Source-Free Domain Adaptation for Medical Image Segmentation via Selectively Updated Mean Teacher

Ziqi Wen, Xinru Zhang, and Chuyang Ye(✉)

School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing, China
chuyang.ye@bit.edu.cn

Abstract. Automated medical image segmentation is valuable for disease diagnosis and prognosis, and it has achieved promising performance with deep neural networks. However, a segmentation model trained on a source dataset may not perform well on a different target dataset when the distribution shift or even modality alteration exists between them. To address this problem, domain adaptation techniques can be applied to train the model with the help of the unannotated target dataset. Often when the target data is available, only a segmentation model trained on the source dataset is provided without the source data, and in this case, *source-free domain adaptation* (SFDA) is needed. In this work, we focus on the development of SFDA techniques for medical image segmentation, where the given source model is updated based on the target data. Since no annotations are available for the target dataset, we propose to leverage the consistency of predictions on the target data when different perturbations are made, and adopt the mean teacher framework that can effectively exploit the consistency. Moreover, we assume that the update of the entire model in vanilla mean teacher is suboptimal because when no annotated data is available the knowledge learned for segmentation in the source model can be easily forgotten. Therefore, we propose *selectively updated mean teacher* (SUMT), which seeks to adapt the source model parameters that are sensitive to domain variance and retain the parameters that are invariant to domains. In SUMT, we develop a progressive layer update strategy with channel-wise weight restoration that alleviates forgetting. To evaluate the proposed method, experiments were performed on three datasets, where the source and target data used different modalities for segmentation, or their images were acquired at different sites. The results show that our method improves the segmentation accuracy compared with other SFDA approaches.

Keywords: Source-free domain adaptation · medical image segmentation · selectively updated mean teacher

A. Frangi et al. (Eds.): IPMI 2023, LNCS 13939, pp. 225–236, 2023.
https://doi.org/10.1007/978-3-031-34048-2_18

1 Introduction

Automated segmentation of medical images can provide a valuable tool for the diagnosis and prognosis of disease and enhance our understanding of disease and treatment planning [13,15]. The use of *deep neural networks* (DNNs) has allowed remarkable improvement of the segmentation accuracy [9]. However, the segmentation model trained on a source dataset may not generalize well to a target dataset that is acquired at a different site on a different scanner due to the domain shift caused by inter-scanner variability [1]. Moreover, the target dataset may even use a different modality for segmenting the same anatomical structure or lesions [2,13], which further increases the difficulty of generalizing the trained model to the target data. In these cases, the segmentation quality for the target dataset can be severely degraded, and it is desirable to develop segmentation approaches that adapt well to different target datasets.

To address the generalization problem, domain adaptation techniques are developed, which exploit both the annotated source training data and unannotated target data [1,2,6,7]. For example, in [6] and [7] adversarial learning is applied to align the features of the source and target domains. In [2], a synergistic fusion of adaptations from both image and feature perspectives is proposed when the source and target domains use different image modalities. AdaEnt [1] uses an additional class ratio predictor for domain adaptation by assuming that the class ratio is invariant between the source and target domains. These methods are shown to allow better adaptation of a DNN-based model to target datasets.

The domain adaptation methods described above assume access to both the annotated source dataset and the unannotated target dataset during model training. However, in real-world scenarios, when the segmentation model is trained with the source data the target data may not be available due to privacy concerns or even not be acquired yet; and it is also not guaranteed that the source dataset can be shared with the target dataset for model retraining. In these cases, *source-free domain adaptation* (SFDA) should be considered, where only the model trained on the source data is provided without the source data, and this given source model is updated based on the unannotated target data. For example, for classification problems, SHOT [12] is developed to align the hypothesis of the source model to the target domain with entropy minimization and diversity regularization, but this method cannot be directly adapted to segmentation; TENT [24] updates the batch statistics and affine parameters in the batch normalization layers of the source model via entropy minimization on the unlabeled target data. More specifically for medical image segmentation, OSUDA is proposed in [13] based on batch normalization statistics under the assumption that scaling and shifting operations in batches are domain shareable. OSUDA explicitly enforces a channel-wise optimization objective, where the domain-specific batch mean and variance are updated incrementally. However, only adapting the batch normalization layers of the source model is generally insufficient for optimal performance. Therefore, the development of SFDA methods for medical image segmentation is still an open problem.

