# IPLC: Iterative Pseudo Label Correction Guided by SAM for Source-Free Domain Adaptation in Medical Image Segmentation

Guoning Zhang[1], Xiaoran Qi[2], Bo Yan[1(✉)], and Guotai Wang[3,4(✉)]

[1] School of Information Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China
yanboyu@uestc.edu.cn
[2] School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, China
[3] School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China
[4] Shanghai Artificial Intelligence Laboratory, Shanghai, China
guotai.wang@uestc.edu.cn

**Abstract.** Source-Free Domain Adaptation (SFDA) is important for dealing with domain shift without access to source data and labels of target domain images for medical image segmentation. However, existing SFDA methods have limited performance due to insufficient supervision and unreliable pseudo labels. To address this issue, we propose a novel Iterative Pseudo Label Correction (IPLC) guided by the Segment Anything Model (SAM) SFDA framework for medical image segmentation. Specifically, with a pre-trained source model and SAM, we propose multiple random sampling and entropy estimation to obtain robust pseudo labels and mitigate the noise. We introduce mean negative curvature minimization to provide more sufficient constraints and achieve smoother segmentation. We also propose an Iterative Correction Learning (ICL) strategy to iteratively generate reliable pseudo labels with updated prompts for domain adaptation. Experiments on a public multi-site heart MRI segmentation dataset (M&MS) demonstrate that our method effectively improved the quality of pseudo labels and outperformed several state-of-the-art SFDA methods. The code is available at https://github.com/HiLab-git/IPLC.

**Keywords:** Source-Free Domain Adaptation · Segment Anything Model · Heart MRI.

## 1 Introduction

Deep learning has achieved excellent performance in medical image segmentation, which plays an essential role in clinical applications like computer-aided

G. Zhang and X. Qi—Contributed equally to this work.

diagnosis [11, 16, 17]. This success is highly dependent on the assumption that training and testing data follow the same distribution. However, the distribution gap (called domain shift) is widespread between training (i.e., source domain) and testing (i.e., target domain) datasets in real-world clinical scenarios due to different scanners, imaging protocols, image qualities, etc [8]. This gap between the source and target domains usually leads to drastic performance degradation when a model is deployed to an unseen target domain dataset [7].

Domain Adaptation (DA) that aims to address the domain shift between source and target domains has been attracting increasing attention [10]. However, obtaining expert-level annotations is time-consuming and expensive, therefore, Unsupervised Domain Adaptation (UDA) methods are proposed to address this issue with unlabeled target domain images and labeled source domain images [20, 22, 23]. Most existing UDA methods require access to source and target domain images simultaneously for adaptation [15, 20]. However, due to privacy and transmission concerns with medical images, the source domain images are usually unavailable, which limits the application of UDA methods. This motivates research of Source-Free Domain Adaptation (SFDA) methods that adapt a pre-trained model to target domains without accessing the source data [1, 3, 5].

Recently, a few SFDA methods have been proposed. PTBN [12] uses Batch Normalization (BN) with statistics recalculated on the batch at prediction time to mitigate the effect of domain shift. However, only updating the BN layers can not sufficiently address the domain shift, which leads to limited performance for SFDA. TENT [19] adapts the model by minimizing the entropy of its predictions in the target domain. SAR [14] uses a sharpness-aware and reliable optimization scheme to minimize the entropy and its sharpness for adaptation. These methods using entropy minimization as supervision can not provide sufficient constraints for adaptation, which easily leads to over-confident yet incorrect predictions. EATA [13] uses sample-adaptive identification and other methods [6, 18, 21, 24] use auxiliary branches and pseudo labels for adaptation, which does not consider the noise in pseudo labels, therefore, misleading the model adaptation. Existing SFDA methods have limited performance due to the effect of insufficient supervision and unreliable pseudo labels, thus, it is significant to provide sufficient constraints and obtain high-quality pseudo labels for domain adaptation. SAM-Med2D [4] is a fine-tuned SAM [9] with medical 2D images that yields satisfactory performance, which has the potential to generate reliable pseudo labels and provide sufficient supervision for SFDA.

