



# UPL-TTA: Uncertainty-Aware Pseudo Label Guided Fully Test Time Adaptation for Fetal Brain Segmentation

Jianghao Wu<sup>1,2</sup> , Ran Gu<sup>1</sup> , Tao Lu<sup>3</sup>, Shaoting Zhang<sup>1,2</sup>,  
and Guotai Wang<sup>1,2</sup>  

<sup>1</sup> School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

[guotai.wang@uestc.edu.cn](mailto:guotai.wang@uestc.edu.cn)

<sup>2</sup> Shanghai Artificial Intelligence Laboratory, Shanghai 200030, China

<sup>3</sup> Department of Radiology, Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu 610072, China

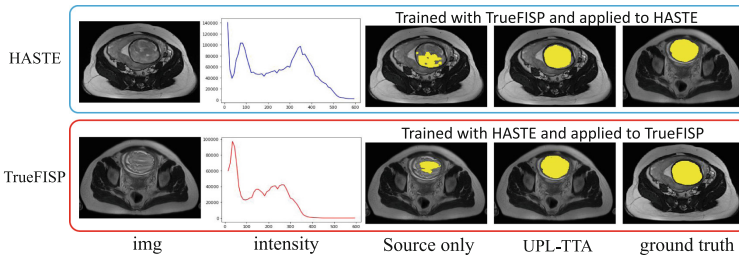
**Abstract.** Test Time Adaptation (TTA) is promising to improve a deep learning model's robustness when encountering images from an unseen domain. Existing TTA methods are with low performance due to the insufficient supervision signal from unannotated target domain images, or limited by specific requirements on the pre-training strategy and network structure in the source domain. We aim to separate the pre-training in the source domain and adaptation in the target domain, in order to achieve high-performance and more generalizable TTA without assumptions on the pre-training strategy. To solve this problem, we propose UPL-TTA, an Uncertainty-aware Pseudo Label guided fully Test Time Adaptation method. Specifically, we introduce Test Time Growing (TTG) to duplicate the prediction head of the source model with perturbations at image and feature levels in the target domain. The different predictions obtained in these duplicated prediction heads are used to obtain pseudo labels for the unlabeled target domain images as well as their uncertainty maps, which can identify reliable pseudo labels. Pixels with unreliable pseudo labels are regularized by imposing entropy minimization on the mean prediction of the multiple heads. UPL-TTA was validated bidirectionally on a cross-modality fetal brain segmentation dataset. Compared with no adaptation, it significantly improved the average Dice in the two different target domains by 3.95% and 6.12%, respectively, and outperformed several state-of-the-art TTA methods.

**Keywords:** Test time adaptation · self-training · fetal brain MRI

## 1 Introduction

Benefiting from high-precision and large-scale annotations, deep learning with Convolutional Neural Networks (CNNs) has achieved excellent performance in

medical image segmentation tasks [11]. However, due to the low cross-domain generalizability of existing methods, their performance will decrease largely when applied to images with a new distribution, i.e., an unseen modality [3]. For example, in the practical application, a pre-trained model can hardly maintain robustness when deployed to a new medical center where the data distribution may be different from the training set due to different scanning instruments used or different imaging sequences [5, 7, 8, 20]. Figure 1 shows such an example, where the image intensity and contrast are quite different in two sequences of fetal brain Magnetic Resonance Imaging (MRI): half-Fourier acquisition single-shot turbo spin-echo (HASTE) and true fast imaging with steady state precession (TrueFISP). A model trained with HASTE images has a poor performance on TrueFISP images, and vice versa.

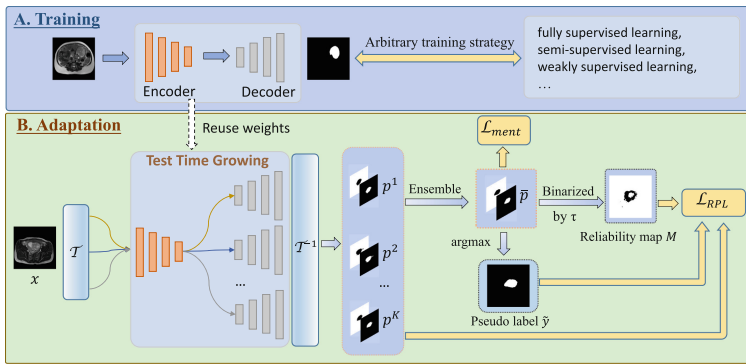


**Fig. 1.** The domain shift between HASTE and TrueFISP of fetal brain MRI. Our UPL-TTA largely improves the model’s robustness on a different sequence at testing time.

