# A Chebyshev Confidence Guided Framework in Source-Free Domain Adaptation for Medical Image Segmentation

Jiesi Hu, Yanwu Yang, Xutao Guo, Jinghua Wang*, Ting Ma*

***Abstract*— Unsupervised domain adaptation approaches aim to adapt models trained on a labeled source domain to an unlabeled target domain. However, privacy concerns, especially in medical imaging scenarios involving patient information, have highlighted the practical significance of source-free domain adaptation (SFDA). State-of-the-art SFDA methods primarily rely on pseudo-label (PL) based self-training. Unfortunately, PLs suffer from accuracy deterioration caused by domain shift limiting the effectiveness of the adaptation process. To address these issues, we propose a Chebyshev confidence guided framework to generate more accurate and clean PLs for self-training. We estimate the Chebyshev confidence of PLs by calculating the alignment probability lower bound of the current PLs with the model, serving as a measure of reliability. Then, we build a novel teacher-student joint training scheme (TJTS), supplemented by a confidence weighting module, to obtain more precise PLs. The framework also includes two confidence-guided denoising methods: direct denoising and prototypical denoising. Through the collaboration of TJTS and the proposed denoising methods, our framework effectively prevent label noise propagation while continuously generating more accurate PLs. Extensive experiments in diverse domain scenarios validate the effectiveness of our proposed framework and establish its superiority over state-of-the-art SFDA methods. Our paper contributes to the field of SFDA by providing a novel approach for precisely estimating the reliability of pseudo-labels and a framework for obtaining high-quality PLs, resulting in improved adaptation performance.**

***Index Terms*— Source-free domain adaptation, Image segmentation, Self-training, Pseudo-label denoising,**

This work was supported in part by grants from the National Natural Science Foundation of P.R. China (62276081, 62106113), Innovation Team and Talents Cultivation Program of National Administration of Traditional Chinese Medicine (NO:ZYYCXTD-C-202004), Basic Research Foundation of Shenzhen Science and Technology Stable Support Program (GXWD20201230155427003-20200822115709001) and The Major Key Project of PCL (PCL2021A06). (Corresponding author: Jinghua Wang, Ting Ma.)

Jiesi Hu , Yanwu Yang, and Xutao Guo are with School of Electronics and Information Engineering, Harbin Institute of Technology at Shenzhen, and The Peng Cheng Laboratory. (e-mail: 405323011@qq.com, yangyanwu1111@gmail.com, 18B952052@stu.hit.edu.cn)

Jinghua Wang is with School of Computer Science, Harbin Institute of Technology at Shenzhen. (e-mail: )

Ting Ma is with School of Electronics and Information Engineering, Harbin Institute of Technology at Shenzhen, The Peng Cheng Laboratory, Guangdong Provincial Key Laboratory of Aerospace Communication and Networking Technology, Harbin Institute of Technology, Shenzhen, and International Research Institute for Artifcial Intelligence, Harbin Institute of Technology, Shenzhen. (e-mail: tma@hit.edu.cn)

## I. Introduction

Deep neural network (DNN) models have achieved remarkable success in a wide range of visual recognition tasks [1]–[4]. However, these models often suffer significant performance degradation when confronted with distribution or domain shifts which often exist between the training (source) and test (target) domains [5], [6]. This issue is particularly evident in medical imaging. For example, a model trained on data captured from one clinical centre may exhibit poor performance when applied to data captured from another clinical centre.

To overcome the problem induced by domain shifts, numerous algorithms have been developed in the field of Unsupervised Domain Adaptation (UDA) [7], [8], [45]. The applicability of most UDA techniques is limited by an underlying assumption that the labeled source domain data is available for training, as this assumption is often impractical and restrictive in clinical applications due to concerns regarding privacy and security [47]. Consequently, source-free domain adaptation (SFDA) has gained significant interest, particularly in medical image analysis [5], [9]–[12], [44]. SFDA considers only the availability of a pre-trained model on the source domain and unlabeled target domain data. It enables clinical centres to adapt a well-trained model to their own data without exchanging sensitive health information, making it a practical and efficient approach.

Current SFDA methods employ various techniques, such as leveraging batch normalization parameters, contrastive learning, and target-to-source data transformation [5], [10], [12]–[14]. Among these methods, the widely utilized approach is self-training guided by pseudo-labels (PLs) which are generated by the model's predictions on unlabeled target data [31]. However, during the early stages of adaptation, the PLs can be highly misleading or noisy, as shown in Fig. 1. Using such PLs can propagate erroneous knowledge and hinder domain adaptation effectiveness. Therefore, obtaining accurate and noise-free PLs for self-training is crucial and requires proper attention [12], [15].

Obtaining high-quality PLs have two major challenges. Firstly, the PLs may deviate significantly from the ground truth in the target domain due to domain shift. The inaccurate PLs hampers the model's ability to achieve precise segmentation in the target domain. Consequently, there is an urgent need

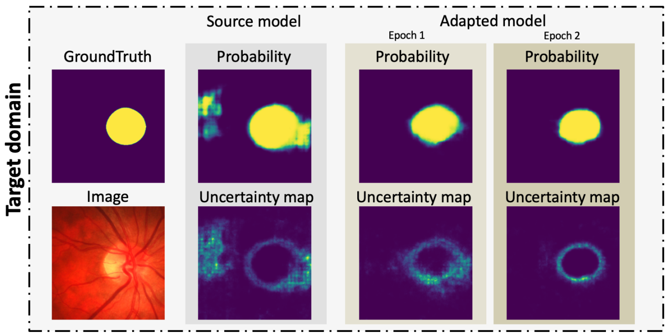


Fig. 1: The figure illustrates the evolution of segmentor outputs during the adaptation process. The outputs from the source model exhibit noticeable noise, which should be minimized in self-training. In contrast, the adapted model demonstrates improvement in predictions and reduced uncertainty. Leveraging the predictions from the adapted model enables the generation of more accurate pseudo-labels. Moreover, the probabilities and uncertainties offer insights into the model's reliability in various aspects of prediction. It is crucial to effectively combine these measures to determine the model's confidence in the pseudo-labels, which is a primary objective of this study.

for generating more accuract PLs, but the existing methods remain insufficient. Secondly, PLs inevitably contain noise, necessitating a critical denoising step in the self-training process. Typically, individuals employ probability [5], entropy [6] or uncertainty [12], [15] to estimate the reliability of the PLs and remove unreliable samples. However, solely considering probability neglects the inherent uncertainty in the prediction, specifically the fluctuation of probability. Conversely, solely considering uncertainty overlooks the predicted probability that determines the value of PLs. A more comprehensive and accurate method is required to assess the reliability of PLs. These problems are particularly critical in the field of medical imaging where the model is required to be highly stable and accurate. Designing a SFDA approach that generate high-quality PLs is a challenging task and a primary focus of this study.

Our research centers around the attainment of more precise and dependable PLs. To achieve this objective, we introduce a Chebyshev confidence guided framework that incorporates the proposed Chebyshev confidence estimation as a fundamental component. This Chebyshev confidence estimation method computes the probability lower bound with which the model aligns with the current PL.

To improve the accuracy of PLs, we propose a teacher-student joint training scheme. This scheme facilitates knowledge exchange between the student and teacher models while filtering out poor knowledge using Chebyshev confidence, ensuring continuous refinement of the PLs. For PL denoising, we employ two methods: direct denoising and prototypical denoising, both based on the proposed Chebyshev confidence. Direct denoising removes pixels with low confidence using Chebyshev confidence as an indicator. Prototypical denoising considers correlation among samples of the same class in the feature space, accurately estimating prototypes by weighting the samples with Chebyshev confidence. By combining these complementary denoising methods, we obtain clean PLs. Additionally, to prevent overconfidence, we incorporate a diversity loss term. The modules in our framework complement each other, synergistically enhancing adaptation performance.