In this work, we propose a novel Iterative Pseudo Label Correction (IPLC) guided by SAM framework to address the limitations of existing SFDA methods for medical image segmentation. Our contributions are summarized as follows: (1) With a pre-trained source model and SAM-Med2D, we propose Multiple Random Sampling (MRS) and Entropy Weight Estimation (EWE) to obtain robust pseudo labels and mitigate the noise of pseudo labels. Specifically, we simultaneously use point prompts obtained from MRS and mask prompts obtained from the model to generate pseudo labels through multiple inferences by SAM-Med2D, which improves the robustness of pseudo labels. We calculate

the entropy of predictions to weight pseudo labels for reliable supervision. (2) We introduce mean negative curvature minimization for the edges of the model predictions to provide more sufficient constraints and achieve smoother segmentation. (3) We also propose the Iterative Correction Learning (ICL) strategy to optimize the model with the supervision of pseudo labels and mean negative curvature, which iteratively generates reliable pseudo labels from SAM-Med2D with iteratively updated prompts from the model. This strategy can effectively improve the quality of pseudo labels and achieve performance improvements for source-free domain adaptation.

Extensive experiments on multi-site heart MRI segmentation showed that our method can effectively improve the quality of pseudo labels and adapt the model from a source domain to target domains. It outperformed several state-of-the-art SFDA methods for medical image segmentation.

## 2 Method

The proposed IPLC framework is depicted in Fig. 1. With a pre-trained model and SAM-Med2D, we simultaneously use point prompts obtained from MRS and mask prompts obtained from outputs of the model to generate pseudo labels through $k$-times inference by SAM-Med2D. Entropy-based weight maps are calculated to assign weights to the supervision of pseudo labels. To provide regularization to the model, we minimize the mean negative curvature, which is derived from the model's edge detection result. Reliable pseudo labels are iteratively refined using updated prompts, enhancing the quality of pseudo labels for adaptation.

### 2.1 Pre-trained Model and SAM-Med2D Model

Let $f_{\Theta^o}$ be the model pre-trained with source domain images, and $f_{\Theta^s}$ be SAM-Med2D model [4]. $\boldsymbol{\mathcal{D}} = \{(x_i,), i = 1, \ldots, N\}$ represent unlabeled images in the target domain. $\delta_S$ and $\delta_T$ are image distributions in the source and target domain, respectively. Note that $\delta_S \neq \delta_T$, and $f_{\Theta^s}$ is frozen.

### 2.2 Multiple Random Sampling and Entropy Weight Estimation

Existing SFDA methods have limited performance due to the noise of pseudo labels. Considering this issue, we propose Multiple Random Sampling (MRS) to generate robust pseudo labels for SFDA, as shown in Fig. 1. To mitigate the prediction noise of $f_{\Theta^o}$ and provide sufficient constraints for $f_{\Theta^s}$, we simultaneously use point prompts obtained from MRS and mask prompts obtained from outputs of $f_{\Theta^o}$ to generate reliable pseudo labels for adaptation.

Specifically, for an input image $x \in \mathbb{R}^{H \times W}$ in the target domain, where $H$ and $W$ are the height and width, respectively. We send it into $f_{\Theta^o}$ to generate a probability map $p_o \in \mathbb{R}^{C \times H \times W}$ with $C$ channels obtained by Softmax, where $C$ is the class number. Through argmax and one-hot conversion, we obtain the

