UPL-SFDA: Uncertainty-aware Pseudo Label Guided Source-Free Domain Adaptation for Medical Image Segmentation

Jianghao Wu, Guotai Wang, Ran Gu, Tao Lu, Yinan Chen, Wentao Zhu, Tom Vercauteren, Sébastien Ourselin, Shaoting Zhang

arXiv:2309.10244v1 [cs.CV] 19 Sep 2023

Abstract—Domain Adaptation (DA) is important for deep learning medical image segmentation models to deal with testing images from a new target domain. As the source-domain data are usually unavailable when a trained model is deployed at a new center, Source-Free Domain Adaptation (SFDA) is appealing for data and annotation-efficient adaptation to the target domain. However, existing SFDA methods have a limited performance due to lack of sufficient supervision with source-domain images unavailable and target-domain images unlabeled. We propose a novel Uncertainty-aware Pseudo Label guided (UPL) SFDA method for medical image segmentation. Specifically, we propose Target Domain Growing (TDG) to enhance the diversity of predictions in the target domain by duplicating the pre-trained model's prediction head multiple times with perturbations. The different predictions in these duplicated heads are used to obtain pseudo labels for unlabeled target-domain images and their uncertainty to identify reliable pseudo labels. We also propose a Twice Forward pass Supervision (TFS) strategy that uses reliable pseudo labels obtained in one forward pass to supervise predictions in the next forward pass. The adaptation is further regularized by a mean prediction-based entropy minimization term that encourages confident and consistent results in different prediction heads. UPL-SFDA was validated with a multi-site heart MRI segmentation dataset, a cross-modality fetal brain segmentation dataset, and a 3D fetal tissue segmentation dataset. It improved the average Dice by 5.54, 5.01 and 6.89 percentage points for the three tasks compared with the baseline, respectively, and outperformed several state-of-the-art SFDA methods.

Index Terms—Source-free domain adaptation, self-training, fetal MRI, heart MRI, entropy minimization.

Jianghao Wu, Guotai Wang, Ran Gu and Shaoting Zhang are with the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China. Guotai Wang and and Shaoting Zhang are also with Shanghai AI laboratory, Shanghai, 200030, China. (e-mail: guotai.wang@uestc.edu.cn, zhangshaoting@uestc.edu.cn)

Tao Lu is with the Department of Radiology, Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu, 610072, China

Yinan Chen is with SenseTime Research, Shanghai, 200233, China

Wentao Zhu is with Research Center for Healthcare Data Science, Zhejiang Laboratory, Hangzhou, 311100, China

Tom Vercauteren and Sébastien Ourselin are with the School of Biomedical Engineering & Imaging Sciences, King's College London, London, WC2R 2LS, UK.

I. INTRODUCTION

DEEP learning has achieved excellent performance in medical image segmentation tasks in recent years [1], [2]. Its current success is highly dependent on the assumption that training and testing images are from the same distribution. However, in practice, a model trained with images from one certain source domain may be used to deal with images in an unseen target domain with different image appearances, which is usually caused by different scanning devices, imaging protocols, patient groups or image qualities, etc. Failing to deal with the gap between the source and target domains will lead to a dramatic performance decrease [3]. As it is impossible to collect images from all the potential target domains during training, it is essential to make the model adapted to images in the unseen target domain after deployment.

Domain Adaptation (DA) that aims to solve the domain gap between training and testing data is attracting increasing attentions recently [4]. Though collecting a set of annotated images in the target domain to fine-tune the pre-trained model can make it adapted to the target domain, the annotations are expensive to obtain and usually unavailable in the target domain for model deployment. Therefore, many researchers have investigated Unsupervised Domain Adaptation (UDA) [4] that uses unannotated images in the target domain for adaptation. Most existing UDA methods require access to source-domain and target-domain images simultaneously for training [5], [6]. However, due to concerns on privacy, bandwidth and other issues, it is not always possible to access source-domain data and target-domain data simultaneously.

Source-Free Domain Adaptation (SFDA) [7]–[9] aims to adapt a model pre-trained with source-domain images to fit the target data distribution without access to the source data. Due to the absence of annotations in the target domain, the main challenge for SFDA is the lack of sufficient supervision for the model in the target domain. To deal with this problem, some existing works designed auxiliary tasks such as rotation prediction [9], image normalization [10] and auto-encoder-based image reconstruction [11] to assist adaptation in the target domain. However, these works introduce an extra sub-network for the auxiliary task that needs to be trained in the

source domain in advance, which makes these SFDA methods only work for a model pre-trained in a specified way in the source domain and cannot be applied to models pre-trained in other manners, e.g., standard supervised learning without auxiliary tasks.

In this work, we explore a more flexible approach for SFDA, where only a pre-trained segmentation model and unannotated images are available in the target domain, without restrictions on how the model has been pre-trained in the source domain, and we call it fully SFDA. Note that fully SFDA is independent of the pre-training process, and is more general than the auxiliary task-based methods [9]–[11] that require special pre-training strategies and network structures.

To deal with unannotated images in the target domain for fully SFDA, several researchers have investigated some regularization methods, such as entropy minimization for the predictions in the target domain [12], [13], which are inspired by entropy minimization in the UDA [14]–[16] and semi-supervised learning tasks [17]–[19]. However, only using entropy minimization as supervision cannot provide sufficient constraints, which makes the model tend to give high-confidence but incorrect predictions in the target domain. To deal with this problem, some researchers also proposed self-training, which fine-tunes the pre-trained model using its predictions on the target-domain images as pseudo labels [20]–[22]. However, due to the change in the target domain distribution, it is hard to obtain accurate pseudo labels, which brings challenges to achieving good performance [23].

To overcome these problems, we propose a novel Uncertainty-aware Pseudo Label guided Source-Free Domain Adaptation (UPL-SFDA) framework for medical image segmentation. Differently from many existing methods that require a special pre-training strategy in the source domain [9]–[11], our method is agnostic to the training stage and has a minimal requirement on the network structure, which is applicable in wider scenarios. Given a pre-trained network, we propose Target Domain Growing (TDG) that duplicates the prediction head K times in the target domain, and add random perturbations (e.g., dropout, spatial transformation) to obtain K different segmentation predictions. The ensemble of these predictions leads to more robust pseudo labels with efficient uncertainty estimation, which helps to distinguish reliable pseudo labels from unreliable ones. To avoid model degradation commonly faced by self-training, we introduce Twice Forward pass Supervision (TFS) that uses reliable pseudo labels obtained in one forward pass to supervise predictions in a following forward pass. In addition, unlike existing works imposing entropy minimization on each single prediction head [12], [21], we impose entropy minimization on the mean prediction across the K heads instead, which additionally introduces an implicit multi-head consistency regularization to obtain more robust results. Our contributions are summarized as follows:

- We propose a Source-Free Domain Adaptation method based on uncertainty-aware pseudo labels for medical image segmentation, which adapts a model to the target domain without specific requirements on the pre-training strategy and network structure in the source domain.

- We introduce Target Domain Growing (TDG) to expand a pre-trained model with perturbed multiple prediction heads in the target domain, which increases the quality of pseudo labels and obtains uncertainty estimation efficiently.

- A Twice Forward pass Supervision (TFS) is introduced for self-training, which is combined with a mean prediction-based entropy minimization to robustly learn from pseudo labels in SFDA.

Extensive experiments on three applications (multi-site heart MRI segmentation, cross-modality fetal brain segmentation, and fetal tissue segmentation) showed that our method can effectively adapt the model from a source domain to one or multiple target domains. It outperformed several existing SFDA methods for medical image segmentation, and was comparable and even better than supervised training in the target domain.

II. RELATED WORKS

A. Unsupervised Domain Adaption

UDA aims to transfer the knowledge learned from labeled source-domain data to an unlabeled target domain. Current UDA methods mainly adapt the model to the target domain in three aspects. The first is image appearance alignment that translates a target-domain image into a source-domain-like image [24]–[27], so that the domain gap is alleviated. The second is feature alignment that minimizes the distance of feature distribution between the source and target domains to learn domain-invariant representations [28]. For example, for cardiac image segmentation, Wu et. [29] used Variational Auto-Encoders (VAEs) to align the features in the source and target domains, and Chen et al. [30] used Generative Adversarial networks (GANs) to align the features. The third is output alignment, i.e., using the source model to generate pseudo labels in the target domain for adaptation [6]. However, even relying on unpaired and unsupervised domain translation techniques, these UDA methods require access to source domain images, which is hardly guaranteed at a testing site due to the concerns on privacy, computational cost and bandwidth. Therefore, source-free DA is highly desirable in practice.