Ideally, SFDA should adapt the domain-specific parameters in the source model according to the target data and retain the domain-invariant parameters. Since no annotations are available for the target dataset, to update the source model, we propose to leverage the consistency of predictions on the target data when different perturbations are made. This idea is common in *semi-supervised learning* (SSL) [5,14,23], where the *mean teacher* (MT) framework [23] has been mostly used for the purpose. However, unlike SSL, in SFDA the source model is updated purely based on the consistency information without any annotated data. This can easily lead to knowledge forgetting, which impairs the domain-invariant knowledge in the source model that is necessary for accurate segmentation.

To avoid this issue, we propose a *selectively updated mean teacher* (SUMT) framework for SFDA-based medical image segmentation. In SUMT, a student model and a teacher model are both initialized by the source model. Since state-of-the-art DNNs for medical image segmentation generally use an encoder-decoder architecture [3,18,21], we also assume that the segmentation model has both an encoder and decoder. First, as earlier layers are more likely to be domain-specific [12], instead of updating all layers in the teacher model, only its encoding layers are updated with *exponential moving average* (EMA) [23] based on the student model. Then, a *channel-wise weight restoration* (CWR) strategy is developed to preserve the domain-invariant knowledge, where the network weights of the encoder of the teacher model that are likely to be domain-invariant are identified, and the identified weights are restored to their initial values. Next, the whole teacher model is updated based on the student model, and CWR is applied to the decoding layers to further alleviate forgetting. Finally, the teacher model is updated again and used for segmentation. To evaluate the proposed method, experiments were performed on three datasets, where the source and target data used different modalities for segmentation or used images acquired at different sites. The results show that our method improves the segmentation accuracy compared with other SFDA approaches.

2 Method

2.1 Problem Formulation and Method Overview

Suppose we are given a DNN-based segmentation model $\mathcal{M}$ with an encoder-decoder architecture trained on a source dataset, but the source training data is not accessible. We seek to perform segmentation on a set $\mathcal{X}$ of N images from a different target dataset, where the i-th target image is denoted by x_i. The target images are acquired differently from the source images, e.g., with different intensity distributions or even different modalities. Due to domain shift, direct application of $\mathcal{M}$ to $\mathcal{X}$ leads to suboptimal performance [12,13,24]. Therefore, the aim of this work is to adapt $\mathcal{M}$ based on $\mathcal{X}$ so that the segmentation performance is improved, which is an SFDA problem.

Since no annotated data is available for the target dataset, we choose to adapt the source model based on the prediction consistency when the target

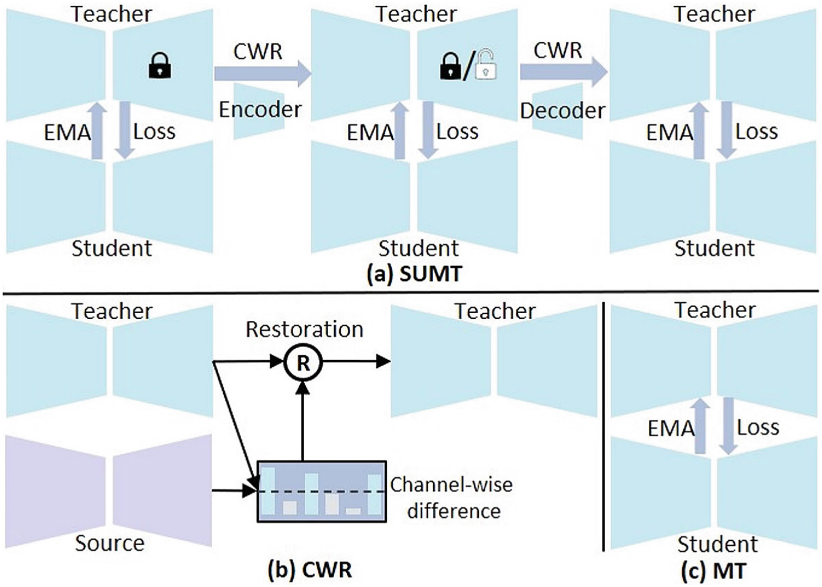

Fig. 1. Method overview: (a) the complete SUMT framework, (b) the CWR strategy in SUMT, and (c) the standard MT for comparison.