In summary, our paper provides the following contributions:

- We propose a novel technique for estimating the reliability of PLs, called Chebyshev confidence, which computes the probability lower bound of aligning with the current PLs.
- We introduce a teacher-student joint training scheme that facilitates the continuous improvement of PLs and reduces the weight of unreliable PLs.
- We propose two effective denoising methods, namely direct denoising and prototypical denoising, based on the estimated confidence. These methods utilize pixel and category information, respectively, for denoising.
- We conduct extensive experiments across various domain scenarios, including cross-centre and cross-modality settings. The results demonstrate that our model outperforms other state-of-the-art SFDA methods. Modules in our model complement each other, leading to further improvement in performance.

## II. Related Works

### A. Unsupervised Domain Adaptation

Unsupervised Domain Adaptation (UDA) has received extensive attention in the literature for visual recognition tasks [16], [17]. Previous works on UDA have utilized popular techniques such as adversarial learning [8], [19], [34], [45], image-to-image translation [8], [20], [21], and cross-domain divergence minimization [22]–[24]. Self-training methods [6], [25]–[27] have also gained prominence in UDA, where a student model is iteratively trained using labeled source data and pseudo-labeled target data generated by a teacher model. With the increasing demand for automated medical image analysis, domain adaptation models have received considerable attention in the field of medical imaging [45], [48], [49] due to the time-consuming nature of manual labeling and the requirement for specialized knowledge.

However, most existing UDA approaches rely on continued access to labeled source domain data during domain adaptation training, which is often impractical in real-world scenarios due to data privacy concerns. To address this limitation, the setting of SFDA has garnered significant interest, as it does not require access to the source data during adaptation.

### B. Source-Free Domain Adaptation

Due to concerns regarding data privacy, SFDA has emerged as an approach to achieve adaptation using only unlabeled target data and a source model, without relying on the source data. In recent years, several approaches have been proposed to address the challenges of SFDA in both natural and medical imaging domains [5], [10], [12], [28], [44]. SHOT [10] employs a centroid-based method to generate PLs for self-training and freezes the last few layers during adaptation. Tent [28] freezes parameters, except for batch normalization, and utilizes entropy minimization to update weights. OSUDA [5] applies constraints to batch normalization parameters and utilizes predicted probability for PL denoising and selection. DPL [12] performs denoising on PLs at both the pixel and class levels based on uncertainty estimation. Additionally, [44] leverages parameters in batch normalization to transform target images into source-style images and incorporates contrastive learning for self-training.

In SFDA, PL are widely used [5], [10], [12], [44] but often contain noise, which can potentially mislead the model. Therefore, accurately estimating the reliability of PLs is crucial. Prior works, such as [12] and [30], filter out PL samples based on uncertainty. Meanwhile, [5] and [29] primarily utilize predicted probability to assess the reliability of PLs. However, considering only probability or uncertainty alone is insufficient. In this study, we comprehensively leverage both probability and uncertainty to assess the model's confidence on PLs, enabling a more accurate denoising. Besides, we introduce a teacher-student scheme to further enhance the quality of the PLs.

## III. METHODOLOGY

### A. Problem Setting

We denote the source domain dataset by $D_s = \{(x_s^i, y_s^i)\}_{i=1}^{N_s}$ consists of $N_s$ labeled samples in the source domain. Here, $y_s^i$ denotes the label for the image $x_s^i$. The source model, denoted as $f_{\theta_s} : x_s^i \to y_s^i$, is trained on $D_s$. The target domain dataset $D_t = \{(x_t^i)\}_{i=1}^{N_t}$ contains $N_t$ unlabeled samples. In unsupervised SFDA, we do not have access to the target labels $\{(y_t^i)\}_{i=1}^{N_t}$ throughout the entire adaptation process. Both $D_s$ and $D_t$ follow the same underlying label distribution, with a common label set $C = \{1, 2, ..K\}$. In the segmentation task, we consider medical images in different scenarios, such as fundus and brain MRI, where $x^i \in \mathbb{R}^{H \times W \times C}$ and $y^i \in \{1, 0\}^{H \times W \times K}$. Here, $C$ represents the number of channels in the input image, and $K$ denotes the number of classes. In this study, we focus on the SFDA problem and $D_s$ is unavailable when adapting $f_{\theta_s}$ on $D_t$.

### B. Chebyshev Confidence Estimation

Firstly, we present our approach for estimating the confidence of PLs, which serves as a fundamental component of our framework.

In the PL-based method, the performance of the resulting model is directly determined by the quality of the PLs. Thus, it is crucial to assess the reliability of the PLs and guide the training procedure using only the reliable ones. Whether it is evaluating the weights of PLs, denoising, or estimating prototypes, the assessment of PL reliability cannot be circumvented [12], [29], [35]. To address this challenge and achieve a comprehensive and accurate assessment of reliability, we propose to estimate the reliability of PLs from a probabilistic perspective. Specifically, we compute the probability of the model aligning with the current PL.

Given an input image $I$ to a DNN, the output corresponding to a pixel $I_{i,j}$ is denoted as $Z_{i,j}$. For simplicity, we use $Z$ to represent $Z_{i,j}$. Treating $Z$ as a random variable, we turn on the dropout module in the model as described in [36] to obtain multiple samples $\{z_1, z_2, ..., z_n\}$ of $Z$. Then, we calculate the mean $p$ and uncertainty $\sigma$. In our experiments, we utilize dropout uncertainty [36] which computes the standard deviation of the predicted probability. The corresponding PL is defined as $\hat{y} = \mathbb{1}[p \geq \mathcal{T}]$, where $\mathcal{T} \in (0, 1)$ is a probability threshold used to generate binary PLs. Ideally, we aim to estimate the probabilities $P(\mathbf{Z} > \mathcal{T}|\hat{y} = 1)$ and $P(\mathbf{Z} < \mathcal{T}|\hat{y} = 0)$ to evaluate the confidence of the PL. However, directly calculating these probabilities is difficult. To address this, we propose a novel technique that utilizes the one-tailed variant of Chebyshev's inequality, incorporating commonly applied statistics mean $p$ and uncertainty $\sigma$, as shown in Fig. 3. We first obtain the following expression for any real number $k > 0$:

$$P(\mathbf{Z} - p \geq k\sigma) \leq \frac{1}{1 + k^2}. \tag{1}$$

Considering $\hat{y} = 0$, let $k = \frac{1}{\sigma}(\mathcal{T} - p)$. Then, we have:

$$P(\mathbf{Z} \geq \mathcal{T}|\hat{y} = 0) \leq \frac{\sigma^2}{\sigma^2 + (\mathcal{T} - p)^2}. \tag{2}$$

Since $P(\mathbf{Z} < \mathcal{T}|\hat{y} = 0) = 1 - P(\mathbf{Z} \geq \mathcal{T}|\hat{y} = 0)$, we obtain

$$P(\mathbf{Z} < \mathcal{T}|\hat{y} = 0) \geq \frac{(\mathcal{T} - p)^2}{\sigma^2 + (\mathcal{T} - p)^2}. \tag{3}$$

Similarly, for $\hat{y} = 1$, we derive the following expression:

$$P(\mathbf{Z} > \mathcal{T}|\hat{y} = 1) \geq \frac{(p - \mathcal{T})^2}{\sigma^2 + (p - \mathcal{T})^2}. \tag{4}$$

It is worth noting that the right sides of (3) and (4) estimate the lower bound of the probabilities and are actually the same. Thus, we define an indicator $l = \frac{1}{\sigma^2 + (p - \mathcal{T})^2}(p - \mathcal{T})^2$ to serve as a measure of the Chebyshev confidence.