B. Source-Free Domain Adaption

Source-Free Domain Adaption (SFDA) deals with domain adaption without access to source-domain data [7], [9], [21], [31]. Yang et al. [31] proposed a Fourier-style mining-guided framework, which comprises a generation stage and an adaptation stage for adapting the source model to the target domain using paired source-like and target images. Sun et al. [9] introduced an auxiliary branch to predict the rotation angle in the target domain. Karani et al. [10] introduced a shallow image normalization network before the segmentation model, and fine-tuned the normalization network in the target domain based on predictions refined by a Denoising Auto-Encoder (DAE). However, these methods require the segmentation model's structure to be modified in advance to support the auxiliary task and pre-trained with a specified strategy, which is inapplicable to general segmentation models that
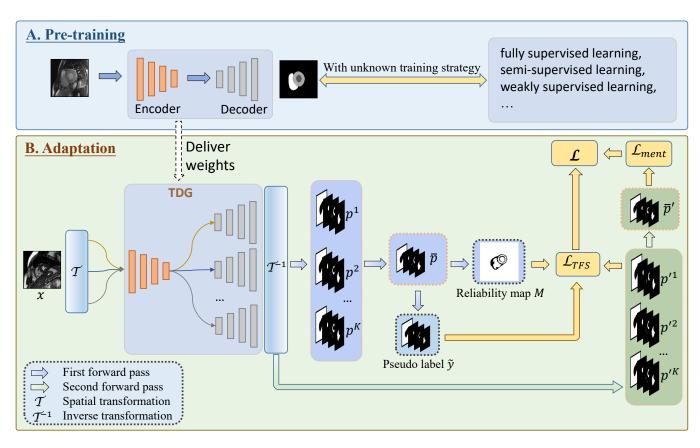
Fig. 1. Overview of our proposed Uncertainty-aware Pseudo Label guided Source-Free Domain Adaptation (UPL-SFDA) framework. In the pre-training stage, the model can be trained in the source domain with an arbitrary strategy. We use Target Domain Growing (TDG) to extend the pre-trained model with multiple prediction heads with perturbations in the target domain. Note that the pseudo label and reliability map obtained in one forward pass are used to supervise the predictions in the next forward pass in the Twice Forward pass Supervision (TFS) loss.

are unaware of the adaptation process during pre-training. Recently, some methods [12], [32] avoid the coupling between training in the source and target domains, so that the adaptation process does not set a prerequisite for training methods in the source domain, which is more general to arbitrary pre-trained models. Wen et al. [7] proposed a selectively updated Mean Teacher for SFDA, where predictions from a teacher model based on exponential moving average is used to supervise the student. Nado et al. [32] proposed Prediction-Time Batch Normalization (PTBN) that recalculates statistics of batch normalization layers according to the images in the target domain. TENT [12] updates the parameters in batch normalization layers to minimize the entropy of predictions in the target domain. In addition to entropy minimization [12], other loss functions, such as regional nuclear-norm loss with contour regularization [33] and consistency regularization [34], have been proposed for the setting. However, due to the lack of annotations, achieving good performance for SFDA methods is still challenging.

III. METHOD

Fig. 1 shows an overview of our proposed Uncertainty-aware Pseudo Label guided Source-Free Domain Adaptation (UPL-SFDA). It is independent of the pre-training stage in the source domain, so it can deal with a model pre-trained in an arbitrary strategy. In UPL-SFDA, we introduce Target

Domain Growing (TDG) to extend the source model into a multi-head prediction structure by duplicating the pre-trained prediction head K times, and then get pseudo labels based on an ensemble of the prediction heads with perturbations using dropout and spatial transformation. Pseudo labels obtained in one forward pass are used to supervise the prediction of the next forward pass, which acts as a consistency regularization between the two forward passes, and they are weighted by the reliability (confidence). For unreliable pixels, we use a mean prediction-based entropy minimization regularization that improves confidence of the predictions and inter-head consistency.

A. Pre-trained Model from the Source Domain

Let S and T be the source and target domains, respectively. Let $\mathbf{X}_S = \{(\boldsymbol{x}_i^s, y_i^s), i = 1, ..., N_s\}$ be the training images and their labels in the source domain, and $\mathbf{X}_T = \{(\boldsymbol{x}_i^t,), i = 1, ..., N_t\}$ represent unlabeled images in the target domain for adaptation, where N_s and N_t are the number of samples in the two domains, respectively. Note that the data distributions in S and T are different, and we assume that the label has the same distribution across the two domains, i.e., the same type of structure for segmentation.

For a general CNN-based segmentation model, it has a feature extractor g and a prediction head h, and the parameters of the segmentation model are denoted as $\{\theta_g, \theta_h\}$, where θ_g

and θ_h denote the parameters of g and h, respectively. As encoder-decoder networks are widely used for medical image segmentation [35], [36], we consider g as an encoder and h as a decoder in this work, respectively. The model is pre-trained in the source domain via:

$$\theta_g^0, \theta_h^0 = \arg\min_{\theta_g, \theta_h} \frac{1}{N_s} \sum_{i=1}^{N_s} L_s\Big(h\big(g(\boldsymbol{x}_i^s)\big), y_i^s\Big) \qquad (1)$$

where L_s donates a certain type of supervision loss in the source domain, which might be implemented by fully supervised learning, semi-supervised learning and weakly supervised learning, etc., based on the type of the available labels in the source domain. θ_g^0 and θ_h^0 denote the optimized parameter values in the source domain, and they are used as initial parameters for the adaptation process in the target domain.

B. Target Domain Growing in the Target Domain

With the pre-trained feature extractor g and prediction head h, the model can be applied to a target-domain image to obtain a prediction as the pseudo label. However, due to the gap between source and target domains, directly applying the pre-trained model will lead to a very low quality of pseudo labels. To improve the quality of pseudo labels for a higher adaptation performance, we propose Target Domain Growing (TDG) to extend the source model, i.e., we duplicate the prediction head (i.e., decoder) h by K times in the target domain, and they are initialized as the pre-trained prediction head with parameter values of θ_h^0. These prediction heads are connected to the same pre-trained feature extractor g in parallel, as shown in Fig. 1.

Let h^k denote the k-th prediction head in the target domain. As they have the same initial parameter values with the same architecture, their outputs will be the same for a given input. To obtain diversity, we introduce perturbations for the prediction heads so that they produce different results for more robust ensemble. Specifically, we use random spatial transformation and dropout to improve the diversity of predictions.

First, for an input image $\boldsymbol{x} \in \mathcal{R}^{H \times W}$ in the target domain, where H and W are the height and width, respectively, we send it into the network K times, each time with a random spatial transformation and for a different prediction head h^k. The segmentation prediction result for the k-th head is:

$$\boldsymbol{p}^k = \mathcal{T}_k^{-1} \circ h^k\big(g(\mathcal{T}_k \circ x)\big) \qquad (2)$$

where $\mathcal{T}_k$ is a random spatial transformation and $\mathcal{T}_k^{-1}$ is the corresponding inverse transformation. $\boldsymbol{p}^k \in \mathcal{R}^{C \times H \times W}$ is the output segmentation probability map with C channels obtained by Softmax, where C is the class number for segmentation. In this paper, we set $\mathcal{T}_k$ as random flipping, random rotation with $\pi/2$, π and $3\pi/2$ for efficient implementation.

Second, we add a dropout layer before each of the prediction head h^k, so that the prediction heads take different random subsets of the features as input. Due to the image-level and feature-level perturbations, the K predictions are different for an input image. We then average across the K predicted segmentation probability maps for ensemble:

$$\bar{\boldsymbol{p}} = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{p}^k \qquad (3)$$

C. Twice forward pass supervision with Reliable Pseudo Labels

With the average probability prediction $\bar{\boldsymbol{p}}$, we take an argmax to obtain the pseudo label for the input $\boldsymbol{x}$. To reduce noises, we post-process it by only keeping the largest component for each foreground class in segmentation tasks where each foreground class has only one component (e.g, heart structure and fetal brain segmentation in this work). Then the post-processed pseudo label is converted into a one-hot representation, which is denoted as $\tilde{y} \in \{0, 1\}^{C \times H \times W}$.