Domain Adaptation (DA) is promising to solve the above problem of domain gap between training and testing data [1]. To avoid time-consuming annotations in the target domain, Unsupervised Domain Adaptation (UDA) [14] methods are proposed to align the source and target distributions at image, feature, or output levels [22]. These methods all require simultaneous access to source and target domain data to make the model perform well. However, in practice, source data is often unavailable when the model is deployed to a new center due to the constraints on computation, bandwidth and privacy.

Source-free Domain Adaptation [10, 17, 19] aims to adapt a pre-trained model to a new target data distribution without access to the source data. In the literature, Test Time Training (TTT) [17] adds an auxiliary branch to predict the rotation by self-supervision, and adapts the shared encoder in the target domain. DTTA [8] and ATTA [5] optimize an auto-encoder during the training of source model to learn shape priors, and align feature distributions for adaptation. However, these methods require the insertion of specific modules, such as an auxiliary prediction branch and auto-encoders, before the training of the source model. They also require that the model should have been pre-trained with a specific strategy in the source domain, which limits their applicability when dealing with a pre-trained model that does not satisfy the training requirements.

In practice, a target domain may be given a pre-trained model that has been trained with an arbitrary strategy. Therefore, it is desirable to achieve fully test time adaptation that does not need the pre-trained model to have a specific structure and training strategy in the source domain [6]. PTBN [12] updates the statistics of Batch Normalization (BN) layers on the target domain data, and TENT [20] tunes BN layers by minimizing the entropy of predictions in the target domain. However, these methods were originally designed for natural images, and they simply assume that the domain shift can be sufficiently alleviated by updating the BN layers, which leads to limited performance in TTA for medical image segmentation [18]. URMA [2] is a method that aids the adaptation process using pseudo labels generated in one branch and uncertainties from multiple branches. However, its pseudo label may contain obvious errors and mislead the model adaptation.



**Fig. 2.** Overview of our UPL-TTA, where the  $p^k$  is the soft prediction of  $k$ -th branch,  $\tau$  is the confidence threshold. It does not require a specific training strategy in the source domain, and uses pseudo labels based on test time growing for adaptation.

In this work, we propose Uncertainty-aware Pseudo Label guided Fully Test Time Adaptation (UPL-TTA) for medical image segmentation, which does not require the pre-trained model to be trained with an extra auxiliary branch or a specific strategy in the source domain before adaptation to a target domain. For a given pre-trained model, we first introduce Test Time Growing (TTG) to duplicate the prediction head (e.g., the decoder in widely used UNet-like CNNs [15]) of the source model several times for the target domain, and add a range of random perturbations (e.g., dropout, spatial transform) to their input image and feature map to obtain several different segmentation predictions. Then pseudo labels for target domain images are obtained by an ensemble of these predictions. To suppress the effect of potentially incorrect pseudo labels, we introduce ensembling-based and MC dropout uncertainty estimation to obtain a reliability map. The pseudo labels of reliable pixels are used to supervise the output of each prediction head, and the predictions of unreliable pixels are regularized

by entropy minimization on the average prediction map. Experiments on bidirectional cross-modality adaptation between HASTE and TrueFISP of the fetal brain showed that our UPL-TTA significantly improved the model’s performance on the target domain, and outperformed several existing TTA methods.

## 2 Method

The proposed UPL-TTA framework is depicted in Fig. 2. Without assumptions on the training strategy in the source domain, we duplicate the prediction head of the pre-trained model several times and add perturbations to obtain multiple predictions, which leads to pseudo labels in the unannotated target domain and the corresponding reliability maps to supervise the model for adaptation.