When the values of $p$ and $\mathcal{T}$ are close, the resulting PL is often unreliable, resulting in a low Chebyshev confidence. When $\sigma^2$ is large, indicating high uncertainty in the model's predictions, the estimated Chebyshev confidence diminishes. In summary, this formula for Chebyshev confidence implicitly covers various scenarios and provides a concise representation, considering both the model's prediction and uncertainty. Additionally, this method can be extended to multi-class tasks without a pre-defined $\mathcal{T}$ by substituting $\mathcal{T}$ with the second highest predicted probability among all classes.

### C. Teacher-Stduent Joint Training Scheme

We employ a PL-based self-training mechanism for SFDA. In most existing methods, the teacher model ($f_{\theta_{te}}$) generates PLs to guide the training of the student model ($f_{\theta_{st}}$) [10], [12], [44]. However, since both models are initialized with the same weights ($\theta_s$) and the teacher model remains fixed during adaptation, the resulting PLs have a higher error rate, limiting the model's performance and potentially introducing erroneous knowledge.

The student model adapts rapidly to the target domain through backpropagation with target domain data, leading to improved PLs, as supported by previous studies [5], [15], [33]. In contrast, the teacher model exhibits more stable and consistent performance. Leveraging this observation, we introduce a teacher-student joint training scheme which combines PLs generated by both models for self-training. In this scheme, knowledge can be continuously transmitted between

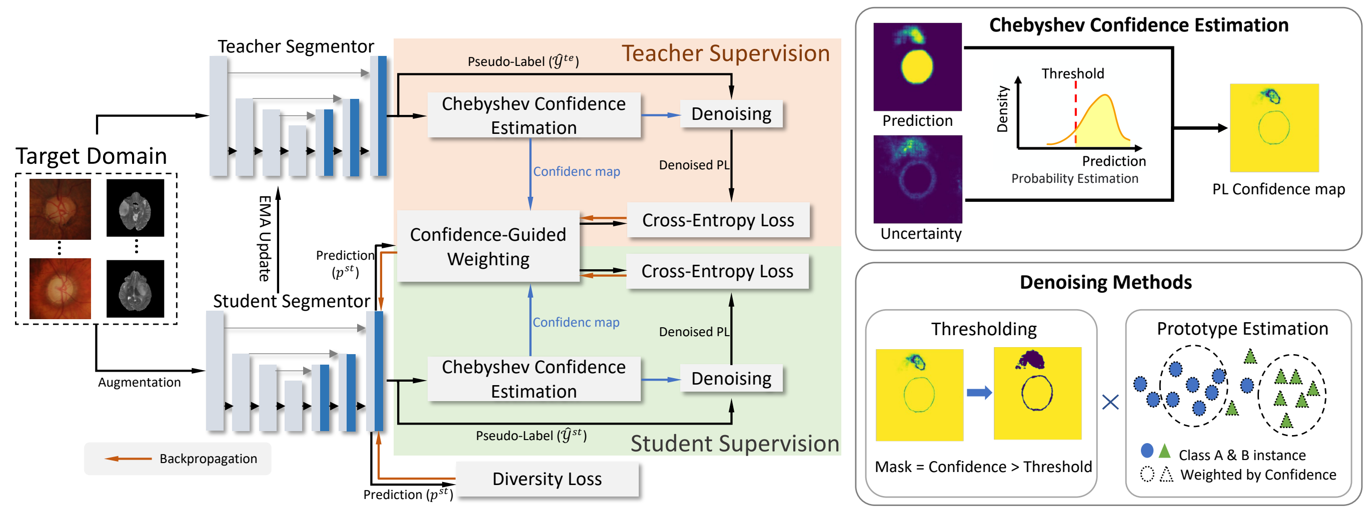


Fig. 2: Overview of our proposed framework where we use predictions from both the teacher and student for self-training. We incorporate confidence guidance to assign higher weights to more accurate predictions. In the upper-right part of the figure, we present the Chebyshev Confidence Estimation module, which estimate the probability lower bound of the model aligning with the current pseudo-label, taking into account the model's uncertainty and predicted probability. Based on this estimated confidence, we apply the direct and prototypical denoising methods to refine the pseudo-labels, as shown in the lower-right part of the figure.

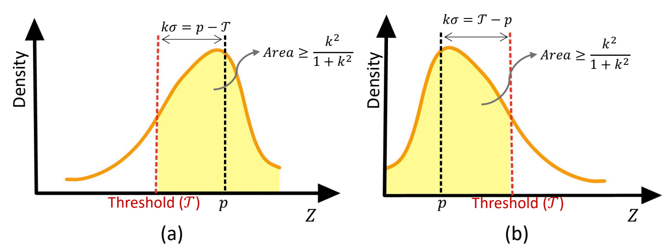


Fig. 3: Probability density function of model's prediction for a pixel. (a) and (b) show the estimation of Chebyshev Confidence when the pseudo-label is 1 and 0, respectively. $p$ and $\sigma$ denote the mean predicted probability and corresponding uncertainty.

the student and teacher models, and poor knowledge can be filtered out using Chebyshev confidence.

In our framework, illustrated in Fig. 2, both the teacher model ($f_{\theta_{te}}$) and the student model ($f_{\theta_{st}}$) are employed to process the data from the target domain. Cross-entropy (CE) loss from the teacher supervision is as follows:

$$\mathcal{L}_i^{te} = -\sum_v [\hat{y}_v^{te} \cdot log(p_v^{st}) + (1 - \hat{y}_v^{te}) \cdot log(1 - p_v^{st})]. \quad (5)$$

Here, $\mathcal{L}_i^{te}$ represent the CE loss for the $i$-th sample using PLs from the teacher model. $p_v^{st}$ represents the predicted probability of the student model for the $v$-th pixel in the $i$-th sample. $\hat{y}_v^{te}$ are the corresponding PLs generated by the teacher model. The CE loss from the student supervision is obtained by replacing the PLs from the teacher model with the PLs from the student model, as shown in the following:

$$\mathcal{L}_i^{st} = -\sum_v [\hat{y}_v^{st} \cdot log(p_v^{st}) + (1 - \hat{y}_v^{st}) \cdot log(1 - p_v^{st})]. \quad (6)$$

Given the varying quality of PLs generated by the student and teacher models across different input data, we employ Chebyshev confidence weighting to assign a higher weight to PLs of superior quality. Drawing inspiration from the approach in [34], we employ the following weight calculation formula:

$$w_i^{te} = \frac{e^{-\gamma L_i^{te}}}{e^{-\gamma L_i^{te}} + e^{-\gamma L_i^{st}}}, \quad w_i^{st} = 1 - w_i^{te}, \quad (7)$$

where $L_i^{te} = \mathbb{E}_{v \in x_i}(l_v^{te})$ and $L_i^{st} = \mathbb{E}_{v \in x_i}(l_v^{st})$ are the mean values of the Chebyshev confidence maps for sample $i$ from the teacher and student models, respectively. The hyperparameter $\gamma$ adjusts the effect of Chebyshev confidence. The segmentation loss is then built as:

$$\mathcal{L}_{ce} = \sum_i^{N_t} w_i^{te} \mathcal{L}_i^{te} + \sum_i^{N_t} w_i^{st} \mathcal{L}_i^{st}. \quad (8)$$

To continuously refine the teacher model, we utilize the exponential moving average (EMA) approach to slowly update the teacher parameters ($\theta_{te}$) using student parameters ($\theta_{st}$). The update equation from iteration $j$ to $j+1$ is:

$$\theta_{te}^{j+1} = \beta \theta_{te}^j + (1 - \beta) \theta_{st}^j, \quad (9)$$

where $\beta$ is a smoothing factor that controls the degree of change. Additionally, our framework incorporates augmented data inputs for the student model to enhance its generalization capabilities.