As the domain gap may limit the quality of the pseudo label $\tilde{y}$, directly using $\tilde{y}$ to supervise the network will lead to a limited performance. To deal with this problem, we use the uncertainty information in $\bar{\boldsymbol{p}}$ to identify pixels with reliable pseudo labels and only use the reliable region to supervise the network. To achieve this, we define a binary reliability map $M \in \{0, 1\}^{H \times W}$ for $\tilde{y}$, and each element in M is defined as:

$$M_n = \begin{cases} 1 & \text{if } \bar{\boldsymbol{p}}_{c^*, n} > \tau \\ 0 & \text{otherwise} \end{cases} \qquad (4)$$

where $n = 1, 2, ..., HW$ is the pixel index. $c^* = \arg\max_c(\bar{\boldsymbol{p}}_{c,n})$ is the class with the highest probability for pixel n, and $\bar{\boldsymbol{p}}_{c^*, n}$ represents the confidence for the pseudo label at that pixel. $\tau \in (1/C, 1.0)$ is a confidence threshold.

For pseudo label-based self-training, the model may be biased towards its own prediction in each iteration. To avoid this problem, Chen et al. [37] introduced cross supervision where two networks with different predictions guide each other to reduce the bias. However, the use of two networks would increase the computational and memory cost, and it is not suitable for SFDA where only one pre-trained model is provided. Inspired by Chen et al. [37] and to improve the robustness of pseudo label-based SFDA, we introduce Twice Forward pass Supervision (TFS) for robust adaptation.

Specifically, for a batch of data in the training set, before each gradient back-propagation, we perform two consecutive forward passes. We employ the pseudo label $\tilde{y}$ and its associated reliability map M obtained in the first forward pass to supervise the prediction heads in the second forward pass. Let $\boldsymbol{p}'^k$ denote the output of the k-th prediction head in the second forward pass. Due to the use of random spatial transformation and dropout as mentioned above, the outputs of the two forward passes are different despite the same parameter values. Using $\tilde{y}$ to supervise $\boldsymbol{p}'^k$ can introduce a consistency regularization under perturbations, which improves the robustness of the network. The TFS loss is:

$$\mathcal{L}_{TFS} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}_{w-dice}(\boldsymbol{p}'^k, \tilde{y}, M) \qquad (5)$$

where $\mathcal{L}_{w-dice}$ is the reliability map-weighted Dice loss for a single head. Here we use a Dice-based loss for pseudo label supervision, as Dice loss can better deal with class

TABLE I
DETAILS OF DATASETS USED FOR EXPERIMENTS. THE VALUES REPRESENT VOLUME NUMBERS.

Dataset	M&MS Dataset				FB Dataset		FeTA Dataset	
Domain	A	B	C	D	Source	Target	Source	Target
	Siemens	Philips	General Electric	Canon	HASTE	trueFISP	IRTK	mialSR
Training	135	177	105	70	47	30	28	28
Validation	19	25	15	10	7	5	4	4
Testing	38	50	30	20	14	9	8	8
Overall	192	252	150	100	68	44	40	40

imbalance in segmentation tasks than cross entropy [38], and the segmentation performance is usually evaluated by Dice.

$$\mathcal{L}_{w-dice} = 1 - \frac{1}{C} \sum_{c=1}^{C} \frac{\sum_n 2 M_n \boldsymbol{p}'^k_{c,n} \tilde{y}_{c,n}}{\sum_n M_n (\boldsymbol{p}'^k_{c,n} + \tilde{y}_{c,n}) + \eta} \quad (6)$$

where n is the pixel index and $\eta = 10^{-5}$ is a small number for numeric stability.

D. Mean Prediction-based Entropy Minimization

Entropy minimization is widely used for regularization in semi-supervised learning [19] and SFDA [13], [39], [40], which improves the model's confidence by minimizing the entropy of the class distribution in a prediction output. However, existing entropy minimization methods for SFDA are applied to networks with a single prediction head. For our method with multiple prediction heads, enforcing entropy minimization for each prediction head respectively may lead to sub-optimal results when different predication heads obtain opposite results with high confidence. For example, for binary segmentation, when h^k and h^{k+1} predict one pixel as being the foreground with probability of 0.0 and 1.0 respectively, both branches have the lowest entropy, but their average has a high entropy. To overcome this problem, we apply entropy minimization to the mean prediction across the K heads:

$$\mathcal{L}_{ment} = -\frac{1}{HW} \sum_{n=1}^{HW} \sum_{c=1}^{C} \bar{\boldsymbol{p}}'_{c,n} log(\bar{\boldsymbol{p}}'_{c,n}) \quad (7)$$

where $\bar{\boldsymbol{p}}'$ is the mean probability prediction obtained by the K heads in the second forward of TFS. Compared with minimizing the entropy of each prediction head respectively, minimizing the entropy of their mean prediction $\bar{\boldsymbol{p}}'$ can not only reduce the uncertainty of each head, but also encourage the consistency between them. Thus, it helps to improve the robustness of the network on samples in the target domain.

E. Adaptation by Self-training

Our adaptation method adopts a self-training process on unlabeled images in the target domain. Based on the pseudo labels obtained by TDG, the overall loss function for tuning the network with TFS in the target domain is:

$$\mathcal{L} = \mathcal{L}_{TFS} + \lambda \mathcal{L}_{ment} \quad (8)$$

where λ is a hyper-parameter to control the weight of $\mathcal{L}_{ment}$. Note that there are two forward passes for each parameter update step, where the first forward pass obtains pseudo labels,

and the loss function is calculated in the second pass for parameter update with back-propagation.

IV. EXPERIMENTS

A. Datasets and Implementation

We used three datasets for experiments: 1) the public Multi-centre, multi-vendor and multi-disease cardiac image segmentation (M&MS) dataset [41], where the images were acquired by devices with four different vendors, 2) an in-house Fetal Brain (FB) segmentation dataset that contains two different MRI sequences, and 3) a public Fetal Tissue Annotation (FeTA) dataset that contains two different super-resolution methods [42]. A summary of these three datasets is listed in Table I.

1) Cardiac Image Segmentation Dataset (M&MS): The M&MS dataset [41] consists of 345 cardiac MRI volumes collected from six different hospitals, using four different scanner vendors, namely Siemens, Philips, General Electric, and Canon. The imaging devices were MAGNETOM Avanto for hospital 1, Achieva for hospital 2 and 3, Signa Excite, Vantage Orian, and MAGNETOM Skyra for hospital 4, 5 and 6, respectively. Following [41], we divide the dataset into four domains: Domain A for Siemens, comprising data from hospitals 1 and 6; Domain B for Philips, comprising data from hospitals 2 and 3; Domain C for General Electric, comprising data from hospital 4; and Domain D for Canon, comprising data from hospital 5. The slice number per volume varied from 10 to 13. The in-plane resolution ranged from 0.85 to 1.45 mm with slice thickness 9.2-10 mm. Following the setting in [40], we used domain A as the source domain, and B, C and D as the target domains. The target tissues for segmentation are the Left Ventricle (LV), Right Ventricle (RV) and Myocardium (MYO). We randomly split images in each domain into 70%, 10% and 20% for training, validation and testing, respectively, and abandoned labels for the training sets in the target domains.

2) Fetal Brain (FB) Segmentation Dataset: The FB dataset had fetal MRI with two imaging protocols acquired from a single center, including 68 volumes of half-Fourier acquired single turbo spin-echo (HASTE) and 44 volumes of true fast imaging with steady state precession (TrueFISP). The slice number for each volume varied from 11 to 22, and the gestational age ranged in 21-33 weeks. The two sequences had an in-plane resolution of 0.64 to 0.70 mm and 0.67 to 1.12 mm respectively, with slice-thickness 6.5 - 7.15 mm and 6.5 mm, respectively. HASTE and TrueFISP were used as the source and target domains, respectively. We randomly split the images in each domain into 70%, 10% and 20% for training,

TABLE II

DICE (%) OF DIFFERENT METHODS ON THE M&MS DATASET FOR CARDIAC STRUCTURE SEGMENTATION IN THE TARGET DOMAINS. THE BOLD FONT HIGHLIGHTS THE BEST VALUES IN THE FIRST AND SECOND SECTIONS, RESPECTIVELY. ASTERISKS INDICATE STATISTICAL SIGNIFICANCE WHEN COMPARING THE METHODS WITH THE SOURCE ONLY (*: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$) USING A PAIRED STUDENT'S T-TEST.