### 2.1 Pre-trained Model from the Source Domain

Let  $S$  with data distribution  $\mu_S(\mathbf{x})$  be the source domain and  $T$  with data distribution  $\mu_T(\mathbf{x})$  be the target domain. Let  $\mathbf{X}_S = \{(\mathbf{x}_i^s, y_i^s), i = 1, \dots, N_s\}$  be the training images and their labels in the source domain, and  $\mathbf{X}_T = \{(\mathbf{x}_i), i = 1, \dots, N_t\}$  represent unlabeled images in the target domain for adaptation. Note that  $\mu_S(\mathbf{x}) \neq \mu_T(\mathbf{x})$ . The pre-training stage in the source domain is represented as:

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

where  $g$  and  $h$  are the feature extractor and prediction head of a segmentation network, respectively.  $\theta_g^0$  and  $\theta_h^0$  are their trained weights, respectively.  $L_s$  is the training 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.

### 2.2 Test-Time Growing for Adaptation

When the source model is deployed to a new center, as access to the source domain data is limited, we consider the problem of adapting the pre-trained model to the target domain based on  $\mathbf{X}_T$  and  $\{\theta_g^0, \theta_h^0\}$ . For the pre-trained feature extractor  $g$  and prediction head  $h$ , we propose Test-Time Growing (TTG) to duplicate  $h$  by  $K$  times in the target domain, as shown in Fig. 2. The weights of shared feature extractor  $g$  and duplicated prediction heads  $\{h^k\}(k = 1, 2, \dots, K)$  are initialized as  $\theta_g^0$  and  $\theta_h^0$ , respectively. Note that the weights in different prediction heads will become different due to the inconsistency of the gradients generated by the different predictions under perturbations. An ensemble of these  $K$  heads is used to obtain pseudo labels of target domain images that are unannotated. To encourage the different heads to obtain diverse results for better ensemble, we introduce random perturbations on the input image and dropout on features [21].

First, for an input image  $\mathbf{x} \in \mathcal{R}^{H \times W}$  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:

$$\mathbf{p}^k = \mathcal{T}^{-1} \circ h^k(g(\mathcal{T} \circ \mathbf{x})) \quad (2)$$

where  $\mathcal{T}$  is a random spatial transformation and  $\mathcal{T}^{-1}$  is the corresponding inverse transformation.  $\mathbf{p}^k \in \mathcal{R}^{C \times H \times W}$  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  $\mathcal{T}$  as random flipping, rotation with  $\pi/2$ ,  $\pi$  and  $3\pi/2$ .

Second,  $K$  different dropout layers are applied in parallel after the feature extractor  $g$ , so that the prediction heads take different random subsets of the features as input. We then average across the  $K$  different predicted segmentation probability maps for ensemble:

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

### 2.3 Supervision with Reliable Pseudo Labels

Based on the average probability map  $\bar{\mathbf{p}}$ , a pseudo label is obtained by taking the argmax across channels. To reduce noises, it is post-processed by only keeping the largest connected component for each foreground class (e.g., fetal brain segmentation in this work). Then the post-processed pseudo label is converted into a one-hot representation, which is denoted as  $\tilde{\mathbf{y}} \in \{0, 1\}^{C \times H \times W}$ . Due to the existence of domain gap, the pseudo labels have a limited accuracy. Directly using the pseudo labels of all pixels for self-training would limit the model's performance.

To deal with these problem, it is important to highlight the reliable pseudo labels and suppress unreliable ones during adaptation. Therefore, we use the uncertainty information of  $\bar{\mathbf{p}}$  to identify pixels with reliable pseudo labels and only use the reliable region to supervise the model for adaptation. Specifically, a binary reliability map  $M \in \{0, 1\}^{H \times W}$  is calculated for the pseudo label  $\tilde{\mathbf{y}}$ , and each element in  $M$  is defined as:

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

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

Then the reliability map is used as a mask to exclude unreliable pixels for supervision, and a Reliable Pseudo Label (RPL) loss is denoted as:

$$\mathcal{L}_{RPL} = \frac{1}{K} \sum_{k=1}^K \mathcal{L}_{w-dice}(\mathbf{p}^k, \tilde{\mathbf{y}}, M) \quad (5)$$

where  $\mathcal{L}_{w-dice}$  is the reliability map-weighted Dice loss for the  $k$ -th head:

$$\mathcal{L}_{w-dice}(\mathbf{p}^k, \tilde{y}, M) = 1 - \frac{1}{Z} \sum_{c=1}^C \sum_{n=1}^{HW} \frac{2M_n \mathbf{p}_{c,n}^k \tilde{y}_{c,n}}{\mathbf{p}_{c,n}^k + \tilde{y}_{c,n} + \epsilon} \quad (6)$$

where  $n$  is the pixel index and  $\epsilon = 10^{-5}$  is a small number for numeric stability.  $Z = C \sum_n M_n$  is a normalization factor.

## 2.4 Mean Prediction-Based Entropy Minimization

Entropy minimization is widely used as a regularizer in test time adaptation [9, 13, 18], which reduces the uncertainty of the system by reducing the entropy of model predictions. However, in our method with multiple prediction heads, applying entropy minimization to each head respectively may lead to sub-optimal results when different heads predict confident while opposite results. For example, in the binary segmentation problem, when branch  $k$  predicts a certain pixel being the foreground with a probability of 0.0 and branch  $k + 1$  predicts it with a foreground probability of 1.0, both branches have the lowest entropy, but their average result has a high entropy. To deal with this problem, we propose to apply entropy minimization to the mean prediction across the  $K$  heads:

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

where  $\bar{\mathbf{p}}$  is the mean probability prediction obtained by the  $K$  heads of TTG. Compared with minimizing the entropy of each prediction head respectively, minimizing the entropy of their mean prediction  $\bar{\mathbf{p}}$  can not only reduce the uncertainty of a single prediction head, but also make the predictions of the  $K$  heads for the same test sample tend to be consistent, therefore improving the prediction robustness of the model for unseen test samples.

## 2.5 Adaptation by Self-training

Our adaptation process adopts a self-training paradigm based on the pseudo labels and mean prediction-based entropy minimization. We obtain the average prediction  $\bar{\mathbf{p}}$ , pseudo label  $\tilde{y}$  and the reliability map  $M$  based TTG for a test sample, and then calculate  $\mathcal{L}_{RPL}$  and  $\mathcal{L}_{ment}$ . The overall loss for TTA is:

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

where  $\lambda$  is a hyper-parameter to control the weight of  $\mathcal{L}_{ment}$ .

### 3 Experiment and Results

#### 3.1 Experimental Details

**Dataset.** We used a Fetal Brain (FB) segmentation dataset to evaluate our UPL-TTA, and it consisted of fetal brain MRI with two imaging sequences: 1) 68 volumes acquired by HASTE with size of  $640 \times 520$ , in-plane resolution of 0.64 to 0.70 mm and slice-thickness of 6.5–7.15 mm; 2) 44 volumes acquired by TrueFISP with size of  $384 \times 312$ , in-plane resolution of 0.67 to 1.12 mm and thickness of 6.5 mm. The gestational age ranged from 21–33 weeks. As shown in Fig. 1, the intensity distribution and contrast are different in these two sequences, leading to a large domain gap. In addition, the different gestational age leads to varying appearance of the fetal brain, which increases the difficulty for robust segmentation. We performed bidirectional TTA for experiments: 1) HASTE to TrueFISP, where HASTE was used as the source domain and TrueFISP as the target domain; 2) TrueFISP to HASTE. We randomly split the images for each domain into 70%, 10% and 20% for training, validation and testing, respectively, and abandoned the labels of training images in the target domain.

**Implementation Details.** For preprocessing, we clip the intensities by the 1-st and 99-th percentiles, and linearly normalized them to  $[-1, 1]$ . Each slice was resized to  $256 \times 256$ . Due to the large inter-slice spacing, we used slice-by-slice segmentation with 2D CNNs and stacked the results into a 3D volume. The segmentation network for our method is flexible, and we selected the widely used UNet [15], as most medical image segmentation models are based on UNet-like structures [11, 16]. The feature extractor  $g$  and prediction head  $h$  were implemented by the encoder and decoder of UNet [15], respectively. During pre-training in the source domain, we trained UNet [15] 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 weight with the best performance on the validation set in the source domain was used for adaptation. For adaptation in the target domain, we duplicated the decoder of UNet [11] for  $K$  times, and updated all the model parameters for 20 epochs with Adam optimizer and a fixed learning rate of  $10^{-4}$ . In the training and adaptation stages, we set all the slices in a single volume as a batch. The hyper-parameter setting was  $K = 4$ ,  $\lambda = 1.0$ , and  $\tau = 0.9$  based on the best performance on the validation set.