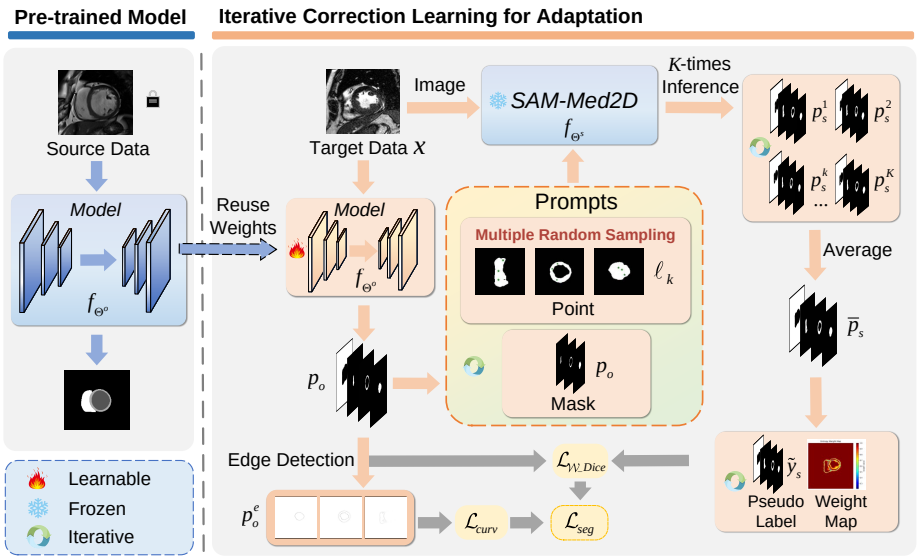


**Fig. 1.** Overview of our IPLC framework. We use $k$-th point prompts $\ell^k$ obtained from multiple random sampling and mask prompts $p_o$ to generate $p_s^k$ and then obtain pseudo labels $\tilde{y}_s$. $\mathcal{W}$ is the entropy-based weight map of $\tilde{y}_s$ obtained from $\bar{p}_s$. Based on the supervision of $\mathcal{L}_{\mathcal{W}\_Dice}$ and $\mathcal{L}_{curv}$, we use iterative correction learning for adaptation.

binary map of $p_o$. We random sample $M$ points for each foreground class in the binary map $K$ times to generate $K$ different 2D coordinates, which represent point prompts. $\ell^{k,c} \in [\mathbb{N}, \mathbb{N}]^{1\times M}$ are the coordinates of $k$-th set of point prompts for class $c$. We set the $c$-th channel of $p_o$ as mask prompts for $f_{\Theta^s}$. The $k$-th inference foreground probability map for class $c$ obtained by $f_{\Theta^s}$ is:

$$p_s^{k,c} = f_{\Theta^s}(x, p_o^c, \ell^{k,c}) \tag{1}$$

where $x$ is the target domain image, $p_o^c$ and $\ell^{k,c}$ are prompts for $f_{\Theta^s}$, $p_s^{k,c} \in \mathbb{R}^{H\times W}$. The $k$-th inference probability map $p_s^k$ for $x$ is obtained by concatenation for the $C$-channels probability map. We then average the $K$ different $p_s^k$ to obtain the mean probability map $\bar{p}_s$ and take the argmax across the class channel to obtain the pseudo label $\tilde{p}_s$:

$$\tilde{p}_s = \arg\max\left(\frac{1}{K}\sum_{k=1}^{K} p_s^k\right) \tag{2}$$

To mitigate the noise of $\tilde{p}_s$, we propose Entropy Weight Estimation (EWE) to weight pseudo labels with the entropy-based weight map of $\bar{p}_s$. The entropy-based weight map is given by:

$$\mathcal{W}(\bar{p}_s, C) = \frac{\log_2 C - (-\sum_{c=1}^{C} \bar{p}_s^c \cdot \log_2 \bar{p}_s^c)}{\log_2 C} \tag{3}$$

where $\log_2 C$ represents the maximum entropy of $\bar{p}_s$ with $C$ class segmentation. Through Eq.(3), the weight map is normalized to $[0, 1]$, where higher entropy values are converted to lower weights. Formally, the weight-based Dice loss is given by:

$$\mathcal{L}_{\mathcal{W}\_Dice}(p_o, \tilde{y}_s, \mathcal{W}) = 1 - \frac{1}{C}\sum_{c=1}^{C}\frac{\sum_{n=1}^{H\times W} 2\cdot(\mathcal{W}_n \cdot p_o^{c,n})\cdot(\mathcal{W}_n \cdot \tilde{y}_s^{c,n})}{\sum_{n=1}^{H\times W}\mathcal{W}_n\cdot(p_o^{c,n}+\tilde{y}_s^{c,n})+\sigma} \tag{4}$$

where $\tilde{y}_s \in \{0,1\}^{C\times H\times W}$ is a one-hot representation converted from $\tilde{p}_s$ and $\sigma = 10^{-5}$ is a small number for numeric stability.