data is perturbed, so that the model can accommodate the target domain. The MT framework [23] has been shown to effectively exploit the prediction consistency in the SSL setting. However, in SFDA the knowledge necessary for image segmentation can be easily forgotten if the model is updated solely based on the prediction consistency. For example, the model can simply resort to a degenerate solution that produces the same result for all inputs. Therefore, we propose SUMT that improves upon the MT framework for the SFDA setting. An overview of SUMT is shown in Fig. 1, where the CWR strategy in SUMT is also illustrated and the standard MT is described for comparison. The detailed design of SUMT is presented below.

2.2 Selectively Updated Mean Teacher

To effectively leverage the data from the target domain, SUMT seeks to adapt the source model parameters that are sensitive to domain variance and retain the parameters that are invariant to domains. Like standard MT, in SUMT a teacher model $\mathcal{M}^{\mathrm{t}}$ and a student model $\mathcal{M}^{\mathrm{s}}$ are constructed. $\mathcal{M}^{\mathrm{t}}$ and $\mathcal{M}^{\mathrm{s}}$ share the same network structure, and they are both initialized by the source model $\mathcal{M}$.

In standard MT, $\mathcal{M}^{\mathrm{t}}$ makes predictions on perturbed target images, which are considered pseudo-labels, and $\mathcal{M}^{\mathrm{s}}$ learns from the pseudo-labels based on differently perturbed target images to update the model parameters. Then, all model weights of $\mathcal{M}^{\mathrm{t}}$ are in turn updated based on $\mathcal{M}^{\mathrm{s}}$ with EMA. However, the joint update of all weights can be problematic for SFDA as it may cause forgetting of domain-invariant knowledge for segmentation due to the lack of annotated data. To address this problem, in SUMT we propose to selectively update the model weights of $\mathcal{M}^{\mathrm{t}}$, where the following steps are applied.

Since the earlier layers are more likely to be domain-specific [12], we propose to first update the encoding layers of $\mathcal{M}^{\mathrm{t}}$ while fixing its decoding layers when $\mathcal{M}^{\mathrm{t}}$ is updated based on $\mathcal{M}^{\mathrm{s}}$ with EMA. Formally, we denote the encoders/decoders of $\mathcal{M}$, $\mathcal{M}^{\mathrm{t}}$, and $\mathcal{M}^{\mathrm{s}}$ by E/D, $E_{\mathrm{t}}/D_{\mathrm{t}}$, and $E_{\mathrm{s}}/D_{\mathrm{s}}$, respectively. At the t-th iteration of the update, the student model $\mathcal{M}^{\mathrm{s}}$ is updated by minimizing a consistency loss based on the target data. Specifically, suppose the prediction given by $\mathcal{M}^{\mathrm{s}}$ for x_i is c_i, and the corresponding pseudo-label given by $\mathcal{M}^{\mathrm{t}}$ is d_i.[1] The consistency loss $\mathcal{L}_{\mathrm{c}}$ for updating $\mathcal{M}^{\mathrm{s}}$ with fixed $\mathcal{M}^{\mathrm{t}}$ is defined as

$$\mathcal{L}_{\mathrm{c}} = \sum_{i=1}^{N} \left(\mathcal{L}_{\mathrm{ce}}(c_i, d_i) + \mathcal{L}_{\mathrm{Dice}}(c_i, d_i)\right), \tag{1}$$

where $\mathcal{L}_{ce}$ and $\mathcal{L}_{\mathrm{Dice}}$ are the cross-entropy loss and Dice loss [20], respectively. Then, the teacher model is updated as

$$E_{\mathrm{t}} \leftarrow E_{\mathrm{t}} \cdot \sigma + E_{\mathrm{s}} \cdot (1 - \sigma) \quad \text{and} \quad D_{\mathrm{t}} \leftarrow D, \tag{2}$$

where σ is the EMA decay rate to be specified. This partial update of $\mathcal{M}^{\mathrm{t}}$ reduces the risk of forgetting high-level semantic knowledge in the decoding layers while adapting the extraction of low-level features in the encoding layers.