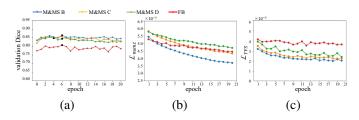| Method	Target domain B			Target domain C			Target domain D		
	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV
Source only	87.54±10.40	75.50±10.97	81.50±16.97	86.47±7.61	77.46±7.71	80.55±11.44	88.04±6.71	75.88±8.57	77.76±17.88
Source only-Esb	88.62±8.19	77.22±9.99	82.82±19.22	86.60±7.94	78.54±9.53	84.26±9.14	88.22±7.56	77.53±8.06	79.35±17.56
Fine-tune valid	90.34±6.30	81.68±6.53	85.30±12.13	89.54±6.06	82.82±4.67	86.20±8.20	88.38±8.46	81.08±4.10	83.47±11.69
Fine-tune train	90.90±5.36	83.76±5.48	**87.63±6.11**	**89.59±5.69**	**83.98±4.85**	**87.46±5.38**	**90.68±5.40**	**83.89±4.29**	**85.93±5.57**
Target only	**91.13±6.37**	**84.37±6.56**	87.27±8.86	89.40±7.57	82.67±5.66	82.99±7.85	88.69±8.35	81.60±5.35	83.41±11.25
PTBN [32]	89.62±7.11	79.99±6.40**	82.31±15.73	86.06±8.96	79.62±7.07	83.76±7.34	88.19±6.51	79.03±4.92	81.01±11.30
TENT [12]	89.03±8.46	79.96±6.44**	83.72±11.42	84.97±10.98	78.68±6.76	84.65±7.18	84.28±10.62	79.08±4.07	82.34±9.70
TTT [9]	89.41±7.06	79.50±6.99**	82.72±13.27	85.89±9.12	79.57±7.10	83.62±6.56	88.13±6.93	79.91±4.60	82.31±9.63
URMA [21]	90.38±5.64	82.09±5.39***	84.30±7.27	88.44±6.29*	81.73±6.21*	86.52±4.90*	88.94±5.92	80.69±4.74	83.01±7.61
Ours w/o Esb	90.70±5.38*	81.82±5.81***	85.73±9.22	89.74±3.98**	83.10±5.61**	86.69±5.50*	89.09±6.00	80.89±3.91	83.87±9.37
Ours	**91.02±5.50 ***	**82.77±5.25***	**87.33±7.87***	**89.64±4.03***	**84.00±5.04***	**88.73±4.51***	**89.13±6.04**	**81.84±4.27***	**85.30±9.53**

(a)   (b)   (c)

Fig. 2. Evolution of validation Dice, $\mathcal{L}_{ment}$ and $\mathcal{L}_{TFS}$ during adaptation. The black squares mark the epoch with the highest validation Dice.

validation and testing, respectively, and abandoned the labels of training images in the target domain.

3) Fetal Tissue Annotation (FeTA) Challenge Dataset: The FeTA Dataset [42] used in this study was from the FeTA2022 challenge[1] that aims to segment seven different tissues, namely External Cerebrospinal Fluid (ECF), Grey Matter (GM), White Matter (WM), Ventricles (Ven), Cerebellum (Cer), Deep Grey Matter (DGM), and Brain Stem (BS). The official dataset has 120 samples, but only 80 samples are publicly available after the challenge, and they were acquired from the University Children's Hospital Zurich (Kispi) using 1.5T and 3T clinical GE whole-body scanners. T2-weighted single-shot Fast Spin Echo sequences were acquired with an in-plane resolution of 0.5mm × 0.5mm and a slice thickness of 3 to 5 mm. To obtain high-resolution fetal brain reconstructions, the mialSR super-resolution (SR) method [43] was used for 40 cases, while the Simple IRTK method [44] was used for the other 40 cases. We used the 40 cases reconstructed by Simple IRTK as the source domain, and the other 40 cases reconstructed by mialSR as the target domain. For each domain, the 3D SR volumes were divided into training, validation, and testing sets in the ratio of 70%, 10% and 20%, respectively.

4) Implementation Details: All the experiments were implemented with PyTorch, using an NVIDIA GeForce RTX 2080Ti GPU. Our code is made available online[2]. For M&MS dataset and FB datasets that have a large slice thickness, we selected the widely used 2D UNet [35] to demonstrate the effectiveness of our method, as most medical image segmentation models are based on UNet-like structures [1]. The image intensity was

[1] https://feta.grand-challenge.org/
[2] https://github.com/HiLab-git/UPL-SFDA

clipped by the 1st and 99-th percentiles, and linearly normalized to [-1,1]. Each slice in the M&MS dataset was center cropped to 256×256, and the slices in the FB dataset were resized to 256×256. For the FeTA dataset, we cropped the 3D volumes based on the brain region during preprocessing, and used the 3D U-Net architecture [45] for implementation. Due to memory limitation, we cropped the images to a patch size of [32, 64, 64]. In the inference stage, we applied a sliding window using the same patch size with a stride of 50% to obtain the final segmentation results. During pre-training in the source domain, we trained the source model for 400 epochs with Dice loss, Adam optimizer and initial learning rate of 0.01 that was decayed to 90% every 4 epochs. The model parameters with the best performance on the validation set in the source domain were used for adaptation. For adaptation in each target domain, we duplicated the decoder for K times, and updated all the model parameters for 20 epochs with Adam optimizer and a fixed learning rate of 10^{-4}.

The hyper-parameter setting was determined based on the labeled validation set of the target domain. Specifically, $K = 4$ and $\lambda = 1.0$. τ was set to 0.95 for the M&MS and FeTA dataset, and 0.9 for FB dataset, respectively. In the adaptation stages, for M&MS and FB dataset, we set all the slices in a single volume as a batch, and for FeTA dataset, the batch size was set to 4. After training, we used the checkpoint with the best performance on the validation set for inference. Fig. 2 shows the evolution of validation Dice, $\mathcal{L}_{ment}$ and $\mathcal{L}_{TFS}$. It can be observed that the loss functions converge in 20 epochs, and the best checkpoint was obtained at epoch 6 for M&MS B, C, 4 for M&MS D and 6 for the FB dataset, respectively.

For quantitative evaluation of the volumetric segmentation results, we adopted the commonly used 3D Dice score and Average Symmetric Surface Distance (ASSD). As the slice thickness was large (6-10 mm) in the M&MS and FB datasets, we calculated ASSD values with unit of pixel.

B. Comparison with State-of-the-art Methods

To verify the effectiveness of our proposed UPL-SFDA, we compared it with four state-of-the-art SFDA methods: 1) **PTBN** [32] that updates batch normalization statistics based on unlabeled training images in the target domain without loss

TABLE III
ASSD (PIXELS) OF DIFFERENT METHODS ON THE M&MS DATASET FOR CARDIAC STRUCTURE SEGMENTATION IN THE TARGET DOMAINS. THE
BOLD FONT HIGHLIGHTS THE BEST VALUES IN THE FIRST AND SECOND SECTIONS, RESPECTIVELY. ASTERISKS INDICATE STATISTICAL
SIGNIFICANCE WHEN COMPARING THE METHODS WITH THE SOURCE ONLY (*: P $\leq$ 0.05, **: P $\leq$ 0.01) USING A PAIRED STUDENT'S T-TEST.

Method	Target domain B			Target domain C			Target domain D		
	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV
Source only	0.55±0.46	0.64±0.45	0.88±1.10	0.58±0.37	0.57±0.21	1.18±1.38	0.54±0.34	0.59±0.30	1.59±2.55
Source only-Esb	0.49±0.43	0.61±0.48	0.77±1.54	0.54±0.30	0.54±0.22	0.64±0.56	0.53±0.38	0.57±0.32	0.85±0.91
Fine-tune valid	0.43±0.36	0.55±0.46	0.49±0.44	0.46±0.30	0.49±0.18	0.68±0.68	0.53±0.43	0.50±0.23	0.69±0.60
Fine-tune train	**0.43±0.44**	**0.50±0.45**	**0.43±0.29**	**0.41±0.20**	**0.40±0.12**	**0.53±0.38**	**0.36±0.16**	**0.39±0.11**	**0.48±0.22**
Target only	0.52±0.77	0.55±0.61	0.54±0.95	0.40±0.24	0.44±0.16	1.31±0.99	0.57±0.55	0.51±0.24	0.88±0.69
PTBN [32]	0.51±0.43	0.60±0.46	0.79±1.33	0.65±0.44	0.53±0.16	1.03±0.92	0.63±0.53	0.55±0.26	1.21±1.41
TENT [12]	0.61±0.69	0.62±0.54	0.67±0.99	0.71±0.65	0.59±0.24	0.87±0.73	0.88±0.72	0.59±0.25	0.55±0.31
TTT [9]	0.48±0.38	0.59±0.47	0.82±1.23	0.71±0.55	0.55±0.19	1.16±0.89	0.64±0.53	0.53±0.21	1.01±1.20
URMA [21]	**0.41±0.25***	**0.51±0.35**	0.50±0.23*	0.45±0.23*	0.46±0.13**	0.53±0.32*	**0.47±0.22**	0.50±0.16	0.52±0.22
Ours w/o Esb	0.46±0.52	**0.58±0.57**	0.54±0.37*	0.42±0.24**	0.46±0.18*	0.62±0.43	0.56±0.39	0.51±0.21	0.59±0.46
Ours	0.45±0.54	0.54±0.57	**0.40±0.38***	**0.40±0.16****	**0.43±0.16****	**0.39±0.26****	0.48±0.30	**0.47±0.19**	**0.46±0.29**

TABLE IV
QUANTITATIVE COMPARISON OF DIFFERENT SFDA METHODS FOR FB
SEGMENTATION. * DENOTES SIGNIFICANT DIFFERENCE (p-VALUE $\leq$
0.05) FROM SOURCE ONLY USING A PAIRED STUDENT'S T-TEST.

