Unsupervised Domain Adaptive Fundus Image Segmentation with Few Labeled Source Data

Qianbi Yu
qiyu2312@uni.sydney.edu.au
Dongnan Liu
dongnan.liu@sydney.edu.au
Chaoyi Zhang
czha5168@uni.sydney.edu.au
Xinwen Zhang
xzha9145@uni.sydney.edu.au
Weidong Cai
tom.cai@sydney.edu.au

School of Computer Science
University of Sydney
Sydney, Australia

Abstract

Deep learning-based segmentation methods have been widely employed for automatic glaucoma diagnosis and prognosis. In practice, fundus images obtained by different fundus cameras vary significantly in terms of illumination and intensity. Although recent unsupervised domain adaptation (UDA) methods enhance the models' generalization ability on the unlabeled target fundus datasets, they always require sufficient labeled data from the source domain, bringing auxiliary data acquisition and annotation costs. To further facilitate the data efficiency of the cross-domain segmentation methods on the fundus images, we explore UDA optic disc and cup segmentation problems using few labeled source data in this work. We first design a Searching-based Multi-style Invariant Mechanism to diversify the source data style as well as increase the data amount. Next, a prototype consistency mechanism on the foreground objects is proposed to facilitate the feature alignment for each kind of tissue under different image styles. Moreover, a cross-style self-supervised learning stage is further designed to improve the segmentation performance on the target images. Our method has outperformed several state-of-the-art UDA segmentation methods under the UDA fundus segmentation with few labeled source data.

1 Introduction

Glaucoma is a chronic eye condition that causes progressive damage to the optic nerve and eventually leads to blindness if left untreated [1]. In clinical practice, accurate examination of the head of the optic nerve i.e. cup-to-disc ratio is crucial for early detection and treatment of glaucoma diagnosis [26]. Recently, deep learning-based models have been widely used for

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

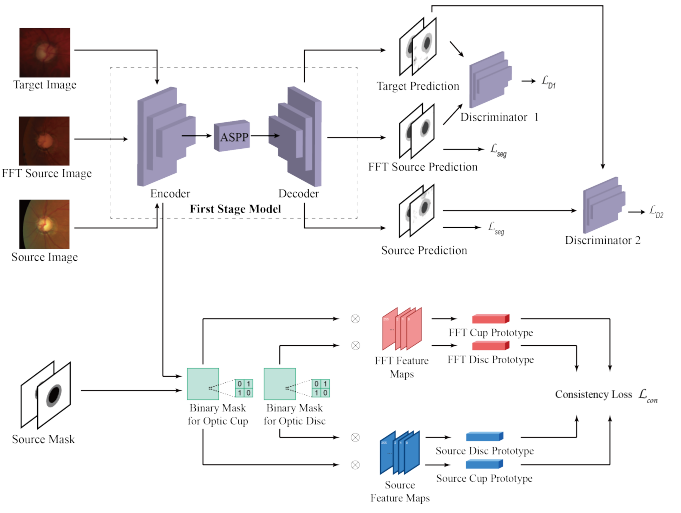

Figure 1: Overview of the proposed framework for cross-domain optic disc and cup segmentation with few labeled data. The first stage model is trained using target and synthesis images with additional discriminative and consistency loss. The customized consistency loss utilizes the mask pooling technique and computes the cosine similarity of the class-prototype vectors.

automatic optic cup and disc segmentation in fundus images [6, 30] and achieved appealing performance. Nevertheless, these methods will suffer from performance drop when validated on new datasets with unseen distributions due to the domain shift issue [8, 28]. To this end, unsupervised domain adaptation (UDA) methods have been proposed to enhance the models' generalization ability, by transferring the knowledge from the labeled source data to the unlabeled target data [10, 14, 17, 21, 22, 24].

Recently, several methods have been proposed to further facilitate the data-efficiency of the UDA medical image segmentation [3, 18, 43]. [18] proposed to train the UDA model with few target images, and [3] explored a source-free UDA setting without sharing the raw source data with the target domain. However, they still require the model to be optimized with sufficient labeled source images, which might be challenging to fulfill in practical applications. Acquiring the pixel-level annotations for the optic discs and cups is time-consuming and error-prone for automatic fundus segmentation datasets [34]. On the other side, directly training the UDA models with insufficient labeled source data can easily cause over-fitting problems and limit the segmentation performance. To further reduce the amount of data required and maintain high-performance cross-domain segmentation, we explore the UDA fundus image segmentation problem with few labeled source data. Although [43] also explores a similar problem as ours via a teacher-student framework, they ignore the category relationship of the foreground tissues. Given the extremely insufficient supervision (e.g., given no more than 10% of the labeled training data), the model would suffer from misalignment due to the lack of semantic-level information learning.