### D. Denoising Methods Based on Pseudo-Label Confidence

To accurately identify noise, we propose direct denoising and prototypical denoising that leverage pixel and category information, respectively.

*1) Direct Denoising:* One common approach after obtaining the Chebyshev confidence is to apply a threshold, denoted as $\eta$, to remove points with low confidence. Specifically, we define a binary mask $m \in \{0, 1\}$ as follows:

$$m = \mathbb{1}[l \geq \eta], \quad (10)$$

where $l$ represents the estimated Chebyshev confidence. Pixels with $l$ less than $\eta$ are considered unreliable and excluded from the loss function. Considering the initially poor quality of the PL [5], we linearly decrease $\eta$ from 0.99 to 0.95 during the adaptation process.

*2) Prototypical Denoising:* Prototype estimation is a commonly used technique to capture the characteristics of the same category and guide PL refinement [10], [12], [37]. By evaluating the consistency between the PL and the corresponding class prototype, unreliable PLs can be identified.

However, using unreliable PLs leads to inaccurate prototype estimation. To address this, we propose a novel confidence-weighted prototype estimation method based on Chebyshev confidence. The prototype for class $k$ is computed as:

$$z^k = \frac{\sum_{v \in \Omega} e_v l_v \mathbb{1}[\hat{y}_v = k]}{\sum_{v \in \Omega} l_v \mathbb{1}[\hat{y}_v = k]}. \tag{11}$$

Here, $\hat{y}_v$ is the PL for pixel $v$, and $e_v$ denotes the corresponding interpolated output feature of the backbone. The estimated confidence $l_v$ is used to reduce the weight of unreliable pixels. By using this approach, we can reduce the influence of outliers when computing prototypes. $\Omega$ is a set of pixels in one mini-batch. Based on the distance from prototypes, we compute the prototypical PL:

$$\hat{y}_v^{proto} = \arg\min_k \|e_v - z^k\|. \tag{12}$$

Inconsistent prototypical PLs indicate a misalignment between a pixel's position in the feature space and the corresponding prediction, suggesting a higher likelihood of labeling errors. Then, we remove them by updating the mask to $m_v = \mathbb{1}[\hat{y}_v = \hat{y}_v^{proto}]\mathbb{1}[l_v \geq \eta]$. Adding the denoising mask to formula (8), we get:

$$\mathcal{L}_{ce} = \sum_i^{N_t} w_i^{te} \sum_v m_v^{te} \mathcal{L}_{i,v}^{te} + \sum_i^{N_t} w_i^{st} \sum_v m_v^{st} \mathcal{L}_{i,v}^{st}. \tag{13}$$

## E. Diversity Loss

Self-training methods may blindly trust false labels and exhibit bias towards easier classes [10], [14]. To mitigate this issue and maintain prediction diversity, we introduce a regularization term in the loss function:

$$\mathcal{L}_{div} = \sum_k \overline{p}^k log \overline{p}^k, \tag{14}$$

where $\overline{p}^k = \mathbb{E}_{x_t^i \in D_t}(p_i^k)$ represents the average predicted probability of class $k$ for the entire target domain. This regularization term helps to prevent overconfidence on certain predictions and promotes a more balanced and diverse output. This diversity loss term is combined with the cross-entropy loss, resulting in the overall training loss as follows:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{div}, \tag{15}$$

where $\lambda$ is a hyperparameter that balances the weight of diversity loss.

# IV. Experiments And Results

## A. Datasets

Comprehensive experiments were conducted on both fundus and brain MRI images segmentation to evaluate the proposed SFDA framework, encompassing cross-centre and cross-modality domain scenarios.

*1) Fundus Image Segmentation:* We performed SFDA for optic disc and cup segmentation of retinal fundus images using datasets from different clinical centres. Our source domain comprised 400 annotated training images from the REFUGE challenge [39]. As distinct target domains, we utilized the RIM-ONE-r3 [40] and Drishti-GS [41] datasets. Following the experimental protocol outlined in [12], the target domains consisted of training/testing subsets of 99/60 and 50/51 images, respectively. All images were cropped to a 512×512 disc region, which served as the input for our models.

*2) Brain Tumor Segmentation:* We conducted cross-modality SFDA on the BraTS2020 dataset [42], with a specific focus on whole tumor segmentation using T1, T1ce, T2, and FLAIR modalities. The image volumes in BraTS originally had different resolutions, and were co-registered and interpolated to a standard resolution. We targeted low-grade glioma cases and performed cross-modality SFDA on two pairs of MRI modalities: FLAIR↔T2 and T1↔T1ce, which exhibit relatively small appearance discrepancies. The target domains consists of randomly split training/testing subsets, with 53/23 cases and corresponding 3439/1487 slices. Each slice was resized to 512 × 512.

## B. Implementation Details

Given the target images $D_t$ and the source model $\theta_s$, the student and teacher models were initialized with $\theta_s$. Subsequently, data from $D_t$ were inputted into the models, and PLs were generated. Denoising techniques were then applied to the PLs, which were used to train the student model. Our network backbone was based on a MobileNetV2 adapted DeepLabv3+ architecture [38], which was pretrained on ImageNet [43]. During the training of the source model, the segmentation network was first trained on labeled source data for 100 epochs. We used the Adam optimizer with a learning rate of 1e-3 and momentum of 0.9 and 0.99. In the adaptation stage, the model was also trained using the Adam optimizer with the same momentum. Following the approach in [12], for fundus image segmentation, the target model was trained for 2 epochs with a batch size of 8. For brain tumor segmentation, training was performed for 10 epochs to ensure convergence. The learning rate was set to be $5e^{-4}$ and $1e^{-5}$ for the fundus and brain datasets, respectively. To estimate uncertainty using Monte Carlo Dropout, we set the dropout rate to 0.5 and performed 10 stochastic forward passes to obtain the standard deviation of the output probabilities. The threshold $\mathcal{T}$ was set to 0.75, following [12]. During training, weak augmentations such as Gaussian noise, Gibbs noise, contrast adjustment, and random erasing were applied to slightly perturb the input data. The hyperparameters $\gamma$, $\lambda$, and $\beta$ were set to be 1000, 0.3, and 0.999, respectively. The implementation was based on PyTorch 1.8.2 and utilized an NVIDIA V100 GPU.

For evaluation, we used two commonly-used metrics: the Dice coefficient and the Average Surface Distance (ASD). The Dice coefficient measured the overlap between the predicted segmentation and the ground truth, with higher values indicating better segmentation accuracy. The ASD measured the average distance between the predicted and ground truth

TABLE I: Performance comparison was conducted between our method and state-of-the-art domain adaptation techniques for optic disc and cup segmentation. The evaluation metrics used were Dice coefficient (%) and Average Surface Distance (ASD) in pixels. The best and second best results obtained are highlighted and underlined.