### 2.3 Regularization based on Mean Negative Curvature

Pseudo labels $\tilde{y}_s$ have limited accuracy due to domain shift, directly using $\tilde{y}_s$ for supervision would limit the model's performance. Therefore, we introduce mean negative curvature minimization to improve performance and achieve smoother segmentation. The edge detection map of $p_o$ for class $c$ is given by:

$$p_o^{e,c} = p_o^c \otimes Sobel \tag{5}$$

where $\otimes$ represents $2D$ convolution operation, $Sobel$ is the edge detection operator, we apply $Sobel$ to each channel respectively and exclude the background component. The curvature of $p_o^{e,c}$ is given by:

$$V_{p_o^{e,c}} = \frac{(1+{U_y}^2)\cdot U_{xx} - 2\cdot U_x\cdot U_y\cdot U_{xy} + (1+{U_x}^2)\cdot U_{yy}}{2(1+{U_x}^2+{U_y}^2)^{3/2}} \tag{6}$$

where $U_x$ and $U_y$ denote the derivatives of $p_o^{e,c}$ along the x-axis and y-axis, respectively, and $U_{xx}$ denotes the derivative‘ of $U_x$ along the x-axis, etc. Loss of curvature is given by:

$$\mathcal{L}_{curv} = \sum_{c=1}^{C-1}\Big(\frac{1}{P}\sum_{n=1}^{P}\big(\mathrm{ReLU}(-V_{p_o^{e,c},n})\big)\Big) \tag{7}$$

where $P$ is the number of $V_{p_o^{e,c}}$ pixels, and $C-1$ is the number of foreground class. We use ReLU to retain the negative curvature of $V_{p_o^{e,c}}$. By minimizing the mean negative curvature, we make $V_{p_o^{e,c}}$ between neighboring pixels closer, which reduces the roughness of the $p_o^{e,c}$ and achieves a smoother segmentation result.

### 2.4 Iterative Correction Learning for Adaptation

We propose the Iterative Correction Learning (ICL) strategy to optimize $f_{\Theta^o}$ for domain adaptation. This strategy iteratively generates reliable pseudo labels $\tilde{y}_s$ from $f_{\Theta^s}$ with iteratively updated prompts from $f_{\Theta^o}$ and uses $\mathcal{L}_{\mathcal{W}\_Dice}$ and $\mathcal{L}_{curv}$ as supervision, which can effectively improve the quality of $\tilde{y}_s$ and achieve performance improvements for SFDA. The overall loss is summarized as:

$$\mathcal{L}_{seg} = \mathcal{L}_{\mathcal{W}\_Dice} + \alpha\mathcal{L}_{curv} \tag{8}$$

where $\alpha$ is the hyper-parameter that controls the weights of $\mathcal{L}_{curv}$.

## 3 Experiments and Results

### 3.1 Experimental Details

**Dataset and Metrics** Our experiments used the public Multi-Centre, Multi-Vendor, and Multi-Disease Cardiac Image Segmentation (M&MS) dataset [2]. The M&MS dataset was acquired from four different scanner vendors: Siemens (Domain A), Philips (Domain B), General Electric (Domain C), and Canon (Domain D), contains 192, 252, 150, and 100 cardiac MRI volumes, respectively. We used domain A as the source domain, and B, C, and D as the target domains. The target tissues for segmentation were the Left Ventricle (LV), Right Ventricle (RV), and Myocardium (MYO). In the target domains, we randomly divided the images into 70% for training, 10% for validation, and 20% for testing and discarded labels for the training set. For quantitative evaluation of the results, we adopted the commonly used Dice score and Average Symmetric Surface Distance (ASSD). We implemented slice-level segmentation and stacked all the slices in a volume into a 3D segmentation. As the slice thickness was large (9.2-10 mm), we calculated ASSD values with unit of pixel.

