †	84/5.5 †	83/4.3 †	76/2.0 †	79/2.0 †
Ours (w/o DG restore)	77/2.7 †	**81**/2.9	77/3.2 †	72/4.0 †	72/4.2 †	76/3.9 †	85/4.2 †	84/4.1	75/2.5 †	78/1.8 †
Ours	**79/2.3**	**81/2.6**	**79/2.7**	**74/3.5**	**73/3.8**	**78/3.4**	**87/4.0**	**84/3.9**	**78/1.4**	**80/1.5**
Improv. over source	7/1.6	0/0.2	3/0.9	1/0.1	5/1.0	4/1.0	4/1.1	3/1.7	6/0.7	7/1.5

performance to evaluate their ability for continual adaptation. The performance was quantitatively evaluated by their volume-wise Dice Similarity Coefficient (Dice, in %) and Average Symmetric Surface Distance (ASSD, in mm).

Results. We compared our method against several state-of-the-art general and medical TTA and CTTA methods that require no additional clinical or anatomical information about either domain. General methods include BN Stats [27], Tent [16], and CoTTA [3], and medical methods involve the combination of ATTA [14] and DLTTA [28]. DLTTA was also combined with Tent and CoTTA for a more comprehensive comparison.

The proposed method substantially outperformed other methods on all three tasks and could consistently improve the CiDG-trained source model even in scenarios where other peer methods could not (Table 1). Surprisingly, the CiDG-trained source model outperformed all TTA methods except ours in numerous experiments with its decent performance. We attribute this to the fact that most TTA methods rely on (1) image/feature translation and reconstruction or (2) BN statistics re-estimation. However, DG methods often employ extensive augmentations, which may continuously change the contrast of the source data to allow domain-invariant feature learning. A constantly changing source domain may impede methods such as ATTA performing image or feature-level translation or reconstruction at the adaptation phase. Furthermore, we hypothesize that the running BN statistics of DG-trained models help to stabilize domain-invariant feature extraction at test time. As such, discarding and re-estimating them from test data, as was done by Tent, may be detrimental to the target-domain performance. To test our hypothesis, we disabled the BN statistics re-estimation in Tent and had a 3% improvement of Dice and 0.4 mm improvement on ASSD in the abdominal segmentation task. DLTTA consistently improved

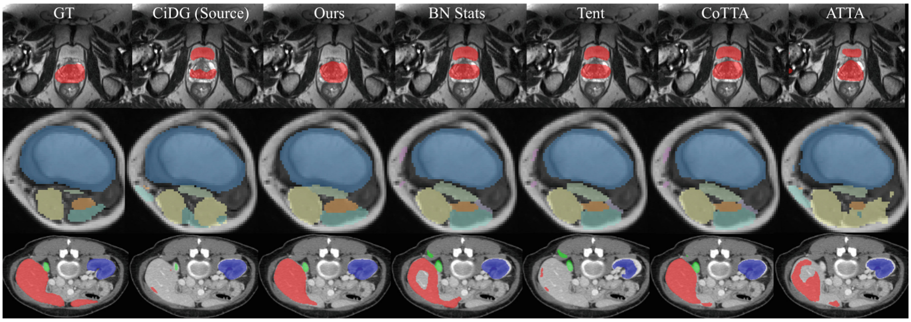

Fig. 3. Qualitative evaluation of selected benchmarked methods on the task of cross-site MRI T_2 prostate segmentation (top), mDixon → MRI T_1 muscle segmentation (middle), and MRI T_2 → CT abdominal segmentation (bottom). Results produced by methods augmented by DLTTA can be viewed in Supplementary Fig. 1.

Tent and ATTA through dynamic learning rates but failed to improve CoTTA at the same rate. CoTTA is a CTTA method highly relevant to ours, and its inconsistency in performance improvement suggests that geometric augmentations may be too strong for test-time learning and highlights the efficacy of the proposed uncertainty and shape-aware mean teacher setup. For adaptation, the uncertainty-aware module ensured only trustworthy predictions from the teacher model were used, and the shape-aware regularization further enhanced the target-domain performance by refining the smoothness of the predicted labels and ensuring the integrity of the anatomical structure of the predicted objects (Fig. 3). A brief ablation study demonstrated the effectiveness of each proposed component (see bottom of Table 1). Our framework also outperformed other methods by a larger margin on the prostate and abdominal segmentation tasks, where an ERM-trained source model was used for adaptation, further showcasing the generalizability of each proposed component (Supplementary Table 1).

The proposed model also demonstrated an equal or higher final performance (in comparison to its running performance) in all experiments, whereas many of its peers demonstrated the opposite (Supplementary Table 2). Equal final performance suggests that the model remembered earlier test data, and a higher final performance indicates its capability to utilize later test samples to improve its earlier performance. On the other hand, a lower final performance suggests that the model forgot about earlier test data and overfitted to later test data. The proposed DG stochastic restore prevented the model from drifting towards later test samples, and the teacher model reduced the likelihood of error accumulation through uncertainty estimation. Together they enabled reliable CTTA for medical images.

4 Conclusion

We proposed a generalizable framework for continual test-time adaption of medical images. Our approach first trains a model on the source domain with domain-invariant shape features before adapting it to the target domain with uncertainty-weighted pseudo-labels and SDF maps. Our method can work with ERM or DG-trained source models and outperformed its peers on three cross-site/cross-domain segmentation tasks without showing performance degradation as the adaptation progressed. Our framework can continuously adapt the source model to unknown test data online for the segmentation task, significantly reducing the cost and bias associated with manual labeling.