Methods	Dice (%)	ASSD (pixel)
Source only	84.09±6.34	1.33±0.49
Source only-Esb	86.39±6.94	1.02±0.41
Fine-tune valid	85.26±5.38	2.09±1.44
Fine-tune train	**89.71±4.87**	**0.86±0.49**
Target only	88.85±4.12	0.91±0.30
PTBN [32]	85.70±4.88	1.85±0.96
TENT [12]	85.75±3.62	1.60±0.71
TTT [9]	85.84±4.52	1.80±0.90
URMA [21]	84.12±6.82	2.18±1.19
Ours w/o Esb	87.95±4.61	1.37±1.21
Ours	**89.10±3.09***	**1.08±0.49**

functions for optimization; 2) **TENT** [12] that only updates the parameters of batch normalization layers by minimizing the entropy of model predictions in the target domain; 3) **TTT** [9] that uses an auxiliary decoder to predict the rotation angle of target-domain images, and the auxiliary task's loss is used to update the model parameters; and 4) **URMA** [21] that uses pseudo labels generated by a frozen main decoder to supervise auxiliary decoders. We also compared our method with four naive methods: 1) **Source only** where the model pre-trained in the source domain is directly used for inference in the target domain, which serves as the lower bound; 2) **Target only** that uses training images and their labels in the target domain to train a model directly, without pre-training in the source domain; and 3) **Fine-tune train** and 4) **Fine-tune valid** that mean the model pre-trained in the source domain is fine-tuned with the annotated training and validation sets in target domain based on fully supervised learning, respectively. In order to investigate the impact of ensembling, we conducted two additional experiments: 1) **Source only-Esb** that refers to ensemble based on spatial transformations of input images for inference with the pre-trained source model; 2) **Ours w/o Esb** where our method did not utilize any spatial transformations and made predictions using only one decoder. We implemented all the compared methods with the same backbone, i.e., UNet [35] for M&MS and FB dataset, and

3D UNet [45] for FeTA dataset for a fair comparison.

1) Result for Cardiac Image Segmentation: For the M&MS dataset, we used domain A as the source domain, and adapted the pre-trained model to domain B, C and D, respectively. Table II and III show the quantitative comparison between the compared methods in terms of Dice and ASSD, respectively. It can be observed the "Target only" outperformed "Source only" substantially, showing the large domain gap between the source and target domains. For example, in target domain B, "Source only" achieved an average Dice of 87.54%, 75.50% and 81.50% for LV, MYO and RV, respectively, and the corresponding Dice values obtained by "Target only" were 91.13%, 84.37% and 87.27% respectively.

The second sections in Table II and III show that all the compared methods outperformed "Source only". PTBN [32], TENT [12] and TTT [9] obtained a moderate improvement from "Source only". For example, in Target domain B, they improved the average Dice for LV from 87.54% to 89.62%, 89.03% and 89.41%, respectively. URMA [21] obtained a higher Dice (90.38%) than these three methods, but it was inferior to our method (91.02%). The average Dice across the three target structures obtained by our method was 87.04%, 87.46% and 85.43% in the three target domains, respectively, compared with the corresponding values of 81.51%, 81.49% and 80.56% achieved by "source only", showing that our method improved the average Dice scores by 5.53, 5.97 and 4.87 percentage points in the three target domains respectively.

In terms of average Dice values, our method outperformed "Fine-tune valid", and was close to "Target only" (p-value > 0.05) in target domain B, and better than "Fine-tune train", "Fine-tune valid" and "Target only" in target domain C. In target domain D, our method also outperformed "Target only". Note that "Target only", and "Fine-tune train" require annotations in the training set of the target domain, while our adaptation method could achieve a similar performance without the annotations. We also analyzed the effectiveness of ensemble of multiple prediction heads with spatial transformations. Taking M&MS B as an example, "Source only-Esb" performed better than "Source only", indicating the positive effect of additional data augmentations for inference. In addition, "Ours w/o Esb" exhibited a decreased performance compared with our com-

TABLE V

DICE (%) OF DIFFERENT SFDA METHODS ON THE FETA DATASET FOR FETAL TISSUE SEGMENTATION. ASTERISKS INDICATE STATISTICAL SIGNIFICANCE WHEN COMPARING THE METHODS WITH SOURCE ONLY (*: $P \leq 0.05$, **: $P \leq 0.01$) USING A PAIRED STUDENT'S T-TEST. THE BOLD FONT HIGHLIGHTS THE BEST VALUES IN THE FIRST AND SECOND SECTIONS, RESPECTIVELY.

Method	ECF	GM	WM	Ven	Cer	DGM	BS	Average
Source only	77.55±5.78	63.68±6.09	83.89±4.16	75.03±10.35	73.91±18.54	52.57±12.54	51.46±28.54	68.30±12.28
Source only-Esb	78.03±5.65	64.54±6.23	83.89±4.19	76.32±8.34	73.88±18.68	47.98±13.45	47.95±27.99	67.51±12.08
Fine-tune valid	79.73±7.95	65.65±6.82	81.83±9.85	77.12±13.25	77.45±12.04	69.86±5.87	50.14±23.58	71.68±11.34
Fine-tune train	85.59±3.23	71.55±5.81	**90.30±2.68**	**85.18±8.03**	**87.78±4.98**	81.49±6.76	**73.17±16.58**	**82.15±6.87**
Target only	**86.16±2.23**	**71.80±5.17**	89.93±3.11	83.11±9.10	84.38±5.37	**82.33±5.00**	71.21±12.77	81.27±6.11
PTBN [32]	77.60±7.70	62.54±7.73	82.32±5.50	74.06±10.63	80.20±12.46	57.91±15.46	52.13±23.86	69.53±11.91
TENT [12]	81.43±4.73*	65.85±5.06*	84.49±4.36	73.85±10.00	80.59±15.97*	62.26±9.12*	60.00±20.33	72.64±9.94*
TTT [9]	80.00±5.98	63.77±6.53	83.06±4.36	74.29±10.39	81.57±12.51	57.57±12.81	56.05±22.42	70.90±10.71
URMA [21]	81.76±5.53*	65.95±5.21*	**84.56±4.71**	73.97±9.54	83.02±13.10*	**64.78±8.26****	62.52±19.89	73.79±9.46*
Ours w/o Esb	84.16±3.32	66.06±5.77	83.56±4.03	75.07±9.14	84.02±10.68	63.80±8.17	**67.36±11.83**	74.86±7.56*
Ours	**84.75±3.15****	**66.98±5.67***	83.96±3.82	**76.66±7.61**	**84.91±8.18***	62.46±8.79*	66.57±13.52*	**75.19±7.25****

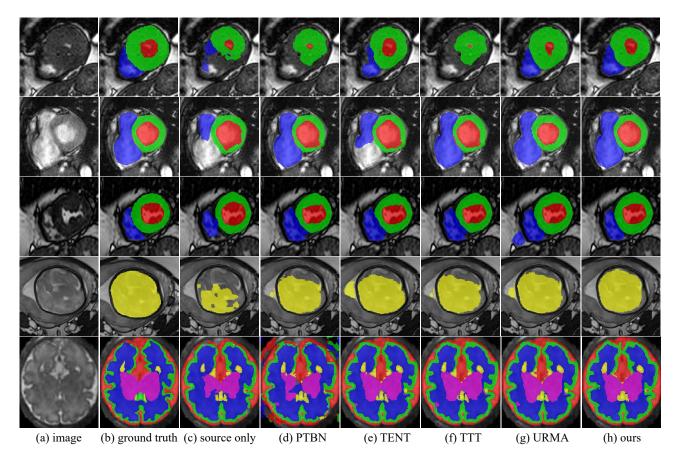| (a) image | (b) ground truth | (c) source only | (d) PTBN | (e) TENT | (f) TTT | (g) URMA | (h) ours |

Fig. 3. Qualitative comparison of different SFDA methods. The top three rows are from domain B, C and D on M&MS dataset respectively. The last two rows are from the target domain of FB and FeTA datasets, respectively.

plete method. This suggests that ensembling during inference plays a beneficial role in our approach. A visual comparison between different SFDA methods is shown in Fig. 3. Note that "Source only" achieved a poor performance, and the results of our method were closer to the ground truth than those of the other methods.