In this work, a novel framework is proposed to transfer the knowledge from the extremely limited labeled source data to the unlabeled target data for cross-domain optic disc and cup segmentation. First, we propose to diversify the domain knowledge by a Searching-based Multi-Style Invariant (SMSI) mechanism, enriching the image distributions by creat-

ing transformed styles based on the synthesized images through searching strategies. Secondly, considering the similarity of the foreground content within the same category under various styles, a new Class-Prototype Consistency (CPC) mechanism is also introduced. Moreover, we design a Cross-Style Self-supervised Learning (CSSL) strategy with pseudo labels to further boost the overall segmentation of the unlabeled target images. Our proposed method is validated on two cross-domain optic disc and cup segmentation experiments with limited labeled source data available. By outperforming other state-of-the-art UDA segmentation methods, our proposed framework is demonstrated to be more effective and can conquer the domain shift under low resource situations, which is, therefore, more practical and important in real-world applications.

2 Related Work

Unsupervised domain adaptation Performing pixel-level domain mapping using image-to-image translation is a typical solution to reduce the domain gap at the appearance level [13, 45]. In addition, feature-level adaptation can also alleviate the cross-domain discrepancy by inducing domain-invariant features learning [9, 21, 32, 33, 41]. Various methods have involved generative adversarial training, but non-GAN-based techniques have also achieved competitive results, especially those via frequency space learning [12, 38, 40]. These methods mostly conduct frequency alignment or frequency modification to achieve image stylization. By introducing little extra computations to the framework, the frequency learning-based methods can achieve style transformation in a more efficient manner than the GAN-based ones. However, current frequency space methods heavily rely on non-learnable parameter selection, such as parameter β in [38] and parameter p in [12]. To avoid massive experiments for selecting the appropriate parameters for the improved synthesized images, we design a SMSI module based on AutoML techniques, which have been widely investigated for efficient medical image analysis [29, 36, 37].

3 Methodology

3.1 Searching-based Multi-Style Invariant Mechanism (SMSI)

To alleviate the domain gap at the appearance level as well as enlarge the data-scarce source domain, we propose a Searching-based Multi-Style Invariant Mechanism (SMSI) for the source domain based on Fourier transform [38]. Specifically, each channel of an input image x is firstly transformed into the frequency space $\mathcal{F}(x)$ via: $\mathcal{F}(x) = \sum_{h,w} x(h,w) e^{-j2\pi(\frac{h}{H}m + \frac{w}{W}n)}$, where $j^2 = -1$. Next, this frequency signal can be further decomposed into an amplitude spectrum $\mathcal{F}^A$ and a phase spectrum $\mathcal{F}^P$, which respectively represent the low-level (e.g., appearance) and high-level (e.g., content) characteristics of each image [15, 38].

By obtaining the amplitude $\mathcal{F}^A_s$ and phase $\mathcal{F}^P_s$ from each source image x_s, its corresponding synthesis image in the target-like style can be generated via:

$$X_{s\to t} = \mathcal{F}^{-1}[(M_\beta \overline{\mathcal{F}^A_t} + (1 - M_\beta)\mathcal{F}^A_s, \mathcal{F}^P_s], \tag{1}$$

where the $\mathcal{F}^{-1}$ is the inverse Fourier transform. To ensure that each synthesis image contains comprehensive appearance-level information under the target distributions, we propose to replace the $\mathcal{F}^A_s$ with the average amplitude spectrum from all the target images, denote as

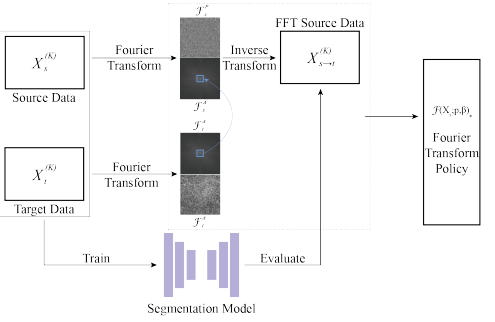

Figure 2: K-folds automatic Fourier transform searching algorithm. Source dataset X_s and target dataset X_t are split in to k-folds, Fourier transform with different proportion parameter β applied to the source dataset, segmentation network trained with target dataset is used to evaluate the transformed data $X_{s\to t}$. The parameters selected from each fold are appended to the final Fourier transform policy list $\mathcal{F}_*$.

$\overline{\mathcal{F}_t^A}$. M_β is used to control the proportion of the target amplitude during synthesis and controlled by a parameter $\beta \in (0,1)$, defined as $M_\beta = \mathcal{I}_{(h,w)\in[-\beta H:\beta H,-\beta W:\beta W]}$. As indicated in previous works [38], different M_β choices can induce distinct domain adaption performance. However, it is cost-intensive to conduct massive experiments for selecting the appropriate parameters for each specific application scenario. To tackle this issue, we propose an efficient searching strategy to find the optimal parameters for the synthesis images which can achieve better cross-domain segmentation performance.