In the partial teacher update above, it is still possible that domain-invariant model parameters in the encoder are inappropriately updated. To address this problem, we further propose the CWR strategy which explicitly restores the knowledge from the source model that may need to be retained. Specifically, suppose the model weights of $\mathcal{M}^{\mathrm{t}}$ at the l-th layer associated with the k-th channel are represented as a set $\mathcal{W}_{l,k}^{\mathrm{t}} = \{w_{l,k,p}^{\mathrm{t}}\}_{p=1}^{P_{l,k}}$, where $P_{l,k}$ is the number of these weights and $w_{l,k,p}^{\mathrm{t}}$ is the p-th weight. The amount of the update of $\mathcal{W}_{l,k}^{\mathrm{t}}$ can indicate whether the weights are associated with domain-specific or domain-invariant features. A greater update amount indicates that the channel is likely to focus on the domain-specific feature and contribute to domain adaptation, whereas a smaller amount indicates that the channel tends to extract domain-invariant features and probably should not be forgotten. To measure the amount of the weight update, we compute the difference $d_{l,k}$ between $\mathcal{W}_{l,k}^{\mathrm{t}}$ and the corresponding weights $\mathcal{W}_{l,k} = \{w_{l,k,p}\}_{p=1}^{P_{l,k}}$ in the source model $\mathcal{M}$ as

$$d_{l,k} = \sum_{p=1}^{P_{l,k}} \left| w_{l,k,p}^{\mathrm{t}} - w_{l,k,p} \right|. \tag{3}$$

Based on $d_{l,k}$, we restore the bottom q_l (percentage) of the weights for the l-th layer of $\mathcal{M}^{\mathrm{t}}$ as

$$\mathcal{W}_{l,k}^{\mathrm{t}} \leftarrow \begin{cases} \mathcal{W}_{l,k}, & d_{l,k} \leq H(\mathcal{D}_l, q_l) \\ \mathcal{W}_{l,k}^{\mathrm{t}}, & d_{l,k} > H(\mathcal{D}_l, q_l) \end{cases}, \tag{4}$$

[1] Noise perturbation and random flips are applied before the teacher or student prediction as in [16].

where $\mathcal{D}_l$ represents the set of all $d_{l,k}$'s at the l-th layer, and $H(\mathcal{D}_l, q_l)$ sorts these $d_{l,k}$'s in ascending order and returns the value that is ranked q_l. Note that here since D_{t} is the same as D, no restoration is needed for the decoding layers.

With the restored encoder, we assume that the earlier layers are better adapted to the target data, and the complete teacher model $\mathcal{M}^{\mathrm{t}}$ including its decoder can now be updated with standard MT, where $\mathcal{M}^{\mathrm{s}}$ is reinitialized by $\mathcal{M}^{\mathrm{t}}$. Note that to avoid incorrectly restored weights, a warmup stage that again only updates E_{t} is inserted before updating the complete teacher model. During the update of the complete teacher model, at the t-th iteration, after $\mathcal{M}^{\mathrm{s}}$ is updated based on $\mathcal{L}_{\mathrm{c}}$, the teacher model is updated with EMA as

$$E_{\mathrm{t}} \leftarrow E_{\mathrm{t}} \cdot \sigma + E_{\mathrm{s}} \cdot (1-\sigma) \quad \text{and} \quad D_{\mathrm{t}} \leftarrow D_{\mathrm{t}} \cdot \sigma + D_{\mathrm{s}} \cdot (1-\sigma). \tag{5}$$

Finally, to further avoid knowledge forgetting in the decoder, CWR is applied to the decoding layers of $\mathcal{M}^{\mathrm{t}}$ based on Eq. (4), and the teacher model is then updated again with standard MT using Eq. (5). After convergence, the teacher model is used as the final segmentation model.

2.3 Implementation Details

We focus on 3D segmentation and use the 3D U-Net [3] implemented in SS4L [16] as the backbone segmentation network, which is a popular choice for semi-supervised medical image segmentation [17,18]. Note that variants of U-Net may also be used and integrated with the proposed method, but it is observed that the performance of these variants is usually on par with the original U-Net [8].