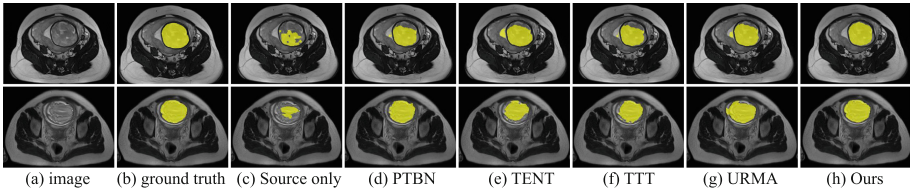
During inference, we computed the argmax of the average prediction generated by the  $K$  heads, and we did not apply any post-processing to the output. All the experiments were implemented with PyTorch 1.8.1, using an NVIDIA GeForce RTX 2080Ti GPU. For quantitative evaluation of the volumetric segmentation results, we adopted the commonly used Dice score (DSC) and Average Symmetric Surface Distance (ASSD). As the slice thickness is large (6–7.15 mm), we calculated ASSD values with unit of pixel.

#### 3.2 Results

**Comparison with Other Methods.** Our UPL-TTA was compared with four state-of-the-art test time adaptation methods on the FB dataset: 1) **PTBN** [12]

**Table 1.** Quantitative comparison of different TTA methods on fetal brain segmentation. † means significant improvement ( $p$ -value  $< 0.05$ ) from “Source only”.

Method	HASTE to TrueFISP		TrueFISP to HASTE	
	Dice (%)	ASSD (pixel)	Dice (%)	ASSD (pixel)
Source only	84.09 $\pm$ 6.34	1.33 $\pm$ 0.49	83.91 $\pm$ 7.39	2.31 $\pm$ 1.96
Target only	88.85 $\pm$ 4.12	0.91 $\pm$ 0.30	94.09 $\pm$ 3.47	0.50 $\pm$ 0.38
PTBN [12]	85.70 $\pm$ 4.88	1.85 $\pm$ 0.96	85.47 $\pm$ 5.65	2.92 $\pm$ 2.55
TENT [20]	85.75 $\pm$ 3.62	1.60 $\pm$ 0.71	88.21 $\pm$ 5.35†	1.16 $\pm$ 1.17
TTT [17]	85.84 $\pm$ 4.52	1.80 $\pm$ 0.90	87.20 $\pm$ 5.33	2.28 $\pm$ 1.99
URMA [2]	84.12 $\pm$ 6.82	2.18 $\pm$ 1.19	81.05 $\pm$ 6.85	5.95 $\pm$ 3.9
UPL-TTA (Ours)	<b>88.04 <math>\pm</math> 4.82†</b>	<b>1.20 <math>\pm</math> 0.74</b>	<b>90.03 <math>\pm</math> 5.28†</b>	<b>0.85 <math>\pm</math> 0.64†</b>

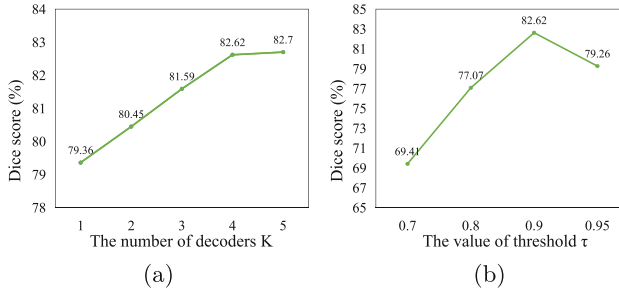
**Fig. 3.** Qualitative comparison of different TTA methods. First row: HASTE to TrueFISP. Second row: TrueFISP to HASTE.

that updates batch normalization statistics on the target data during test time; 2) **TENT** [20] that only updates the parameters of batch normalization layers by minimizing the entropy of model predictions on new test data; 3) **TTT** [17] that uses self-supervision for adaptation, where an auxiliary decoder is used in both of the source and target domains to predict the rotation angle of an image; and 4) **URMA** [2] that uses pseudo labels generated in one branch and uncertainties in multiple branches to aid the adaptation process. We also compared our method with two oracle methods: 1) **Source only** where the pre-trained model was directly used for inference on the target domain dataset, and 2) **Target only** where the model was trained with annotated images in the target domain. All the compared methods were implemented with the same backbone (UNet [15]) for a fair comparison.