**Implementation Details** We adopted classic 2D UNet [16] to demonstrate the effectiveness of our method. The image intensity was linearly normalized to [-1, 1] and each slice was center-cropped to 256×256. With the pre-trained model $f_{\Theta^o}$ in the source domain and SAM-Med2D [4] model $f_{\Theta^s}$, we optimized $f_{\Theta^o}$ for 20 epochs with Adam optimizer and a fixed learning rate of $10^{-4}$ for adaptation. The segmentation class number was 4. The hyper-parameter $K$ and prompts setting were determined based on the training set of target domains, which aims to generate more reliable pseudo labels than $f_{\Theta^o}$ for adaptation. The hyper-parameter $\alpha$ was determined based on the validation set. Specifically, $K = 10$, $\alpha = 0.01$, and the prompts mode were points & mask. All experiments were implemented with PyTorch, using an NVIDIA GeForce RTX 3060 GPU.

### 3.2 Results of Source-Free Domain Adaptation

**Comparision with Other Methods** Our IPLC method was compared with four state-of-the-art methods on the M&MS dataset: 1) **PTBN** [12] that uses BN to mitigate the domain shift. 2) **TENT** [19] that uses entropy minimization to adapt the model. 3) **EATA** [13] that uses sample-adaptive identification to optimize the model. 4) **SAR** [14] that uses a sharpness-aware and reliable optimization scheme to minimize the entropy for adaptation. We also compared our method with two oracle methods: 1) **Source only** that the pre-trained model is directly used for inference on the target domain images. 2) **Target only** that training images with labels in the target domain are used to train the model directly, without pre-training in the source domain. For fairness, all compared methods were implemented using the same backbone of UNet [16].

The quantitative evaluation results are shown in Table 1. It can be observed that "Target only" outperformed "Source only" substantially, showing the large

**Table 1.** Quantitative comparison of different methods on M&MS dataset. Asterisks indicate a significant improvement from the best values obtained by existing methods (*: $p \leq 0.05$, **: $p \leq 0.01$). The best values are highlighted in bold.