2) Results for Fetal Brain Segmentation: We further investigated the performance of the compared methods on FB dataset, with HASTE and TrueFISP as the source and target domains, respectively. The quantitative evaluation results are shown in Table IV. It can be observed that "Source only" and "Target

only" achieved an average Dice of 84.09% and 88.85%, respectively, showing the large gap between the two domains. "Fine-tune train" outperformed "Target only", achieving an average Dice of 89.71%. The existing methods only achieved a slight improvement compared with "Source only", with the Dice values ranging from 84.12% to 85.84%. In contrast, our method largely improved it to 89.10%, which outperformed "Target only" and was close to "Fine-tune train" (p-value > 0.05). Our method achieved an average ASSD of 1.08 pixels, which was lower than those of the other SFDA methods. The qualitative comparison in the penultimate row of Fig. 3 shows
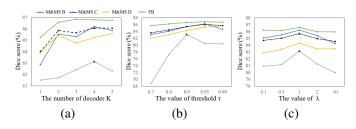
Fig. 4. Performance of our method with different hyper-parameter values on the validation sets of different target domains.

that the existing methods tend to achieve under-segmentation of the fetal brain, while our method can successfully segment the entire fetal brain region with high accuracy.

3) Results for 3D Fetal Tissue Segmentation: Quantitative evaluation results on the FeTA dataset in terms of Dice are shown in Table V. It shows that "Source only" and "Target only" achieved an average Dice of 68.30% and 81.27%, respectively, indicating the large gap between the two domains. Our method increased the average Dice by 6.89 percentage points compared with "Source only", reaching 75.19%. In contrast, the existing methods had a lower performance than ours. The average Dice obtained by PTBN [32], TENT [12], TTT [9] and URMA [21] was 69.53%, 72.64%, 70.90% and 73.79%, respectively. The qualitative comparison in the last row of Fig. 3 demonstrates that our method outperformed the other methods in terms of segmentation performance.

C. Ablation Analysis of Our UPL-SFDA

1) Effect of Hyper-parameters: There are three important hyper-parameters specific to our method: the number of duplicated prediction heads K, the confidence threshold τ to select reliable pseudo labels for supervision, and the loss weight λ. We first investigated the effect of K by setting it to 1 to 5 respectively, and the performance on the validation sets of the two datasets are shown in Fig. 4(a). It can be observed that $K = 1$ performed worse than larger K values, showing the superiority of using TDG. The performance on both datasets improved when K changed from 1 to 4, and $K = 5$ did not further bring performance improvement. Therefore, we set $K = 4$ for our method.

Then we investigated how τ affected the pseudo labels and the SFDA performance. Fig. 5 shows some examples of reliable pseudo labels with different τ values. We found that a higher threshold τ will lead to smaller reliable regions for each class, which helps to avoid the model being affected by unreliable regions of the pseudo labels. Quantitative comparison between different τ values is demonstrated in Fig. 4(b), which shows that the performance on the M&MS dataset was relatively stable with different τ values, and $\tau = 0.95$ performed slightly better than the other values in average. The best τ value on the FB dataset was 0.90 based on performance on the validation set. Therefore, we set τ to 0.95 and 0.9 for the two datasets, respectively. The performance on the validation set with different λ values is shown in Fig. 4(c). It demonstrates that the best λ was 1.0 for the different datasets.

Fig. 6 shows the reliable pseudo labels obtained at different training epochs in the target domains. It can be observed

that the pseudo labels are updated during the self-training process, and their quality gradually improves at different training epochs. In addition, the confidence of the pseudo labels also improves with the increase of training epochs.

2) Ablation study of each component: To evaluate the effectiveness of each of the proposed components in our UPL-SFDA, we set the baseline as updating the source model based on self-training where the network was supervised by its own prediction and an entropy minimization loss. The quantitative results obtained by different variants of our method are shown in Table VI, where M means using the binary reliability map to weight pseudo labels, TDG means using target domain growing with dropout before each prediction head, and $\mathcal{T}$ means using random spatial transformation for each prediction head. $\mathcal{L}_{ment}$ means minimizing entropy of the mean prediction across the K heads, rather than minimizing entropy of each head respectively.

Table VI shows that each component of our method led to a performance improvement. Take the performance on the domain C of M&MS dataset as an example, the average Dice score obtained by "Source only" was 81.47%. The baseline obtained an average Dice of 84.20%, and introducing reliability map weighting for pseudo labels improved it to 85.09%. For TDG, only using dropout for perturbations obtained an average Dice of 85.22%, and additionally using spatial transformation for the prediction heads improved it to 86.16%, showing that the spatial transformation plays an important role in our method. Then, using our Twice Forward pass Supervision (TFS) loss improved it to 86.79%, and our proposed method combining all these modules with $\mathcal{L}_{ment}$ obtained the highest Dice score of 87.46%. Note that by removing the spatial transformation for the prediction heads in our method, the average Dice decreased to 85.43%. We also tried to only combine L_{ment} loss with TDG using the spatial transformations (i.e., removing TFS loss), and the average Dice dropped to 86.94%. In addition, Table VI shows that our method outperformed "Target only" on domains C and D in the M&MS dataset and the target domain of FB dataset in terms of average Dice score.

V. DISCUSSION

Our proposed UPL-SFDA based on TDG and reliable pseudo label supervision has several advantages over existing SFDA methods for domain adaptation without access to source-domain images. First, unlike some existing methods [9]–[11] using auxiliary branches in the network that require special training strategies in the source domain, our method does not require training auxiliary branches in the source domain, and it makes the training methods in the source and target domains independent. This decoupling makes our method more general to a wider range of pre-trained models. Second, compared with existing methods using entropy minimization for regularization [12], our method uses reliable pseudo labels for adaptation, which provides more effective supervision signals for model update. In addition, based on the TDG strategy with perturbations, we obtain multiple predictions that can provide high-quality pseudo labels with

TABLE VI

ABLATION STUDY ON DIFFERENT COMPONENTS OF OUR UPL-SFDA. THE FIRST ROW (BASELINE) IS UPDATING THE SOURCE MODEL WITH PSEUDO LABELS OBTAINED BY ITSELF AND ENTROPY MINIMIZATION. M: USING THE BINARY RELIABILITY MAP TO WEIGHT PSEUDO LABELS. TDG: TARGET DOMAIN GROWING WITH DROPOUT BEFORE EACH PREDICTION HEAD. $\mathcal{T}$: RANDOM SPACIAL TRANSFORMATION FOR EACH PREDICTION HEAD.