References

1. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
2. Bateson, M., Lombaert, H., Ben Ayed, I.: Test-time adaptation with shape moments for image segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13434, pp. 736–745 (2022). https://doi.org/10.1007/978-3-031-16440-8_70
3. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
4. Gong, T., Jeong, J., Kim, T., Kim, Y., Shin, J., Lee, S.,: NOTE: robust continual test-time adaptation against temporal correlation. In: Conference on Neural Information Processing Systems (NeurIPS) (2022)
5. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Mach. Learn. **79**(1), 151–175 (2010)
6. Ouyang, C., Kamnitsas, K., Biffi, C., Duan, J., Rueckert, D.: Data efficient unsupervised domain adaptation for cross-modality image segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 669–677 (2019). https://doi.org/10.1007/978-3-030-32245-8_74
7. Yang, J., Dvornek, N.C., Zhang, F., Chapiro, J., Lin, M.D., Duncan, J.S.: Unsupervised domain adaptation via disentangled representations: application to cross-modality liver segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 255–263. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_29
8. Zeng, G., et al.: Semantic consistent unsupervised domain adaptation for cross-modality medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 201–210. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_19
9. Bateson, M., Kervadec, H., Dolz, J., Lombaert, H., Ayed, I.B.: Constrained domain adaptation for segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 326–334. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_37
10. Bateson, M., Kervadec, H., Dolz, J., Lombaert, H., Ben Ayed, I.: Source-relaxed domain adaptation for image segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 490–499. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_48

11. Krikamol, M., Balduzzi, D., Schölkopf, B.: Domain generalization via invariant feature representation. In: International Conference on Machine Learning (ICML) (2013)
12. Yeh, H.-W., Yang, B., Yuen, P.C., Harada, T.: SoFA: source-data-free feature alignment for unsupervised domain adaptation. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WCAV) (2021)
13. 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
14. He, Y., Carass, A., Zuo, L., Dewey, B.E., Prince, J.L.: Autoencoder based self-supervised test-time adaptation for medical image analysis. Med. Image Anal. **72**, 102136 (2021)
15. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
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 (ICLR) (2021)
17. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
18. Xu, Z., Liu, D., Yang, J., Raffel, C., Niethammer, M.: Robust and generalizable visual representation learning via random convolutions. In: International Conference on Learning Representations (ICLR) (2021)
19. Ouyang, C., et al.: Causality-inspired single-source domain generalization for medical image segmentation. IEEE Trans. Med. Imaging **42**, 1095–1106 (2023)
20. Kim, T., Chai, J.: Pre-processing method to improve cross-domain fault diagnosis for bearing. Sensors **21**, 4970 (2021)
21. Wang, K., et al.: Tripled-uncertainty guided mean teacher network for semi-supervised medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 450–460 (2021). https://doi.org/10.1007/978-3-030-87196-3_42
22. Xue, Y., et al.: Shape-aware organ segmentation by predicting signed distance maps. In: AAAI Conference on Artificial Intelligence (AAAI) (2020)
23. Yang, X., Song, Z., King, I., Xu, Z.: A survey on deep semi-supervised learning. IEEE Trans. Knowl. Data Eng. **35**, 8934–8954 (2022)
24. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? In: Conference on Neural Information Processing Systems (NeurIPS), pp. 5574–5584 (2017)
25. Luo, X., Chen, J., Song, T., Wang, G.: Semi-supervised medical image segmentation through dual-task consistency. In: AAAI Conference on Artificial Intelligence (AAAI) (2021)
26. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
27. Schneider, S., Rusak, E., Eck, L., Bringmann, O., Brendel, W., Bethge, M.: Improving robustness against common corruptions by covariate shift adaptation. In: Conference on Neural Information Processing Systems (NeurIPS), pp. 11539–11551 (2020)
28. Yang, H., et al.: DLTTA: dynamic learning rate for test-time adaptation on cross-domain medical images. IEEE Trans. Med. Imaging **41**, 3575–3586 (2023)

29. Liu, Q., Dou, Q., Heng, P.-A.: Shape-aware meta-learning for generalizing prostate MRI segmentation to unseen domains. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12262, pp. 475–485. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59713-9_46
30. Bloch, N., et al.: NCI-ISBI 2013 challenge: automated segmentation of prostate structures. The Cancer Imaging Archive, vol. 370 (2015)
31. Lemaître, G., Martí, R., Freixenet, J., Vilanova, J.C., Walker, P.M., Meriaudeau, F.: Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: a review. CBM **60**, 8–31 (2015)
32. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: MICCAI multi-atlas labeling beyond the cranial vault-workshop and challenge. In: Proceedings of MICCAI Multi-Atlas Labeling Beyond Cranial Vault-Workshop Challenge (2015)
33. Kavur, A.E., et al.: CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation. Med. Image Anal. **69**, 101950 (2021)
34. Zhu, J., et al.: Deep learning methods for automatic segmentation of lower leg muscles and bones from MRI scans of children with and without cerebral palsy. NMR Biomed. **34**, e4609 (2021)
35. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv:1412.6980 (2014)