| Methods | Access source | RIM-ONE-r3 | | | | Drishti-GS | | | | Avg. Dice↑ | Avg. ASD↓ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Optic Disc | | Optic Cup | | Optic Disc | | Optic Cup | | | |
| | | Dice↑ | ASD↓ | Dice↑ | ASD↓ | Dice↑ | ASD↓ | Dice↑ | ASD↓ | | |
| Source only | | 83.18 | 24.15 | 74.51 | 14.44 | 93.84 | 9.05 | 83.36 | 11.39 | 83.72 | 14.76 |
| Target only | ✓ | 96.8 | - | 85.6 | - | 97.4 | - | 90.1 | - | 92.48 | - |
| BEAL [45] | ✓ | 89.8 | - | **81.0** | - | 96.1 | - | 86.2 | - | 88.28 | - |
| Tent [28] | | 90.61 | 8.54 | 79.43 | **7.97** | 96.41 | 4.05 | 80.77 | 12.81 | 86.81 | 8.34 |
| DPL [12] | | 90.13 | 9.43 | 79.78 | 9.01 | 96.39 | 4.08 | 83.53 | 11.39 | 87.46 | 8.48 |
| FSM [44] | | 87.90 | 13.06 | 79.43 | 8.32 | 95.68 | 5.52 | 82.94 | 11.06 | 86.49 | 9.40 |
| OSUAD [5] | | 88.61 | 10.48 | 78.44 | 8.89 | 95.99 | 4.44 | 82.33 | 11.72 | 86.34 | 8.88 |
| Ours | | **94.45** | **4.80** | 79.85 | 8.29 | **96.52** | **4.04** | **86.30** | **8.95** | **89.28** | **6.52** |

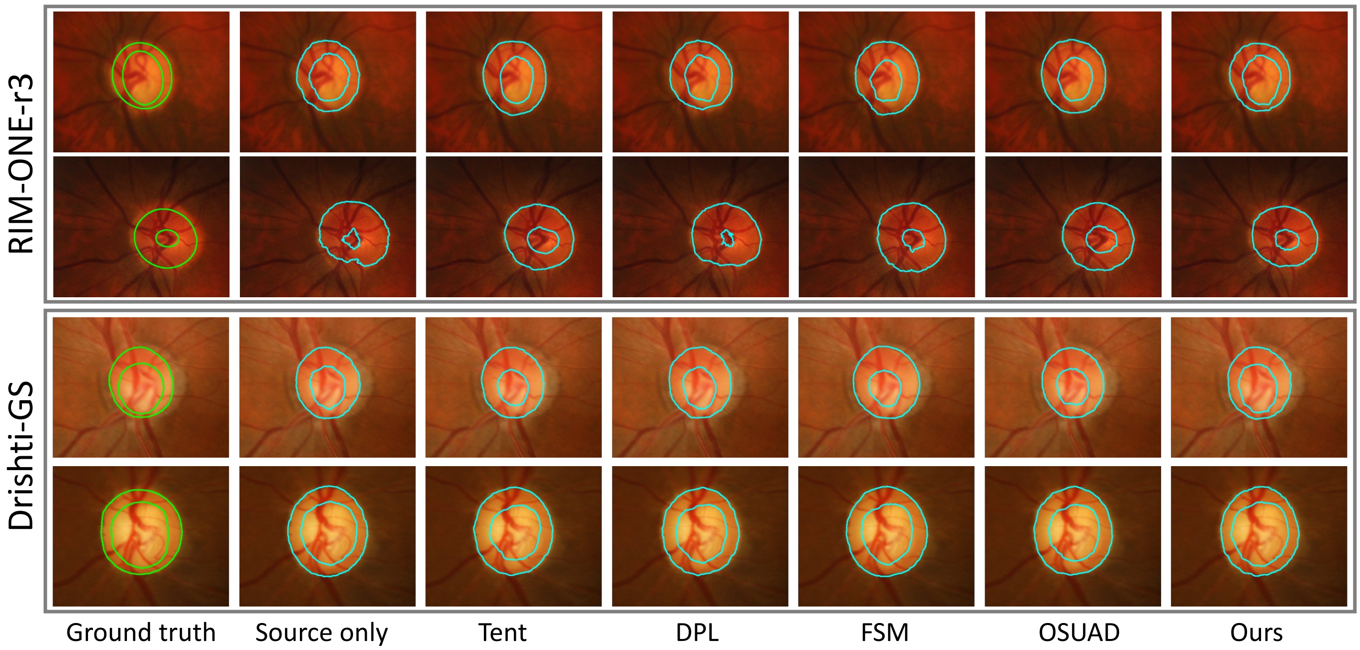


Fig. 4: Qualitative comparison of the adaptation performance using different methods for optic disc and cup segmentation.

surfaces, with lower values indicating better segmentation accuracy.

### C. Competitive State-of-the-Art Domain Adaptation Methods

To assess the effectiveness of our proposed SFDA method, we conducted a comparison with several state-of-the-art methods, namely BEAL [45], Tent [28], DPL [12], FSM [44], and OSUAD [5]. The method proposed by BEAL [45] is an unsupervised domain adaptation (UDA) approach designed specifically for the fundus dataset, utilizing both source and target images during adaptation. On the other hand, DPL [12], FSM [44] and OSUAD [5] are the latest SFDA methods that focus on the medical images. These SFDA methods are provided with a source model but not the source data for target domain adaptation. We also compare the SFDA mode of Tent [28], which primarily focuses on natural datasets. Given that these methods are trained on different tasks with diverse datasets, it is not appropriate to directly apply them to our evaluation. Therefore, to ensure a fair comparison, we trained SFDA methods using the same MobileNetV2 backbone [46]. Additionally, we included the Source only and Target only models, which were trained exclusively on either the source or target data. These models serve as the lower and upper bounds for the domain adaptation problem, respectively.

### D. Experimental Results

*1) Results on Optic Disc and Cup Segmentation:* Table I presents a performance comparison of Optic Disc and Cup Segmentation using Dice and ASD metrics on two target datasets: RIM-ONE-r3 and Drishti-GS. We refer to the results reported in the BEAL paper [45], which developed an unsupervised domain adaptation (UDA) model for cross-domain fundus image segmentation. Additionally, we incorporate the findings from DPL [12], which employed the same experimental setting as ours. We retrained other SOTA methods using the same backbone. Comparing the performance of the models, our approach achieves a significant improvement in Disc segmentation, surpassing other methods by 3.84% and 3.74 in Dice and ASD metrics, respectively, for the RIM-ONE-r3 dataset. On Drishti-GS, our method demonstrates the best performance for both Disc and Cup segmentation. Specifically,

TABLE II: Performance Comparison with State-of-the-Art Domain Adaptation Methods on Brain Tumor Segmentation using Dice (%) and ASD (pixel). Best and second-best results are highlighted and underlined.

| Methods | Access source | T2 $\rightarrow$ FlAIR | | FLAIR $\rightarrow$ T2 | | T1 $\rightarrow$ T1CE | | T1CE $\rightarrow$ T1 | | Avg. Dice↑ | Avg. ASD↓ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Dice↑ | ASD↓ | Dice↑ | ASD↓ | Dice↑ | ASD↓ | Dice↑ | ASD↓ | | |
| Source only | | 62.84 | 20.15 | 73.38 | 12.77 | 66.30 | 15.98 | 71.98 | 14.79 | 68.63 | 15.92 |
| Target only | ✓ | 88.30 | 4.53 | 88.78 | 5.08 | 76.48 | 9.38 | 78.47 | 8.76 | 83.01 | 6.93 |
| Tent [28] | | 69.55 | <u>11.91</u> | 74.06 | 12.72 | 72.99 | 11.41 | 74.93 | 10.10 | 72.88 | 11.54 |
| DPL [12] | | 69.35 | 12.13 | 74.40 | 11.97 | 72.09 | 11.28 | 73.81 | 10.67 | 72.41 | 11.51 |
| FSM [44] | | 70.00 | 13.06 | 74.70 | 12.27 | 72.10 | 11.55 | 73.87 | 10.70 | 72.67 | 11.89 |
| OSUAD [5] | | <u>73.69</u> | 12.26 | <u>77.06</u> | <u>11.91</u> | <u>75.26</u> | <u>10.32</u> | <u>77.19</u> | **9.43** | <u>75.80</u> | <u>10.98</u> |
| Ours | | **74.82** | **11.22** | **78.89** | **11.71** | **76.43** | **10.02** | **78.28** | <u>9.44</u> | **77.10** | **10.60** |