The quantitative evaluation results of bidirectional TTA are shown in Table 1. It can be observed that **Source only** and **Target only** achieved an average Dice of 84.09% and 88.85%, respectively in HASTE to TrueFISP and 83.90% and 94.09%, respectively in TrueFISP to HASTE, showing the large gap between the two domains. The existing methods only achieved a slight improvement or even a decrease compared with **Source only**, with the Dice values ranging from 84.12% to 85.84% for HASTE to TrueFISP and 81.05% to 88.21% for TrueFISP to HASTE, respectively. In contrast, our method largely improved the Dice to 88.04% and 90.03% for the two target domains, respectively. Our method



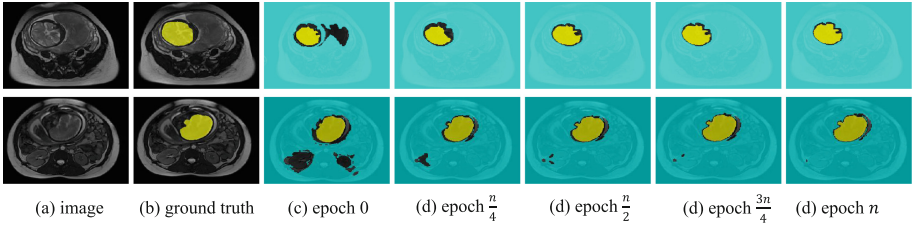
achieved an average ASSD of 1.20 and 0.85 pixels, in the two domains, respectively, which was lower than those of the other TTA methods. The qualitative comparison in Fig. 3 shows that the existing methods tend to achieve under-segmentation of the fetal brain, while our method can successfully segment the entire fetal brain region with high accuracy.



**Fig. 4.** Performance of our method with different hyper-parameter values on the validation set when HASTE and TrueFISP are the source and target domains, respectively.

**Ablation Study.** Our UPL-TTA adds two new hyperparameters: the number of duplicated prediction heads  $K$ , and the confidence threshold  $\tau$  to select reliable pseudo labels. We first investigated the effect of  $K$  by setting it to 1 to 5 respectively, and the performance on the validation set of TrueFISP is shown in Fig. 4(a). It can be observed that  $K = 1$  performed worse than larger  $K$  values, showing the superiority of using Test-Time Growing (TTG). As  $K$  increased, our method performed progressively better, and  $K = 5$  reached a plateau. Therefore, we finally set  $K$  to 4 considering the trade-off between performance and memory consumption. Then we investigated the effect of  $\tau$ . A higher threshold  $\tau$  will result in a smaller reliable region for each class, which helps avoid the model being misled by inaccurate pseudo label, but a too large  $\tau$  will make the reliable pseudo label region too small and thus cannot provide sufficient supervision. Quantitative comparison between different  $\tau$  values in Fig. 4(b) shows that the best performance on the validation set was achieved when  $\tau = 0.9$ .

We further investigated the effect of each component of our UPL-TTA. The baseline was just using the pre-trained model’s predictions as pseudo labels for adaptation, and the introduced components are: 1) single-head entropy minimization (Entropy-min) [4] that is applied to each of the prediction heads respectively; 2) “Reliability map” that uses  $M$  to suppress unreliable pseudo labels; 3) Test Time Growing (TTG) that duplicates the prediction head  $K$  times with feature dropout; 4) random spatial transformation ( $T$ ) further introduced to the  $K$  heads; and 5)  $L_{ment}$  that applies entropy minimization to the mean prediction of the  $K$  heads rather than to each head respectively. The quantitative evaluation results are presented in Table 2. We observed that the baseline (74.98%) performed worse than “Source only” (84.09%). Additionally using entropy minimization (83.44%) was still not better than “Source only”, which indicated