The major hyperparameters in the proposed method are the restoration percentages $\{q_l\}_{l=1}^{L}$, where L is the total number of layers and it is equal to ten for the selected 3D U-Net. We set $q_l = 0.1 * l$ because earlier layers tend to be domain-specific and their restoration is less needed. We set the EMA decay rate $\sigma = 0.999$ according to [16]. The other training configurations, such as the optimizer, learning rate, etc., are set to the default specification in [16].

3 Results

3.1 Data Description and Experimental Settings

To evaluate the proposed method, we performed experiments on three datasets. Their details and experimental settings are given below.

BraTS 2018. The first dataset is the publicly available BraTS 2018 dataset [19], and it was used in this work for whole brain tumor segmentation. The BraTS 2018 dataset contains 285 subjects with four modalities of magnetic resonance imaging, including T1w, T2w, T1ce, and FLAIR images. These images are aligned and have the same voxel size of 1 mm isotropic. For each subject, voxel-wise labels for the enhancing tumor, peritumoral edema, and necrotic and non-enhancing tumor core are given, and they were combined to provide the annotation of the whole tumor. We randomly split the dataset into a training set of 200 subjects and a test set of 85 subjects. To investigate the performance of cross-domain segmentation where the training and test sets use different image modalities for segmentation, we considered the following settings: 1) FLAIR for training and T2w for testing, 2) T2w for training and FLAIR for testing, 3) T1w for training and T1ce for testing, and 4) T1ce for training and T1w for testing.

INBT. The second dataset is an in-house dataset for whole brain tumor segmentation, which is referred to as INBT for convenience. The dataset includes 67 annotated FLAIR images acquired on multiple scanners, and they have been skull-stripped with BET [22]. The voxel size of these images ranges from $0.875\,\text{mm} \times 0.875\,\text{mm} \times 2\,\text{mm}$ to $2\,\text{mm} \times 2\,\text{mm} \times 5\,\text{mm}$. We used INBT to investigate the segmentation performance when the same modality was used for segmentation but the training and test images were acquired on different scanners. Specifically, the FLAIR images of the training subjects in the BraTS 2018 dataset were used for model training, and all subjects in INBT were used as the test set.

MSSEG. The last dataset is the MSSEG dataset [4] for segmenting multiple sclerosis lesions. The dataset contains multimodal images of 15 subjects acquired on three scanners (five subjects for each scanner), including Philips Ingenia 3T (PI3T), Siemens Aera 1.5T (SA1.5T), and Siemens Verio 3T (SV3T). These images have been preprocessed with skull-stripping and co-registration [4]. The resolution of the preprocessed images ranges from 0.5 mm to 1.25 mm in each dimension for different subjects. The annotation of multiple sclerosis lesions was performed for each subject by seven independent clinical experts, and their consensus was used as the final annotation. For demonstration, we used the FLAIR modality for segmentation. The MSSEG dataset was used to investigate the segmentation performance when the training and test images were of the same modality but acquired on different scanners. We considered two experimental settings for MSSEG. First, the training images and test images were acquired on scanners of different vendors, where the images acquired on SA1.5T and SV3T were used for training, and the images acquired on PI3T were used for testing. Second, the training images and test images were acquired with different magnetic fields, where the images acquired on PI3T and SV3T were used for training, and the images acquired on SA1.5T were used for testing.

3.2 Evaluation of Segmentation Accuracy

SUMT was applied to the three datasets separately, and it was compared with four other SFDA methods. The first one is *pseudo-labeling* (PL) [10] that generates pseudo-labels on the target data based on the source model and optimizes the segmentation model with the pseudo-labels. The second one is AdaBN [11] that only updates batch normalization statistics based on the target test data during inference. The third one is TENT [24] that updates both batch statistics and affine parameters in the batch normalization layers via entropy minimization on the target data. The fourth one is OSUDA [13] that is designed for medical image segmentation with an adaptive update of batch-wise normalization statistics. Also, the standard MT framework was included for comparison. In addition, direct application of the source model to the target data without SFDA was considered for comparison, and it is referred to as the baseline. For reference, for the BraTS 2018 dataset that has a large number of training subjects, the *upper bound* (UB) performance that was obtained by training the segmentation model with the target modality of the training subjects was also given (e.g., the model was trained with the FLAIR images of the training subjects when segmentation was to be performed on the FLAIR images of the test subjects). The results on the three datasets are presented next individually.