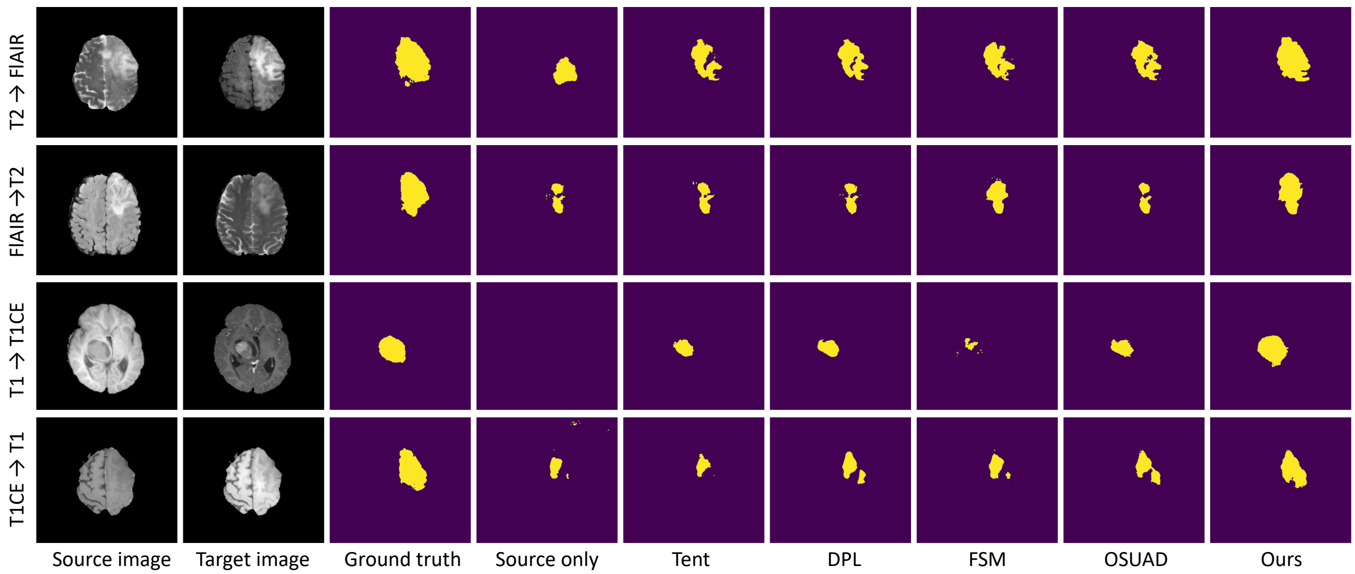


Fig. 5: Qualitative comparison of adaptation performance with different methods on brain tumor segmentation. The input of the adapted model is target domain images, and the first column shows the corresponding source domain images.

the Cup segmentation results show a notable enhancement with a 2.77% increase in Dice compared to other SFDA methods. The improvement in Disc segmentation is minimal due to its proximity to the upper bound, represented by the target-only model.

Overall, our approach outperforms all other methods on average, including the UDA model, highlighting its effectiveness in achieving high segmentation accuracy for Optic Disc and Cup. These improvements can primarily be attributed to the utilization of a teacher-student scheme, which facilitates the generation of more accurate PLs. Additionally, the application of the proposed denoising method is crucial in ensuring model stability and preventing the acquisition of detrimental knowledge.

To further validate the qualitative effectiveness of our proposed method, we provide visualizations of the segmentation in Figure 4. The first, third, and fourth rows demonstrate significant improvement, particularly in Cup segmentation. It can be observed that other models exhibit a higher number of false negatives in cup prediction, while our model accurately identifies the cup's location. This precise recognition ability is learned from high-quality pseudo-labels (PLs). The second row showcases improvements in both Disc and Cup segmentation. In contrast, other models either exhibit inaccuracies or irregular edges, which may be attributed to noise in the PLs. However, our superior denoising method enables our model to achieve smoother and more accurate segmentation.

TABLE III: Ablation study with different experimental settings on the fundus and BraTS datasets. This table presents the average performance of all scenarios on both datasets.

| Diversity Loss | Student Branch | Confidence Weighting | Direct Denoising | Prototypical Denoising | Fundus Dataset | | BraTS Dataset | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Dice↑ | ASD↓ | Dice↑ | ASD↓ |
| | | | | | 86.43 | 8.84 | 70.20 | 12.36 |
| ✓ | | | | | 86.48 | 9.04 | 72.41 | 12.24 |
| ✓ | ✓ | | | | 86.93 | 8.61 | 71.20 | 11.57 |
| ✓ | ✓ | ✓ | | | 87.78 | 7.65 | 71.51 | 11.63 |
| ✓ | ✓ | ✓ | ✓ | | 88.55 | 7.12 | 75.56 | 11.21 |
| ✓ | ✓ | ✓ | | ✓ | 88.64 | <u>6.75</u> | 75.84 | 10.74 |
| | ✓ | ✓ | ✓ | ✓ | <u>88.69</u> | 6.91 | 73.76 | 11.07 |
| ✓ | | | ✓ | ✓ | 87.66 | 7.37 | 76.62 | <u>10.65</u> |
| ✓ | ✓ | | ✓ | ✓ | 87.93 | 7.11 | <u>76.90</u> | 10.77 |
| ✓ | ✓ | ✓ | ✓ | ✓ | **89.28** | **6.52** | **77.10** | **10.60** |

*2) Results on Brain Tumor Segmentation:* Table II presents a comprehensive performance comparison for brain tumor segmentation. We retrained all SOTA methods using the experimental setting. A comprehensive performance comparison for brain tumor segmentation is presented in Table II. Our model consistently achieves the highest Dice values across all scenarios. In most cases, it also achieves the lowest ASD values, except for T1CE $\rightarrow$ T1, where it performs as the second-best. On average, our model outperforms all other models, demonstrating an improvement of 8.47% in Dice and 5.32 in ASD compared to the source-only model. These results emphasize the strong performance of our method compared

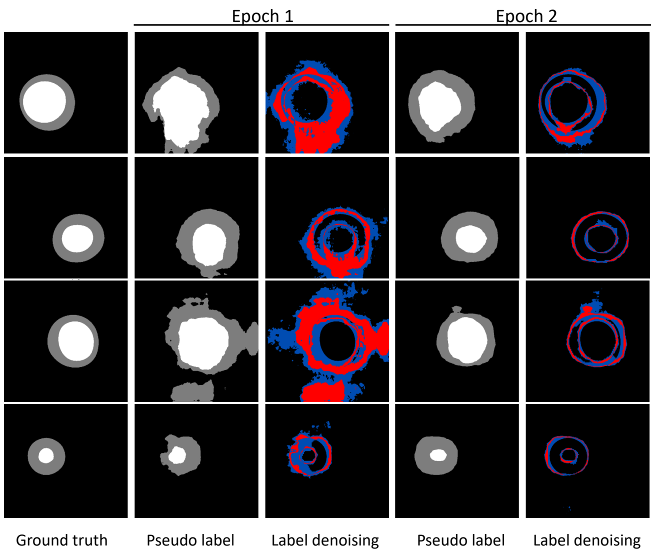


Fig. 6: Examples of pseudo-labels with higher weight during adaptation for fundus images. The correctly and falsely identified noisy pseudo-labels are indicated in red and blue colors.