**Fig. 5.** Pseudo labels at different training steps in self-training. Epoch 0 means “Source only” (before adaptation) and  $n$  is the optimal epoch number on the validation set of the target domain. In (c)–(g), only reliable pseudo labels are encoded by colors. (Color figure online)

**Table 2.** Ablation study of the proposed method. HASTE and TrueFISP were used as the source and target domains, respectively. The baseline was just using the pre-trained model’s predictions as pseudo labels for adaptation. Entropy-min: Entropy minimization for each prediction head.  $L_{ment}$  is entropy minimization on the average prediction of the  $K$  heads.

	Components				Dice (%)	ASSD (pixel)
	Reliability-map	TTG	T	$L_{ment}$		
					$74.98 \pm 10.72$	$4.9 \pm 1.37$
✓					$83.44 \pm 7.38$	$1.39 \pm 0.76$
✓	✓				$85.88 \pm 5.49$	$1.56 \pm 1.01$
✓	✓	✓			$86.77 \pm 4.17$	$0.95 \pm 0.24$
✓	✓	✓	✓		$86.92 \pm 5.41$	<b><math>0.91 \pm 0.36</math></b>
	✓	✓	✓	✓	<b><math>88.04 \pm 4.82</math></b>	$1.20 \pm 0.74$

that the pseudo label from a single prediction head contained a lot of misleading information. In contrast, each component of our introduced reliability map, TTG, spatial transformation and  $L_{ment}$  led to some improvement, showing the effectiveness of our method.

## 4 Discussions

In general, a segmentation model contains a feature extractor and a prediction head, and our method duplicates the prediction head via Test-Time Growing (TTG) in the target domain. This paper implemented TTG with an encoder-decoder structure, as most efficient CNNs for medical image segmentation tasks are UNet-like [3, 15]. However, our method can be easily applied to other segmentation networks, as it has a minimal assumption on the structure of the pre-trained model and how it was trained in the source domain, which is more general than existing methods like TTT [17] and DTTA [8].

Due to the absence of annotations in the target domain, it is important to obtain effective supervision signal and regularization for the TTA task. Our

method uses reliable pseudo labels to deal with the unannotated images, where the TTG improves the quality of pseudo labels, and the introduced reliability map avoids the model being corrupted by inaccurate pseudo labels. URMA [2] also uses pseudo labels to guide the adaptation, but its pseudo labels are obtained from a single decoder, which are less robust than our pseudo labels based on an ensemble of multiple heads. In addition, our mean prediction-based entropy minimization has an implicit consistency regularization on the  $K$  prediction heads, which improves the model’s robustness against perturbations in the target domain.

Despite UPL-TTA’s higher performance than existing TTA methods in the experiment, it is applicable to a moderate domain shift where high-quality pseudo labels can be obtained by TTG. In some other scenarios where the domain gap is extremely large, it may be hard to obtain usable pseudo labels, and our method may not be applicable. In addition, this work only deals with a binary segmentation task, but the pipeline can also be applied for multi-class segmentation and 3D segmentation networks.

## 5 Conclusion

To summarize, we propose a fully test time adaptation method that adapts the source model to an unannotated target domain without knowing the training strategy of the source model. Without access to source domain images, our proposed uncertainty-aware pseudo label-guided TTA generates multiple prediction outputs for the same sample in the target domain via Test Time Growing (TTG). It generates high-quality pseudo labels and the corresponding reliability maps to provide effective supervision in the unannotated target domain. Pixels with unreliable pseudo labels are further regularized by entropy minimization of the mean prediction across the duplicated heads, which also introduces an implicit consistency regularization. Experiments on bidirectionally cross-modality TTA for fetal brain segmentation showed that our method outperformed several state-of-the-art TTA methods. In the future, it is of interest to implement a 3D version of our method and apply it to other segmentation tasks.