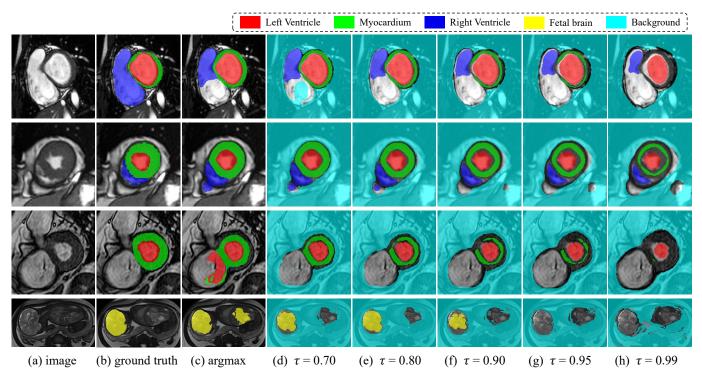| Components					Dice (%)				ASSD (pixel)			
M	TDG	$\mathcal{T}$	TFS	$\mathcal{L}_{ment}$	M&MS B	M&MS C	M&MS D	FB	M&MS B	M&MS C	M&MS D	FB
					84.10±9.67	84.20±6.16	83.37±6.90	83.44±7.38	0.60±0.84	0.64±0.36	0.60±0.38	1.39±0.76
✓					84.49±10.35	85.09±5.79	84.10±6.97	85.88±5.49	0.58±0.80	0.49±0.21	0.56±0.36	1.56±1.01
✓	✓				84.52±10.31	85.22±5.57	84.12±6.89	86.77±4.17	0.56±0.63	0.50±0.22	0.55±0.34	0.95±0.24
✓	✓	✓			86.39±7.90	86.16±6.17	84.98±7.08	86.92±5.41	**0.44±0.43**	0.46±0.25	0.48±0.26	0.91±0.36
✓	✓	✓	✓		86.52±6.63	86.79±4.44	85.20±7.02	88.11±5.06	0.46±0.44	0.42±0.17	0.48±0.26	**0.86±0.34**
✓	✓	✓	✓	✓	**87.04±6.20**	**87.46±4.52**	**85.43±6.61**	**89.10±3.09**	0.46±0.38	**0.40±0.24**	**0.47±0.33**	1.08±0.49
✓	✓	✓		✓	85.16±6.91	85.43±6.54	84.42±6.89	87.57±3.20	0.49±0.32	0.56±0.37	0.54±0.36	1.02±0.27
	✓	✓		✓	85.82±7.95	86.94±5.13	84.09±7.87	84.48±6.88	0.46±0.42	0.41±0.20	0.55±0.39	1.52±0.92
Source only					81.51±12.78	81.47±8.92	80.56±11.05	84.09±6.34	0.69±0.67	0.71±0.65	0.90±1.06	1.33±049
Target only					87.59±7.26	85.02±7.02	84.56±8.31	88.85±4.12	0.53±0.77	0.71±0.46	0.65±0.49	0.91±0.30

Fig. 5. Effect of confidence threshold τ on reliable pseudo labels. The first three rows are from domain B, C and D on M&MS dataset respectively, and the last row is from the target domain of FB dataset. (c) shows pseudo labels obtained by argmax, and (d)-(h) are reliable pseudo labels with different τ values, where uncolored regions are pixels with unreliable pseudo labels.

(a) image (b) ground truth (c) argmax (d) $\tau = 0.70$ (e) $\tau = 0.80$ (f) $\tau = 0.90$ (g) $\tau = 0.95$ (h) $\tau = 0.99$

efficient uncertainty estimation, which prevents the model being corrupted by unreliable pseudo labels. Using entropy minimization on the average prediction across the multiple heads can encourage a consistency between them, which also improves the robustness of our method.

The pseudo label-based supervision loss $\mathcal{L}_{w-dice}$ and the unsupervised regularization loss $\mathcal{L}_{ment}$ have two similarities. First, both of them are based on multi-head agreement. $\mathcal{L}_{w-dice}$ uses relatively consensus regions of the K prediction heads as pseudo labels, and $\mathcal{L}_{ment}$ encourages the K prediction heads to obtain consensus results by minimizing the uncertainty in the average prediction. Second, the two terms will increase the confidence of the predictions. $\mathcal{L}_{w-dice}$ drives the predictions to be closer to the hard pseudo labels, while $\mathcal{L}_{ment}$ directly minimizes the entropy, and both of them will

reduce uncertain predictions. However, they also have several important differences. First, $\mathcal{L}_{w-dice}$ encourages consistency across two different forward passes with feature perturbations, while $\mathcal{L}_{ment}$ is for consistency across prediction heads. Second, $\mathcal{L}_{w-dice}$ is applied to high-confidence pixels (with a threshold of τ), while $\mathcal{L}_{ment}$ is applied to the entire image region. Thirdly, $\mathcal{L}_{w-dice}$ is a pseudo label-based supervision loss, while $\mathcal{L}_{ment}$ is an unsupervised loss for regularization. Therefore, the two terms are complementary to each other.

Introducing perturbations to the K prediction heads in TDG is important for achieving good performance. Without perturbation, the K prediction heads will obtain the same result, which degrades to just using the pre-trained model with a single prediction head. With perturbations, the K prediction results are different and their ensemble is more robust, which

(a) image   (b) ground truth   (c) Source only   (d) epoch $\frac{\epsilon}{4}$ (e) epoch $\frac{\epsilon}{2}$ (f) epoch $\frac{3\epsilon}{4}$ (g) epoch ϵ
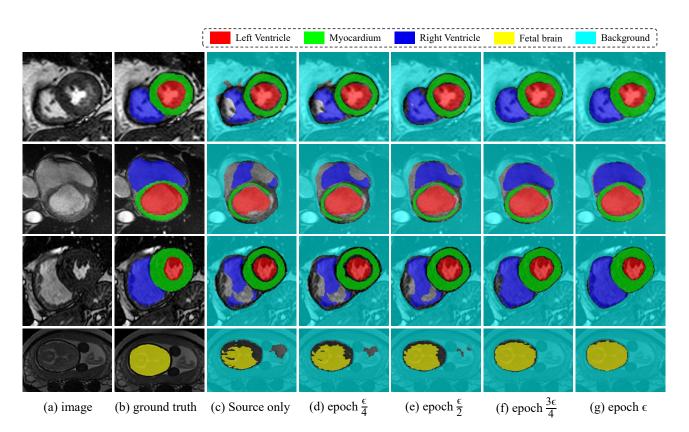
Fig. 6.   Pseudo labels at different training steps in self-training. ϵ means the epoch number with the highest performance on the validation set. The first three rows are from domain B, C and D of M&MS dataset respectively, and the bottom row is from the target domain of FB dataset. In (c)-(g), only reliable pseudo labels are encoded by colors, and pixels without encoded colors will be ignored in the calculation of TFS loss.

can overcome the bias in each prediction head and lead to uncertainty estimation. In addition, we implemented our TDG with an encoder-decoder structure due to that most state-of-the-art CNNs for medical image segmentation have an encoder-decoder structure [35], [36]. It may also be applied to other networks [46] by duplicating the prediction head multiple times with perturbations in the target domain.

In our experiment, a validation set with annotations in the target domain is used to select hyper-parameters for the compared methods. The advantage of using the labeled validation set is that it allows to find the optimal hyper-parameters such as learning rate and weights of loss terms of each compared method. In addition, it allows early stopping and checkpoint selection to avoid over-fitting on the training set in the target domain, which ensures a fair comparison between the different methods. One may also use the validation set to update the model weights by fine-tuning, which could provide more supervision signal directly to the model for parameter optimization. However, it may lead the model to over-fit the validation set that is usually small. In addition, using the validation set for hyper-parameter selection rather than model learning is a work standard in the machine learning community. However, in some cases, the labeled validation set may not be available, making it less practical to use the validation set to fine-tune the pre-trained model.

This work still has some limitations. First, our method involves performing two forward passes for each gradient back-propagation, which takes more time than using a single forward pass. The training time consumption for our method is slightly higher than TENT [12], but lower than URMA [21]. For instance, in M&MS B, our method takes an average of 0.661s per case to train one epoch, while TENT and URMA require 0.342s and 0.944s in average, respectively. The average inference time for our method is 0.342s per case, and slightly higher than TENT's 0.269s. Second, we have employed a labeled validation set in the target domain to select the optimal hyper-parameters. However, in practical applications, acquiring a validation set could be challenging, making it hard to determine hyper-parameters. Additionally, TDG with multiple prediction heads increase the memory cost, which does not allow a large patch size or batch size for dealing with 3D medical images and may limit the performance.

VI. Conclusion

In conclusion, we propose a novel uncertainty-aware pseudo label-guided approach for Source-Free Domain Adaptation (UPL-SFDA) in medical image segmentation, which uses target domain growing to generate multiple predictions for an input to obtain reliable pseudo labels with a weight map based on uncertainty estimation. The network is supervised by the weighted pseudo labels and minimizing the entropy of the average of the multiple predictions. A twice forward pass supervision strategy is also proposed to avoid the network being biased towards its own predictions in self-training. Experimental results on multi-site heart MRI segmentation and cross-modality fetal brain segmentation showed that our

method outperformed existing SFDA methods, and it was comparable to and even better than supervised training in the target domain. In the future, it is of interest to apply our method to other segmentation tasks.

REFERENCES

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[2] Q. Duan, G. Wang, R. Wang, C. Fu, X. Li, M. Gong, X. Liu, Q. Xia, X. Huang, Z. Hu *et al.*, "Sensecare: a research platform for medical image informatics and interactive 3D visualization," *arXiv preprint arXiv:2004.07031*, 2020.