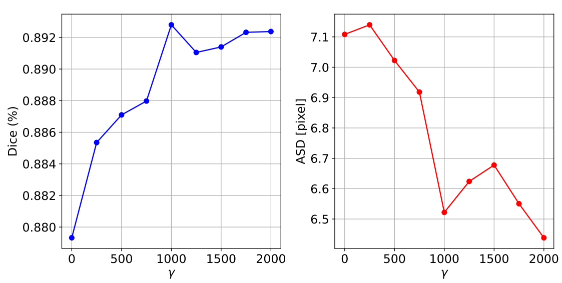


Fig. 7: Average segmentation performance of optic disc and cup on both RIM-ONE-r3 and Drishti-GS target domains with varying $\gamma$ values of the confidence-guided weighting module.

to other approaches. The notable improvements in our model can be attributed to the utilization of an enhanced denoising method and continuously improved PL. Moreover, the incorporation of a diversity loss term plays a crucial role in preventing overconfidence in the model's predictions.

To qualitatively validate the effectiveness of our method, we visualize the segmentation predictions in different scenarios as shown in Figure 5. In these samples, other methods generally demonstrate poor performance with numerous false negatives, whereas our method consistently produces more accurate brain tumor segmentations. Besides, our segmented regions closely resemble regular tumors with no hollow areas or outliers. In addition, the corresponding source images are displayed in the first column, revealing the evident change in object intensity between the source and target images, which highlights the capability of our method.

## E. Ablation Analysis of Key Components

*1) Effectiveness of Teacher-Student Joint Training Scheme:* The teacher-student Scheme can be splited into a student branch module and a confidence-guided weighting module. When the confidence-guided weighting is removed, $\gamma$ is set to 0, ensuring equal weighting between the student and teacher branches.

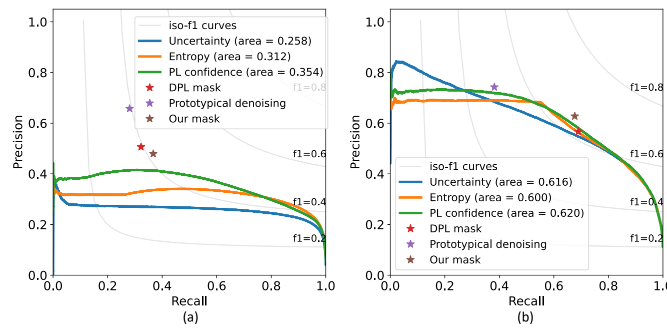


Fig. 8: (a) and (b) depict the precision and recall curves for the denoising performance of fundus and brain images, respectively.

For the fundus dataset, Table III (rows 3, 4, 9, and 10) demonstrates the performance enhancement achieved by incorporating the student branch into the model, and the addition of confidence-guided weighting further improves the model's performance. While the direct inclusion of the student branch yields performance improvements, confidence-guided weighting remains critical. It enables the model to select high-quality PLs for learning and suppresses the impact caused by poor-quality PLs. Figure 7 illustrates the variations in segmentation performance on the fundus dataset for different $\gamma$ values. The graph shows that the segmentation performance continuously improves before $\gamma$ reaches 1000. Subsequently, further increases in $\gamma$ do not significantly boost the model's performance. This is because when $\gamma$ is small, the weight difference is small, thus the effect is not significant. This is because when $\gamma$ is small, the weight difference is small, thus the effect of the scheme is not significant.

For the BraTs dataset, the teacher-student scheme enhances the model's performance (0.58% improvement in Dice) when combined with denoising. However, incorporating the teacher-student scheme without denoising yields negligible improvements. This can be attributed to the higher level of noise present in the modality adaptation within the BraTS dataset. Additionally, the results in rows 9 and 10 of Table III indicate that the BraTS dataset results are not sensitive to changes in $\gamma$.

Fig. 6 displays the PLs generated by the student branch at different adaptation stages. The quality of the PLs progressively improves throughout the adaptation process, accompanied by a reduction in the denoised area. This iterative refinement provides the model with more accurate and fine domain knowledge.

TABLE IV: Average performance of different denoising methods on all scenarios. To calculate the following results, we set the threshold for entropy, uncertainty, and Chebyshev confidence to 0.1, 0.05, and 0.05, respectively.

| Method | F1 score↑ | Accuracy↑ |
|---|---|---|
| Entropy | <u>0.3861</u> $\pm$ 0.0633 | 0.9568 $\pm$ 0.0194 |
| Uncertainty | 0.2704 $\pm$ 0.1861 | 0.9741 $\pm$ 0.0180 |
| Prototypical denoising | 0.2739 $\pm$ 0.1665 | **0.9796** $\pm$ 0.0052 |
| DPL mask | 0.3721 $\pm$ 0.0836 | 0.9731 $\pm$ 0.0136 |
| Our mask | **0.4054** $\pm$ 0.0606 | <u>0.9767</u> $\pm$ 0.0083 |

*2) Effectiveness of Proposed Denoising Method:* Comparing the 2rd and 8th rows, as well as the 4th and 10th rows, of Table III reveals a significant performance improvement when incorporating the proposed denoising method. Specifically, there is an increase of approximately 1.83% and 4.90% in the Dice coefficient on average for the fundus and BraTs datasets, respectively. Notably, the denoising method shows greater significance on the BraTs dataset, possibly due to the larger domain shift that necessitates noise reduction in PLs.

The 5th and 6th rows demonstrate that incorporating either direct denoising or prototypical denoising individually leads to performance enhancements. The simultaneous utilization of both denoising techniques further amplifies the model's performance for both the fundus and BraTs datasets. This finding suggests a certain degree of complementarity between direct denoising and prototypical denoising, and their simultaneous use yields better denoising outcomes.

*3) Comparison of Denoising Methods:* To further evaluate our proposed denoising method, we compared it with commonly used denoising methods, including entropy, uncertainty, and the DPL mask [12]. Denoising can be seen as a binary classification task, where the objective is to classify correct PLs from incorrect ones. Hence, evaluation metrics from binary classification can be employed.

Table IV showcases the performance of different denoising methods based on F1 score and Accuracy. It can be observed that Our mask achieves the highest F1 score, surpassing the second-best method by 1.93%. Furthermore, Our mask achieves a comparable accuracy to the best method, with only a marginal difference of 0.29%. The high F1 score and accuracy indicate that our denoising method performs well in classifying both noisy and non-noisy regions.

The precision and recall curves shown in Fig. 8 were generated by randomly selecting 16 images from each dataset to evaluate different denoising methods. From Fig. 8, it can be observed that our proposed Chebyshev confidence outperforms uncertainty and entropy in terms of classification performance, with respective area values of 0.354 and 0.620 for the fundus and BraTs datasets. This can be attributed to the improved assessment of PL reliability provided by our designed Chebyshev confidence. Furthermore, we can observe that prototypical denoising achieves the highest precision at the same recall value. This highlights the superiority of the prototypical approach. However, it should be noted that prototypical denoising generates a binary mask that cannot be adjusted by manipulating the threshold. Therefore, the combination of prototypical denoising and direct denoising using Chebyshev confidence provides a more flexible denoising approach. We believe that this is one of the reasons why prototypical denoising and direct denoising can complement each other.

*4) Effectiveness of Diversity Loss:* We also investigated the influence of diversity loss on the model's performance. The results, as shown in the second and seventh rows of Table III, indicate that incorporating diversity loss yields performance enhancements for both datasets. Specifically, for the fundus dataset, the model demonstrates improvements of 0.32 in Dice coefficient and 0.10 in ASD on average. In the case of the BraTs dataset, the observed improvements are more pronounced, with gains of 2.78 in Dice coefficient and 0.3 in ASD. The significant impact of diversity loss on the BraTs dataset can be attributed to its longer adaptation over 10 epochs, which may exacerbate the issue of overconfidence. In such scenarios, the inclusion of diversity loss becomes crucial as a means to prevent overconfidence and learning incorrect knowledge.