| Metrics | Method | Target domain B | | | Target domain C | | | Target domain D | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | LV | MYO | RV | LV | MYO | RV | LV | MYO | RV |
| Dice ↑ (%) | Source only | 88.04±8.75 | 76.27±10.28 | 80.04±20.37 | 85.20±9.95 | 77.18±9.42 | 82.65±9.87 | 88.05±6.71 | 75.88±8.57 | 77.76±17.88 |
| | Target only | 90.29±7.04 | 84.62±7.33 | 87.56±7.72 | 90.36±4.50 | 83.80±4.75 | 84.66±7.91 | 89.49±6.91 | 80.83±4.81 | 81.87±11.70 |
| | PTBN [12] | **89.64±6.92** | 79.95±6.64 | 81.99±16.92 | 86.66±8.27 | 80.40±6.62 | 84.40±7.92 | 88.34±6.37 | 79.41±4.73 | 81.32±12.53 |
| | TENT [19] | 89.03±8.45 | 79.86±6.51 | 83.67±11.49 | 85.17±10.83 | 78.78±6.76 | 84.61±7.07 | 84.16±11.17 | 79.15±3.90 | 82.17±9.69 |
| | EATA [13] | 89.23±8.25 | 79.86±6.50 | 83.78±11.13 | 85.28±10.71 | 79.04±6.54 | 84.52±6.98 | 84.26±11.05 | 79.17±3.97 | 82.27±9.29 |
| | SAR [14] | 88.99±8.57 | 79.92±6.50 | 83.63±11.72 | 85.12±10.88 | 78.83±6.73 | 84.70±7.01 | 84.16±10.90 | 79.04±3.98 | 82.03±9.86 |
| | Ours | 89.63±6.16 | **81.93±6.21**** | **84.66±9.25** | **88.30±5.51*** | **82.24±6.23**** | **86.06±6.13*** | **89.34±5.69** | **80.99±4.89**** | **82.77±9.61** |
| ASSD ↓ (pixel) | Source only | 0.56±0.41 | 0.64±0.44 | 1.46±3.02 | 0.63±0.42 | 0.56±0.25 | 1.08±1.21 | 0.54±0.34 | 0.59±0.30 | 1.59±2.55 |
| | Target only | 0.36±0.27 | 0.37±0.21 | 0.62±0.65 | 0.36±0.16 | 0.66±1.30 | 1.06±0.83 | 0.55±0.51 | 0.65±0.36 | 0.99±1.05 |
| | PTBN [12] | 0.48±0.42 | 0.60±0.49 | 0.85±1.46 | 0.62±0.41 | 0.51±0.16 | 0.90±0.87 | 0.61±0.51 | 0.53±0.21 | 1.02±1.24 |
| | TENT [19] | 0.60±0.70 | 0.62±0.53 | 0.67±0.95 | 0.70±0.63 | 0.59±0.24 | 0.87±0.69 | 0.90±0.78 | 0.60±0.28 | 0.63±0.37 |
| | EATA [13] | 0.58±0.66 | 0.62±0.52 | 0.66±0.92 | 0.69±0.62 | 0.57±0.22 | 0.87±0.67 | 0.89±0.77 | 0.60±0.27 | 0.64±0.38 |
| | SAR [14] | 0.62±0.73 | 0.62±0.54 | 0.67±0.98 | 0.70±0.64 | 0.58±0.23 | 0.85±0.68 | 0.91±0.79 | 0.60±0.28 | 0.64±0.39 |
| | Ours | **0.41±0.22*** | **0.51±0.31**** | **0.59±0.60** | **0.48±0.24*** | **0.46±0.15**** | **0.62±0.45*** | **0.46±0.29** | **0.50±0.23** | **0.57±0.40** |

domain shift between the source and target domains. The existing SFDA methods only achieved a moderate improvement or even a decrease compared with "source only". In contrast, our method achieved excellent improvement and remarkably outperformed the vast majority of SFDA methods in terms of Dice and ASSD. For example, compared with "Source only" in target domain C, the average Dice of LV, MYO, and RV obtained by our method was 88.30%, 82.24%, and 86.06%, respectively, showing that our method improved the average Dice by 3.10%, 5.06%, and 3.41%, respectively. In terms of ASSD, our method achieved the lowest pixel-wise ASSD across all tissues within each domain. A visual comparison between different SFDA methods is shown in Fig. 2. Note that "Source only" achieved a poor performance, and the results of our method were closer to the ground truth and smoother than those of the other methods.

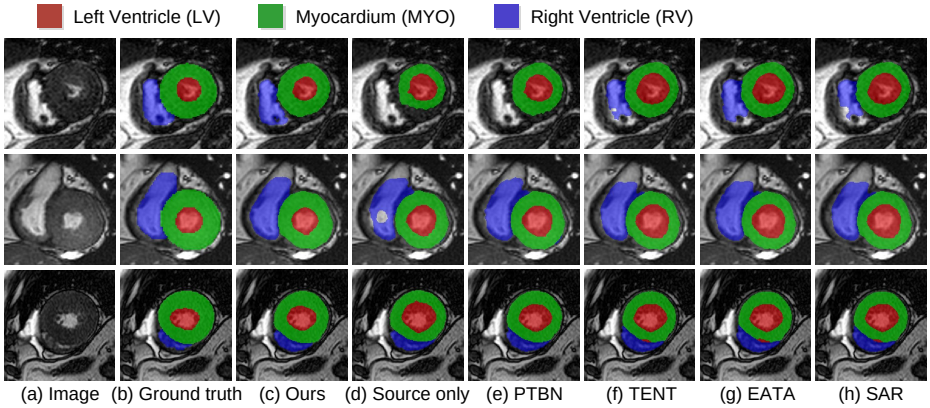


**Fig. 2.** Qualitative segmentation results of different methods. The three rows are from domains B, C, and D of the M&MS dataset, respectively.