BraTS 2018 for Cross-modality Segmentation. For qualitative evaluation, axial views of representative segmentation results on test scans from the BraTS 2018 dataset are shown in Fig. 2(a) for SUMT and each competing method, together with the image for segmentation and the expert annotation. The results are shown for the different settings of test image modalities. We can see that in these different cases SUMT produced segmentation results that better agree with the annotation than the competing methods.

Next, SUMT was quantitatively evaluated by computing the Dice coefficient between the segmentation results on the test set and expert annotation. The means and standard deviations of the Dice coefficients are summarized in Table 1. In all cases, SUMT outperforms the competing methods with higher Dice coefficients. In addition, with paired Student's t-tests we show that the difference between SUMT and the competing methods is statistically significant, and this is also indicated in Table 1.

Moreover, we investigated the individual benefit of the proposed CWR strategy and progressive layer update. To demonstrate the benefit of CWR, we integrated CWR with the standard MT, where all weights were jointly updated, restored according to Eq. (4), and then updated again. This procedure is referred to as MT-CWR. In addition, to show the necessity of our weight restoration design in Eq. (4), we modified MT-CWR by replacing Eq. (4) with stochastic restoration with the same ratio, and this procedure is referred to as MT-SR. To demonstrate the benefit of the proposed progressive layer update, we integrated it with the standard MT, which is equivalent to the application of the proposed method without CWR, and this procedure is referred to as MT-PLU. The means and standard deviations of the Dice coefficients for these cases are summarized in Table 1 as well. We can see that both MT-CWR and MT-PLU are better than MT but worse than SUMT, which confirms that these two individual contributions and their integration are all beneficial. Besides, MT-CWR is better than MT-SR, which shows the benefit of the proposed restoration design.

INBT for Cross-scanner Segmentation. Qualitative evaluation and quantitative evaluation of the results on INBT for cross-scanner segmentation are given in Fig. 2(b) and Table 2 (the part associated with INBT), respectively.