## V. Conclusion

In this paper, we propose a SFDA framework that aims to enhance the accuracy of pseudo-labels and effectively remove noise. The SFDA framework addresses privacy concerns by transferring knowledge only from a well-trained source model to the target domain. We introduce a technique for estimating the Chebyshev confidence of pseudo-labels, which is independent of hyper-parameters and precisely evaluates the reliability of pseudo-labels. Based on the estimated confidence, our proposed teacher-student joint training scheme continuously improves the accuracy of pseudo-labels and prioritizes high-quality ones. For pseudo-label denoising, we propose two confidence-guided methods: direct denoising and prototypical denoising. These methods utilize pixel and category information, respectively, to remove noise. Additionally, we incorporate a diversity loss term to prevent overconfidence. By combining these modules, our framework generate more accurate and cleaner pseudo-labels for self-training, resulting in stable adaptation and better performance. We have conducted experiments on different scenarios, including cross-centre and cross-modality adaptation, and the results have demonstrated the superiority of our framework.

## References

[1] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

[2] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[3] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015.

[4] Chen, Liang-Chieh, et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." IEEE transactions on pattern analysis and machine intelligence 40.4 (2017): 834-848.

[5] Liu, Xiaofeng, et al. "Memory consistent unsupervised off-the-shelf model adaptation for source-relaxed medical image segmentation." Medical Image Analysis 83 (2023): 102641.

[6] Liu, Hong, Jianmin Wang, and Mingsheng Long. "Cycle self-training for domain adaptation." Advances in Neural Information Processing Systems 34 (2021): 22968-22981.

[7] Gu, Xiang, Jian Sun, and Zongben Xu. "Spherical space domain adaptation with robust pseudo-label loss." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[8] Hoffman, Judy, et al. "Cycada: Cycle-consistent adversarial domain adaptation." International conference on machine learning. Pmlr, 2018.

[9] Kundu, Jogendra Nath, Naveen Venkat, and R. Venkatesh Babu. "Universal source-free domain adaptation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[10] Liang, Jian, Dapeng Hu, and Jiashi Feng. "Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation." International Conference on Machine Learning. PMLR, 2020.

[11] Bateson, Mathilde, et al. "Source-relaxed domain adaptation for image segmentation." Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23. Springer International Publishing, 2020.

[12] Chen, Cheng, et al. ”Source-free domain adaptive fundus image segmentation with denoised pseudo-labeling.” Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. Springer International Publishing, 2021.

[13] Yang, Shiqi, et al. ”Generalized source-free domain adaptation.” Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[14] Chen, Dian, et al. ”Contrastive test-time adaptation.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[15] Karim, Nazmul, et al. ”C-SFDA: A Curriculum Learning Aided Self-Training Framework for Efficient Source Free Domain Adaptation.” arXiv preprint arXiv:2303.17132 (2023).

[16] Wang, Mei, and Weihong Deng. ”Deep visual domain adaptation: A survey.” Neurocomputing 312 (2018): 135-153.

[17] uan, Hao, and Mingxia Liu. ”Domain adaptation for medical image analysis: a survey.” IEEE Transactions on Biomedical Engineering 69.3 (2021): 1173-1185.

[18] Long, Mingsheng, et al. ”Conditional adversarial domain adaptation.” Advances in neural information processing systems 31 (2018).

[19] Tzeng, Eric, et al. ”Adversarial discriminative domain adaptation.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[20] Lee, Hsin-Ying, et al. ”Diverse image-to-image translation via disentangled representations.” Proceedings of the European conference on computer vision (ECCV). 2018.

[21] Murez, Zak, et al. ”Image to image translation for domain adaptation.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[22] Chen, Chao, et al. ”Homm: Higher-order moment matching for unsupervised domain adaptation.” Proceedings of the AAAI conference on artificial intelligence. Vol. 34. No. 04. 2020.

[23] Abuduweili, Abulikemu, et al. ”Adaptive consistency regularization for semi-supervised transfer learning.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[24] Shen, Jian, et al. ”Wasserstein distance guided representation learning for domain adaptation.” Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. 2018.

[25] Zou, Yang, et al. ”Unsupervised domain adaptation for semantic segmentation via class-balanced self-training.” Proceedings of the European conference on computer vision (ECCV). 2018.

[26] Toldo, Marco, et al. ”Unsupervised domain adaptation in semantic segmentation: a review.” Technologies 8.2 (2020): 35.

[27] Mei, Ke, et al. ”Instance adaptive self-training for unsupervised domain adaptation.” Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16. Springer International Publishing, 2020.

[28] Wang, Dequan, et al. ”Tent: Fully test-time adaptation by entropy minimization.” arXiv preprint arXiv:2006.10726 (2020).

[29] Xu, Zhe, et al. ”Denoising for relaxing: Unsupervised domain adaptive fundus image segmentation without source data.” Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V. Cham: Springer Nature Switzerland, 2022.

[30] Litrico, Mattia, Alessio Del Bue, and Pietro Morerio. ”Guiding Pseudo-labels with Uncertainty Estimation for Test-Time Adaptation.” arXiv preprint arXiv:2303.03770 (2023).

[31] Rizve, Mamshad Nayeem, et al. ”In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning.” arXiv preprint arXiv:2101.06329 (2021).

[32] Sohn, Kihyuk, et al. ”Fixmatch: Simplifying semi-supervised learning with consistency and confidence.” Advances in neural information processing systems 33 (2020): 596-608.

[33] Lai, Xin, et al. ”DecoupleNet: Decoupled network for domain adaptive semantic segmentation.” Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII. Cham: Springer Nature Switzerland, 2022.

[34] Long, Mingsheng, et al. ”Conditional adversarial domain adaptation.” Advances in neural information processing systems 31 (2018).

[35] Wang, Yuxi, Junran Peng, and ZhaoXiang Zhang. ”Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation.” Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[36] Gal, Yarin, and Zoubin Ghahramani. ”Dropout as a bayesian approximation: Representing model uncertainty in deep learning.” international conference on machine learning. PMLR, 2016.

[37] Zhang, Pan, et al. ”Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[38] Chen, Liang-Chieh, et al. ”Encoder-decoder with atrous separable convolution for semantic image segmentation.” Proceedings of the European conference on computer vision (ECCV). 2018.

[39] Orlando, José Ignacio, et al. ”Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs.” Medical image analysis 59 (2020): 101570.

[40] Fumero, Francisco, et al. ”RIM-ONE: An open retinal image database for optic nerve evaluation.” 2011 24th international symposium on computer-based medical systems (CBMS). IEEE, 2011.

[41] Sivaswamy, Jayanthi, et al. ”A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis.” JSM Biomedical Imaging Data Papers 2.1 (2015): 1004.

[42] Bakas, Spyridon, et al. ”Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge.” arXiv preprint arXiv:1811.02629 (2018).

[43] Deng, Jia, et al. ”Imagenet: A large-scale hierarchical image database.” 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.

[44] Yang, Chen, et al. ”Source free domain adaptation for medical image segmentation with fourier style mining.” Medical Image Analysis 79 (2022): 102457.

[45] Wang, Shujun, et al. ”Boundary and entropy-driven adversarial learning for fundus image segmentation.” Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22. Springer International Publishing, 2019.

[46] Sandler, Mark, et al. ”Mobilenetv2: Inverted residuals and linear bottlenecks.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[47] Yang, Dong, et al. ”Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan.” Medical image analysis 70 (2021): 101992.

[48] Chen, Cheng, et al. ”Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation.” IEEE transactions on medical imaging 39.7 (2020): 2494-2505.

[49] Xing, Fuyong, Tell Bennett, and Debashis Ghosh. ”Adversarial domain adaptation and pseudo-labeling for cross-modality microscopy image quantification.” Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22. Springer International Publishing, 2019.