**Table 2.** Effectiveness of components in our IPLC method. The first row (baseline) uses pseudo labels obtained by the source model for adaptation. $\mathcal{L}_{\mathcal{W}\_Dice}$: Generating SAM-Med2D pseudo labels by MRS and EWE methods to supervise the model. $\mathcal{L}_{curv}$: Using regularization based on mean negative curvature in adaptation. ICL: SAM-Med2D pseudo labels are iteratively refined through continuously updated prompts in domain adaptation.

| Components | | | Dice (%) ↑ | | | | ASSD (pixel) ↓ | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{L}_{\mathcal{W}\_Dice}$ | $\mathcal{L}_{curv}$ | ICL | B | C | D | Average | B | C | D | Average |
| | | | 83.74±9.29 | 84.42±6.39 | 83.63±6.58 | 83.93±7.42 | 0.54±0.45 | 0.55±0.29 | 0.55±0.31 | 0.55±0.35 |
| ✓ | | | 83.69±12.47 | 85.43±6.51 | 83.52±8.81 | 84.21±9.26 | 0.77±1.74 | 0.54±0.31 | 0.59±0.57 | 0.63±0.87 |
| ✓ | ✓ | | 84.15±10.81 | 85.50±6.31 | 84.00±8.39 | 84.55±8.50 | 0.70±1.51 | 0.53±0.31 | 0.55±0.47 | 0.59±0.76 |
| ✓ | | ✓ | 84.98±6.69 | 85.49±5.87 | 83.15±6.66 | 84.54±6.41 | 0.52±0.35 | 0.53±0.27 | 0.61±0.40 | 0.55±0.34 |
| ✓ | ✓ | ✓ | **85.41±7.21** | **85.53±5.96** | **84.37±6.73** | **85.10±6.63** | **0.50±0.37** | **0.52±0.28** | **0.51±0.30** | **0.51±0.32** |

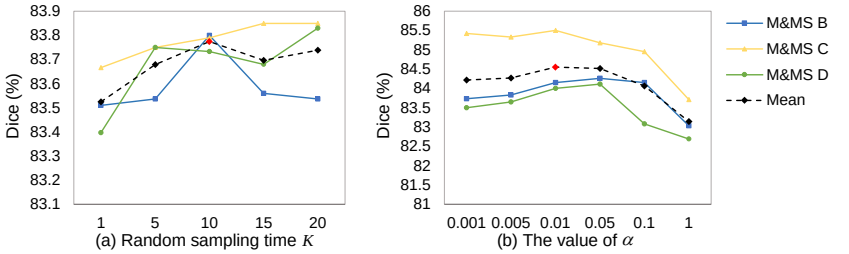


**Fig. 3.** Performance of our method with different hyper-parameters. (a) shows the performance of SAM-Med2D pseudo labels on the training set with different random sampling times $K$. (b) shows the effect of $\mathcal{L}_{curv}$ weight $\alpha$ on the validation set.

**Ablation Study** As shown in Fig. 3, there are two important hyper-parameters specific to our method: random sampling time $K$, and loss weight $\alpha$. We first investigated the effect of different $K$, and the performance in target domains is shown in Fig. 3(a). It can be observed that $K = 10$ achieved the best performance, which demonstrates the superiority of using MRS. Fig. 3(b) shows the performance on the validation set with different $\alpha$ values and the best $\alpha$ was 0.01. To evaluate the effectiveness of components in our IPLC, we further investigated the effect of $\mathcal{L}_{\mathcal{W}\_Dice}$, $\mathcal{L}_{curv}$ and ICL in Table 2. It shows that each component of our method led to a performance improvement. The baseline obtained an average Dice of 83.93%. Only using SAM-Med2D pseudo labels for adaptation obtained the Dice of 84.21%, and additionally using the regularization based on mean negative curvature or iterative correction learning improved it to 84.55% and 84.54%, respectively. Our proposed method combining all these components achieved the highest Dice of 85.10%.