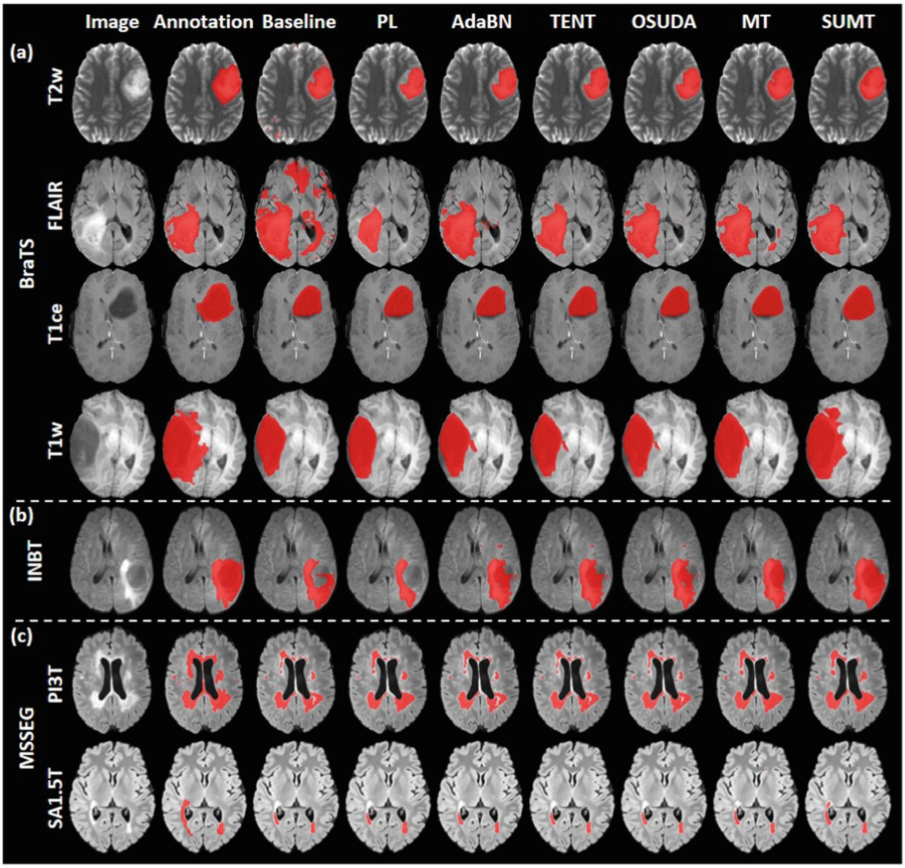

Fig. 2. Axial views of representative segmentation results (red) on test scans for (a) the BraTS 2018 dataset, (b) the INBT dataset, and (c) the MSSEG dataset. The images for segmentation and the expert annotation are also shown for reference. (Color figure online)

From Fig. 2(b), we can see that the segmentation result of SUMT better resembles the expert annotation than the competing methods. Table 2 indicates that SUMT has a higher Dice coefficient than the competing methods and its difference with the competing methods is significant with paired Student's *t*-tests in four out of six cases. Also, the results of MT-SR, MT-CWR, and MT-PLU are shown in Table 2. The observation for them is consistent with the results of BraTS 2018, where MT-CWR and MT-PLU are better than MT and worse than SUMT, and MT-CWR is better than MT-SR.

MSSEG for Cross-scanner Segmentation. Qualitative evaluation and quantitative evaluation of the segmentation results on the MSSEG dataset are given in Fig. 2(c) and Table 2 (the part associated with MSSEG), respectively,

Table 1. Means and standard deviations of the Dice coefficients (%) of the segmentation results on the test set for the BraTS 2018 dataset. Asterisks indicate that the difference between SUMT and the competing method is statistically significant (*: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$) using a paired Student's t-test. The best results are highlighted in bold.

Method	BraTS 2018			
	FLAIR→T2w	T2w→FLAIR	T1w→T1ce	T1ce→T1w
UB	85.0 ± 13.1	81.2 ± 14.8	74.0 ± 19.7	74.6 ± 18.7
Baseline	50.0 ± 30.1***	63.0 ± 26.2***	64.0 ± 20.0***	60.7 ± 28.5***
PL	43.7 ± 33.2***	63.3 ± 25.6***	58.2 ± 25.5***	46.5 ± 28.4***
AdaBN	48.1 ± 30.0***	69.1 ± 24.3***	56.0 ± 25.5***	65.9 ± 24.7**
TENT	48.5 ± 30.4***	72.8 ± 23.4**	56.6 ± 25.0***	64.2 ± 26.0***
OSUDA	47.8 ± 30.5***	72.0 ± 23.7***	55.8 ± 25.9***	64.3 ± 26.1***
MT	53.5 ± 28.6***	74.3 ± 22.7***	65.9 ± 19.8***	67.6 ± 24.8*
MT-SR	54.0 ± 28.1	74.0 ± 24.5	65.4 ± 22.2	66.5 ± 25.4
MT-CWR	54.9 ± 27.7	74.6 ± 23.2	66.4 ± 19.7	68.2 ± 23.1
MT-PLU	55.4 ± 27.0	74.8 ± 21.8	69.0 ± 20.6	70.3 ± 21.3
SUMT	**56.3 ± 26.7**	**76.4 ± 20.4**	**69.7 ± 20.0**	**70.5 ± 21.5**

Table 2. Means and standard deviations of the Dice coefficients (%) of the segmentation results on the test set for the INBT and MSSEG datasets. Asterisks indicate that the difference between SUMT and the competing method is statistically significant (*: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$) using a paired Student's t-test. The best results are highlighted in bold.