**Acknowledgements.** This work was supported by National Natural Science Foundation of China (No. 62271115).

## References

1. Chen, C., Dou, Q., Chen, H., Qin, J., Heng, P.A.: Synergistic image and feature adaptation: towards cross-modality domain adaptation for medical image segmentation. In: AAAI, pp. 865–872 (2019)
2. Fleuret, F., et al.: Uncertainty reduction for model adaptation in semantic segmentation. In: CVPR, pp. 9613–9623 (2021)
3. Gu, R., Zhang, J., Huang, R., Lei, W., Wang, G., Zhang, S.: Domain composition and attention for unseen-domain generalizable medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 241–250. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-87199-4\\_23](https://doi.org/10.1007/978-3-030-87199-4_23)

4. Hang, W., et al.: Local and global structure-aware entropy regularized mean teacher model for 3D left atrium segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 562–571. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-59710-8\\_55](https://doi.org/10.1007/978-3-030-59710-8_55)
5. He, Y., Carass, A., Zuo, L., Dewey, B.E., Prince, J.L.: Autoencoder based self-supervised test-time adaptation for medical image analysis. *Med. Image Anal.* **72**, 102136 (2021)
6. Hu, M., et al.: Fully test-time adaptation for image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 251–260. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-87199-4\\_24](https://doi.org/10.1007/978-3-030-87199-4_24)
7. Karani, N., Chaitanya, K., Baumgartner, C., Konukoglu, E.: A lifelong learning approach to brain MR segmentation across scanners and protocols. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 476–484. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-00928-1\\_54](https://doi.org/10.1007/978-3-030-00928-1_54)
8. Karani, N., Erdil, E., Chaitanya, K., Konukoglu, E.: Test-time adaptable neural networks for robust medical image segmentation. *Med. Image Anal.* **68**, 101907 (2021)
9. Lee, J., Jung, D., Yim, J., Yoon, S.: Confidence score for source-free unsupervised domain adaptation. In: ICML, pp. 12365–12377. PMLR (2022)
10. Li, X., et al.: A free lunch for unsupervised domain adaptive object detection without source data. In: AAAI, vol. 35, pp. 8474–8481 (2021)
11. Litjens, G., et al.: A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017)
12. Nado, Z., Padhy, S., Sculley, D., D’Amour, A., Lakshminarayanan, B., Snoek, J.: Evaluating prediction-time batch normalization for robustness under covariate shift. arXiv preprint [arXiv:2006.10963](https://arxiv.org/abs/2006.10963) (2020)
13. Niu, S., Wu, J., Zhang, Y., Chen, Y., Zheng, S., Zhao, P., Tan, M.: Efficient test-time model adaptation without forgetting. arXiv preprint [arXiv:2204.02610](https://arxiv.org/abs/2204.02610) (2022)
14. Pei, C., Wu, F., Huang, L., Zhuang, X.: Disentangle domain features for cross-modality cardiac image segmentation. *Med. Image Anal.* **71**, 102078 (2021)
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-24574-4\\_28](https://doi.org/10.1007/978-3-319-24574-4_28)
16. Singh, R., Rani, R.: Semantic segmentation using deep convolutional neural network: a review. In: ICICC (2020)
17. Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., Hardt, M.: Test-time training with self-supervision for generalization under distribution shifts. In: International Conference on Machine Learning, pp. 9229–9248. PMLR (2020)
18. Tomar, D., Vray, G., Thiran, J.P., Bozorgtabar, B.: OptTTA: learnable test-time augmentation for source-free medical image segmentation under domain shift. In: Medical Imaging with Deep Learning (2021)
19. Varsavsky, T., Orbes-Arteaga, M., Sudre, C.H., Graham, M.S., Nachev, P., Cardoso, M.J.: Test-time unsupervised domain adaptation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 428–436. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-59710-8\\_42](https://doi.org/10.1007/978-3-030-59710-8_42)
20. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: fully test-time adaptation by entropy minimization. In: ICLR (2021)

21. Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T.: Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing* **338**, 34–45 (2019)
22. Wu, J., Gu, R., Dong, G., Wang, G., Zhang, S.: FPL-UDA: filtered pseudo label-based unsupervised cross-modality adaptation for vestibular schwannoma segmentation. In: ISBI, pp. 1–5. IEEE (2022)