## 4 Conclusion

In this paper, we propose a novel iterative pseudo label correction guided by SAM framework to address the limitations of existing SFDA methods for medical image segmentation. We propose multiple random sampling and entropy

weight estimation that obtain robust pseudo labels and mitigate the noise of pseudo labels. We introduce mean negative curvature minimization, which provides sufficient supervision for the model. We also propose an iterative correction learning strategy to iteratively refine SAM pseudo labels, which improves the quality of pseudo labels for domain adaptation. Extensive experiments on the public M&MS dataset demonstrate the effectiveness of our method, and our method achieves excellent performance over several state-of-the-art methods.

**Acknowledgments.** This work was supported by the National Natural Science Foundation of China under grant 62271115.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Bateson, M., Kervadec, H., Dolz, J., Lombaert, H., Ayed, I.B.: Source-free domain adaptation for image segmentation. Med. Image Anal. **82**, 102617 (2022)
2. Campello, V.M., et al.: Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge. IEEE Trans. Med. Imaging **40**(12), 3543–3554 (2021)
3. Chen, C., Liu, Q., Jin, Y., Dou, Q., Heng, P.A.: Source-free domain adaptive fundus image segmentation with denoised pseudo-labeling. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 225–235. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_22
4. Cheng, J., et al.: SAM-Med2D. arXiv preprint arXiv: 2308.16184 (2023)
5. Ding, N., Xu, Y., Tang, Y., Xu, C., Wang, Y., Tao, D.: Source-free domain adaptation via distribution estimation. In: CVPR. pp. 7212–7222 (2022)
6. Fleuret, F., et al.: Uncertainty reduction for model adaptation in semantic segmentation. In: CVPR. pp. 9613–9623 (2021)
7. 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
8. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Trans. Biomed. Eng. **69**(3), 1173–1185 (2021)
9. Kirillov, A., et al.: Segment anything. In: ICCV. pp. 4015–4026 (2023)
10. Kouw, W.M., Loog, M.: A review of domain adaptation without target labels. IEEE Trans. Pattern Anal. Mach. Intell. **43**(3), 766–785 (2019)
11. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
12. Nado, Z., et al.: Evaluating prediction-time batch normalization for robustness under covariate shift. arXiv preprint arXiv: 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. In: ICML. pp. 16888–16905 (2022)
14. Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., Tan, M.: Towards stable test-time adaptation in dynamic wild world. In: ICLR (2023)
15. Pei, C., Wu, F., Huang, L., Zhuang, X.: Disentangle domain features for cross-modality cardiac image segmentation. Med. Image Anal. **71**, 102078 (2021)

16. 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
17. Shen, S., et al.: Cortical atrophy in early-stage patients with anti-NMDA receptor encephalitis: a machine-learning MRI study with various feature extraction. Cerebral Cortex **34**(2), bhad499 (2024)
18. Varsavsky, T., et al.: 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
19. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: ICLR (2021)
20. Wu, J., Guo, D., Wang, G., Yue, Q., Yu, H., Li, K., Zhang, S.: FPL+: Filtered pseudo label-based unsupervised cross-modality adaptation for 3D medical image segmentation. IEEE Trans. Med. Imaging (2024)
21. Wu, J., et al.: UPL-SFDA: Uncertainty-aware pseudo label guided source-free domain adaptation for medical image segmentation. IEEE Trans. Med. Imaging **42**(12), 3932–3943 (2023)
22. Xian, J., Li, X., Tu, D., Zhu, S., Zhang, C., Liu, X., Li, X., Yang, X.: Unsupervised cross-modality adaptation via dual structural-oriented guidance for 3D medical image segmentation. IEEE Trans. Med. Imaging **42**(6), 1774–1785 (2023)
23. Xu, X., et al.: A novel one-to-multiple unsupervised domain adaptation framework for abdominal organ segmentation. Med. Image Anal. **88**, 102873 (2023)
24. You, F., Li, J., Zhu, L., Chen, Z., Huang, Z.: Domain adaptive semantic segmentation without source data. In: ACM Multimedia. pp. 3293–3302 (2021)