Method	INBT	MSSEG	
		PI3T	SA1.5T
Baseline	63.5 ± 30.5***	60.4 ± 9.6**	38.1 ± 6.5**
PL	70.1 ± 23.2***	57.5 ± 13.2*	46.2 ± 10.0
AdaBN	74.0 ± 22.5*	60.0 ± 9.4*	40.7 ± 9.6**
TENT	74.2 ± 22.2	61.3 ± 9.9**	40.3 ± 8.9**
OSUDA	74.3 ± 22.1	61.6 ± 10.7**	40.4 ± 8.8**
MT	73.3 ± 20.9***	66.3 ± 9.0*	45.6 ± 7.3*
MT-SR	72.7 ± 21.6	65.1 ± 8.7	44.6 ± 7.3
MT-CWR	74.0 ± 19.5	66.3 ± 9.2	45.7 ± 7.6
MT-PLU	75.9 ± 19.6	67.7 ± 8.5	48.0 ± 7.2
SUMT	**76.2 ± 18.3**	**68.1 ± 8.2**	**48.7 ± 7.7**

and the results of MT-SR, MT-CWR, and MT-PLU are shown in Table 2 as well. Like the results of BraTS 2018 and INBT, these results show that SUMT outperforms the competing methods and its difference with the competing methods is significant with paired Student's t-tests in most cases; also, the results of MT-SR, MT-CWR, and MT-PLU in Table 2 confirm the benefit of the integration of CWR and progressive layer update, as well as the weight restoration design.

4 Conclusion

We have proposed SUMT for SFDA-based medical image segmentation. In SUMT, we adapt the mean teacher framework by selectively updating the model parameters to better preserve domain-invariant knowledge. The model update is performed progressively with channel-wise weight restoration. Experimental results on cross-modality and cross-scanner segmentation tasks demonstrate that SUMT outperforms other SFDA methods.

Acknowledgement. This work is supported by the Fundamental Research Funds for the Central Universities.

References

1. 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
2. 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. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 01, pp. 865–872 (2019)
3. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
4. Commowick, O., et al.: Multiple sclerosis lesions segmentation from multiple experts: the MICCAI 2016 challenge dataset. Neuroimage **244**, 118589 (2021)
5. Cui, W., et al.: Semi-supervised brain lesion segmentation with an adapted mean teacher model. In: Chung, A.C.S., Gee, J.C., Yushkevich, P.A., Bao, S. (eds.) IPMI 2019. LNCS, vol. 11492, pp. 554–565. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20351-1_43
6. Ganin, Y., et al.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17**(1), 2096-2030 (2016)
7. Ghafoorian, M., et al.: Transfer learning for domain adaptation in MRI: application in brain lesion segmentation. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 516–524. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_59
8. Gut, D., Tabor, Z., Szymkowski, M., Rozynek, M., Kucybała, I., Wojciechowski, W.: Benchmarking of deep architectures for segmentation of medical images. IEEE Trans. Med. Imaging **41**(11), 3231–3241 (2022)

9. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
10. Lee, D.H.: Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. In: ICML Workshop on Challenges in Representation Learning (2013)
11. Li, Y., Wang, N., Shi, J., Hou, X., Liu, J.: Adaptive batch normalization for practical domain adaptation. Pattern Recogn. **80**, 109–117 (2018)
12. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. In: International Conference on Machine Learning, pp. 6028–6039 (2020)
13. Liu, X., Xing, F., Yang, C., El Fakhri, G., Woo, J.: Adapting off-the-shelf source segmenter for target medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 549–559. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_51
14. Liu, Y.C., Ma, C.Y., Kira, Z.: Unbiased teacher v2: semi-supervised object detection for anchor-free and anchor-based detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9819–9828 (2022)
15. Lu, Q., Ye, C.: Knowledge transfer for few-shot segmentation of novel white matter tracts. In: Feragen, A., Sommer, S., Schnabel, J., Nielsen, M. (eds.) IPMI 2021. LNCS, vol. 12729, pp. 216–227. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-78191-0_17
16. Luo, X.: SSL4MIS (2020). https://github.com/HiLab-git/SSL4MIS
17. Luo, X., Chen, J., Song, T., Wang, G.: Semi-supervised medical image segmentation through dual-task consistency. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 8801–8809 (2021)
18. Luo, X., et al.: Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 318–329. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_30
19. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2014)
20. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: International Conference on 3D Vision, pp. 565–571 (2016)
21. 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
22. Smith, S.M.: Fast robust automated brain extraction. Hum. Brain Mapp. **17**(3), 143–155 (2002)
23. Tarvainen, A., Valpola, H.: Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems, pp. 1195–1204 (2017)
24. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: TENT: fully test-time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726 (2020)