[3] R. Gu, J. Zhang, R. Huang, W. Lei, G. Wang, and S. Zhang, "Domain composition and attention for unseen-domain generalizable medical image segmentation," in *MICCAI*, 2021, pp. 241–250.

[4] H. Guan and M. Liu, "Domain adaptation for medical image analysis: a survey," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 3, pp. 1173–1185, 2021.

[5] C. Pei, F. Wu, L. Huang, and X. Zhuang, "Disentangle domain features for cross-modality cardiac image segmentation," *Medical Image Analysis*, vol. 71, p. 102078, 2021.

[6] J. Wu, R. Gu, G. Dong, G. Wang, and S. Zhang, "FPL-UDA: Filtered pseudo label-based unsupervised cross-modality adaptation for vestibular schwannoma segmentation," in *ISBI*, 2022, pp. 1–5.

[7] Z. Wen, X. Zhang, and C. Ye, "Source-free domain adaptation for medical image segmentation via selectively updated mean teacher," in *IPMI*, 2023, pp. 225–236.

[8] X. Li, W. Chen, D. Xie, S. Yang, P. Yuan, S. Pu, and Y. Zhuang, "A free lunch for unsupervised domain adaptive object detection without source data," in *AAAI*, vol. 35, no. 10, 2021, pp. 8474–8481.

[9] Y. Sun, X. Wang, Z. Liu, J. Miller, A. Efros, and M. Hardt, "Test-time training with self-supervision for generalization under distribution shifts," in *ICML*, 2020, pp. 9229–9248.

[10] N. Karani, E. Erdil, K. Chaitanya, and E. Konukoglu, "Test-time adaptable neural networks for robust medical image segmentation," *Medical Image Analysis*, vol. 68, p. 101907, 2021.

[11] Y. He, A. Carass, L. Zuo, B. E. Dewey, and J. L. Prince, "Autoencoder based self-supervised test-time adaptation for medical image analysis," *Medical image analysis*, vol. 72, p. 102136, 2021.

[12] D. Wang, E. Shelhamer, S. Liu *et al.*, "Tent: Fully test-time adaptation by entropy minimization," in *ICLR*, 2021, pp. 1–15.

[13] S. Niu, J. Wu, Y. Zhang, Y. Chen, S. Zheng, P. Zhao, and M. Tan, "Efficient test-time model adaptation without forgetting," *arXiv preprint arXiv:2204.02610*, 2022.

[14] C. Li, X. Luo, W. Chen, Y. He, M. Wu, and Y. Tan, "Attent: Domain-adaptive medical image segmentation via attention-aware translation and adversarial entropy minimization," in *BIBM*, 2021, pp. 952–959.

[15] X. Liu, F. Xing, C. Yang, G. El Fakhri, and J. Woo, "Adapting off-the-shelf source segmenter for target medical image segmentation," in *MICCAI*, 2021, pp. 549–559.

[16] X. Liu, F. Xing *et al.*, "Self-semantic contour adaptation for cross modality brain tumor segmentation," in *ISBI*, 2022, pp. 28–31.

[17] G. Wang, X. Luo, R. Gu, S. Yang, Y. Qu, S. Zhai, Q. Zhao, K. Li, and S. Zhang, "Pymic: A deep learning toolkit for annotation-efficient medical image segmentation," *Computer Methods and Programs in Biomedicine*, vol. 231, p. 107398, 2023.

[18] Y. Shi, J. Zhang, T. Ling, J. Lu, Y. Zheng, Q. Yu, L. Qi, and Y. Gao, "Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 41, no. 3, pp. 608–620, 2021.

[19] X. Luo, G. Wang, W. Liao, J. Chen, T. Song, Y. Chen, S. Zhang, D. N. Metaxas, and S. Zhang, "Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency," *Medical Image Analysis*, vol. 80, p. 102517, 2022.

[20] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2, 2013, p. 896.

[21] P. Teja S and F. Fleuret, "Uncertainty reduction for model adaptation in semantic segmentation," in *CVPR*, 2021, pp. 9613–9623.

[22] H. Kingetsu, K. Kobayashi, Y. Okawa, Y. Yokota, and K. Nakazawa, "Multi-step test-time adaptation with entropy minimization and pseudo-labeling," in *ICIP*, 2022, pp. 4153–4157.

[23] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang *et al.*, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical Image Analysis*, vol. 63, p. 101693, 2020.

[24] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *CVPR*, 2017, pp. 7167–7176.

[25] R. Dorent, A. Kujawa, M. Ivory *et al.*, "Crossmoda 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation," *Medical Image Analysis*, vol. 83, p. 102628, 2023.

[26] X. Xu, Y. Chen, J. Wu, J. Lu, Y. Ye, Y. Huang, X. Dou, K. Li, G. Wang, S. Zhang, and W. Gong, "A novel one-to-multiple unsupervised domain adaptation framework for abdominal organ segmentation," *Medical Image Analysis*, vol. 88, p. 102873, 2023.

[27] J. Wu, D. Guo, L. Wang, S. Yang, Y. Zheng, J. Shapey, T. Vercauteren, S. Bisdas, R. Bradford, S. Saeed *et al.*, "Tiss-net: Brain tumor image synthesis and segmentation using cascaded dual-task networks and error-prediction consistency," *Neurocomputing*, vol. 544, p. 126295, 2023.

[28] C. Chen, Q. Dou, H. Chen, J. Qin, and P. A. Heng, "Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 7, pp. 2494–2505, 2020.

[29] F. Wu and X. Zhuang, "Unsupervised Domain Adaptation with Variational Approximation for Cardiac Segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3555–3567, 2021.

[30] C. Chen, Q. Dou, H. Chen, J. Qin, and P.-A. Heng, "Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation," in *AAAI*, 2019, pp. 865–872.

[31] C. Yang, X. Guo, Z. Chen, and Y. Yuan, "Source free domain adaptation for medical image segmentation with fourier style mining," *Medical Image Analysis*, vol. 79, p. 102457, 2022.

[32] Z. Nado, S. Padhy, D. Sculley, A. D'Amour, B. Lakshminarayanan, and J. Snoek, "Evaluating prediction-time batch normalization for robustness under covariate shift," *arXiv preprint arXiv:2006.10963*, 2020.

[33] M. Hu, T. Song, Y. Gu, X. Luo, J. Chen, Y. Chen, Y. Zhang, and S. Zhang, "Fully test-time adaptation for image segmentation," in *MICCAI*, 2021, pp. 251–260.

[34] T. Varsavsky, M. Orbes-Arteaga, C. H. Sudre, M. S. Graham, P. Nachev, and M. J. Cardoso, "Test-time unsupervised domain adaptation," in *MICCAI*, 2020, pp. 428–436.

[35] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[36] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, pp. 203–211, 2021.

[37] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *CVPR*, 2021, pp. 2613–2622.

[38] F. Milletari1, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *International Conference on 3D Vision*, 2016, pp. 565–571.

[39] J. Lee, D. Jung, J. Yim, and S. Yoon, "Confidence score for source-free unsupervised domain adaptation," in *ICML*, 2022, pp. 12 365–12 377.

[40] D. Tomar, G. Vray, J.-P. Thiran, and B. Bozorgtabar, "Opttta: Learnable test-time augmentation for source-free medical image segmentation under domain shift," in *MIDL*, 2021, pp. 1–26.

[41] V. M. Campello, P. Gkontra, C. Izquierdo, C. Martín-Isla, A. Sojoudi *et al.*, "Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3543–3554, 2021.

[42] K. Payette, H. Li, P. de Dumast, R. Licandro, H. Ji *et al.*, "Fetal brain tissue annotation and segmentation challenge results," *Medical Image Analysis*, p. 102833, 2023.

[43] S. Tourbier, X. Bresson, P. Hagmann, J.-P. Thiran *et al.*, "An efficient total variation algorithm for super-resolution in fetal brain MRI with adaptive regularization," *NeuroImage*, vol. 118, pp. 584–597, 2015.

[44] M. Kuklisova-Murgasova, G. Quaghebeur, M. A. Rutherford, J. V. Hajnal, and J. A. Schnabel, "Reconstruction of fetal brain mri with intensity matching and complete outlier removal," *Medical image analysis*, vol. 16, no. 8, pp. 1550–1564, 2012.

[45] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *MICCAI*, 2016, pp. 424–432.

[46] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "Cotr: Efficiently bridging CNN and transformer for 3D medical image segmentation," in *MICCAI*, 2021, pp. 171–180.