Specifically, the search space is first defined as $\mathcal{F}(X_s;\beta)$. Given the above search space, the search processes are formulated as: (i) Train a plain segmentation model with original source data X_s. (ii) Initialized by the model from step (i), several models are further optimized on K groups of synthesis images according to Equation 1. In each group, the β for M_β is randomly initialized within $(0,1)$. (iii) Let each model in step (ii) learn the ideal β by searching controllers and appending to the final policy set following [4, 19]. As the Fourier transformation is only changing the style of each image, instead of its content, the segmentation ground truth of the synthesis images is similar to the original one [38]. Therefore, the objective function of the policy search is designed to maximize the validation dice on transformed data $X_{s\to t}$ with original source label:

$$\mathcal{F}_* = argmax\, \mathcal{D}(\theta_G|\mathcal{F}(X_s)), \tag{2}$$

where θ_G is the parameter of the segmentation network used to optimize $\mathcal{L}_{seg}$ and $\mathcal{D}$ is the validation dice. The search controller can be implemented efficiently using the Tree-structured Parzen Estimators algorithm in [2]. Figure 2 indicates the detailed process of the SMSI mechanism. After optimal parameters are determined, each source image can generate $n \times k$ synthesis image by varying the n number of β parameters that control the likelihood of $s \to t$ and taking k average amplitude values from different parts of the target images. This procedure generally expands the source domain dataset and provides a useful regularization technique to increase the diversity of the dataset.

Although training the models with the synthesis images can alleviate the domain gap at the appearance levels, it can still incur domain shifts at the feature level [11, 22]. As such, we introduce feature invariant induction learning based on the searching-based multi-style

synthesis process. Specifically, additional adversarial domain discriminators are utilized to generate domain-invariant features for the synthesis images and target images on top of the traditional supervised loss. Denote that the source dataset is $X_s \subset R^{H\times W\times 3}$ with ground truth C-class segmentation maps $Y_s \subset (1,C)^{H\times W}$, synthesis source dataset is $X_{s\to t}$ with the same ground truth maps. The target dataset is X_t with no ground truth label. A discriminator θ_D is trained adversarially to distinguish between the synthesis source set and target set with discrimination loss $\mathcal{L}_D$. Simultaneously, the segmentation network is trained to fool the discriminator as:

$$\min_{\theta_D} \frac{1}{|X_{s\to t}|}\sum_{x_{s\to t}} \mathcal{L}_D(I_{x_{s\to t}},1) + \frac{1}{|X_t|}\sum_{x_t}\mathcal{L}_D(I_{x_t},0) \qquad \min_{\theta_G} \frac{1}{|X_t|}\sum_{x_t}\mathcal{L}_D(I_{x_t},1) \tag{3}$$

where $I_{x_{s\to t}}$ and I_{x_t} are the weighted self-information maps following ADVENT [33]. Summarily, there are two discriminators D_1 and D_2 implemented to distinguish (i) $X_{s\to t}$ and X_t (ii) X_s and X_t as indicated in the top part of Figure 1.

3.2 Class-Prototype Consistency Mechanism (CPC)

The synthesis images produced from the SMSI mechanism and their corresponding source images should have the same image content but in different styles. Motivated by previous works that the class-aware features under the same category should maintain the same across different domains [44], we propose a class-prototype consistency mechanism for the synthesis images. The class prototype is created by using the high-level feature maps from the model's encoder and the ground truth source masks. The source masks are first resized and converted into binary masks for each class. Then, they are multiplied by the feature maps extracted from synthesis images and source images respectively, generating class-relevant masked feature maps. Taking a global average pooling further converts the feature maps into class prototype vectors. Global average pooling has the ability to sum out the spatial data and enforce the correspondences between feature maps and classes. Denote the mask of class c as M^c and f_s, $f_{s\to t}$ as the feature maps, the class prototypes of c class are defined as:

$$p_s^c = \frac{1}{N_{h\times w}}\sum_c M^c f_s \qquad p_{s\to t}^c = \frac{1}{N_{h\times w}}\sum_c M^c f_{s\to t} \tag{4}$$

where h and w are the height and width of the feature maps. This masked pooling technique enables the network to focus on the target content of images instead of intensity and illumination variation. To narrow the gap between the features under the same class in different synthesis domains, we propose to enlarge the similarity between them. The overall CPC Mechanism is illustrated in the bottom part of Figure 1, where binary masks for optic cups and discs are used. Specifically, the overall consistency loss function can be defined as:

$$\mathcal{L}_{con} = \sum_{c=(0,1)} \left(1 - \frac{p_s^c \cdot p_{s\to t}^c}{max(||p_s^c||_2 \cdot ||p_{s\to t}^c||_2, \varepsilon)}\right), \tag{5}$$

where ε is the small value to avoid division by zero, p_s^c and $p_{s\to t}^c$ are the prototypes of the class c for the features from the FFT synthesized images and source images. The overall optimization function for the segmentation network with SMSI and CPC mechanism is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{seg}(X_s,Y_s) + \mathcal{L}_{seg}(X_{s\to t},Y_s) + \lambda(\mathcal{L}_{D_1}+\mathcal{L}_{D_2}) + \mathcal{L}_{con}. \tag{6}$$

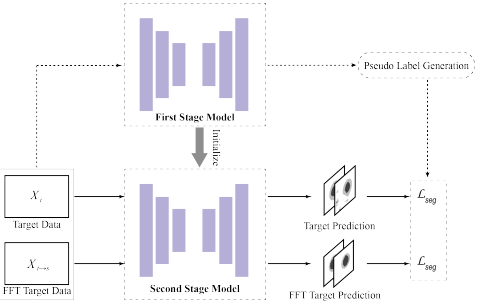

Figure 3: Demonstration of the pseudo label self-supervised learning process. Specifically, a second stage model θ_G^2 is initialized by θ_G^1 and trained over the target data X_t and the FFT target data $X_{t\to s}$ with pseudo labels.

3.3 Cross-Style Self-supervised Learning (CSSL)

When dealing with model adaptation towards the target domain, a consummate resource is the target ground truth masks, which are not available in UDA settings. As compensation, highly-confident pseudo labels can be created for unlabeled target images by using prediction probabilities P^v from the trained model θ_G^1 on the v-th pixel. The pseudo labels can be defined as $\hat{y}_t^v = I[P^v \geq \gamma]$, where I is the indicator function and $\gamma \in (0,1)$ is the probability threshold to determine the binary mask. However, solely training the model with the target pseudo labels brings noise to the optimization process due to the gap between the pseudo and real labels. To stabilize the training process, we propose a cross-style self-supervised learning strategy, to jointly re-train the model using the target images X_t and the Fourier transformed target images with source-like styles $X_{t\to s}$ and their pseudo labels $\hat{y}_t^v$, and $\hat{y}_{t\to s}^v$, respectively. The source-like synthesized images are obtained following the process in Section 2.1. Since the segmentation learning for the model in the first stage is based on the annotated source data, the pseudo labels for the source-like synthesis images contain less noise and therefore can be a complement to the target supervised loss. Figure 3 demonstrates these two supervised segmentation losses. In general, the model learns combined distribution and gets further improvement using self-supervised training in the second stage.

4 Experiments

4.1 Datasets and Implementation Details

The experiments aim to segment the cup and disc components in the multi-center fundus images, which are obtained from various patients using different eye examination equipment. The source dataset REFUGE [27] contains 400 annotated images. There are two target datasets, RIM ONE-r3 [7] contains 99 training images and 60 testing images, and Drishti-GS [31] contains 50 training images and 51 testing images. All datasets used are publicly available. In our experiments with few labeled source data, only 10 random source images are accessible during model training. For the image synthesis process in the first stage, $n = 3$ and $k = 5$ are selected for each source image, with a total of 150 FFT source images generated. In the second stage, $n = 3$ and $k = 1$ are selected for each target image, 150 and 297 FFT target images are generated respectively. The experiment results under other

selections for the source images as well as more detailed experimental and implementation settings are available in the supplementary material.

The network used in our experiments is a MobileNetv2 with a DeepLabv3+ backbone based on the structure in [34]. A semantic segmentation module called Atrous Spatial Pyramid Pooling (ASPP) in DeepLab re-samples a given feature layer at various rates before convolution. The overall model size is 7.62M and the inference time is 30.57s for one image. In the first stage, the segmentation model is trained with Adam optimizer under a 1e-3 learning rate, the discriminators have been trained with SGD optimizer with a 2.5e-5 learning rate, 8 batch size, and 200 training epochs. The weighting factor λ in Equation 6 is set as 0.5. In the second stage, the segmentation model is trained with Adam optimizer with a 2e-3 learning rate, 8 batch size, and 20 training epochs. The probability threshold $\gamma = 0.75$ is used to generate the pseudo labels. Segmentation results are evaluated by the Dice coefficient and Average Surface Distance (ASD). The framework is implemented on Pytorch 1.7.1 using a NVIDIA RTX3090 GPU.

4.2 Comparison Experiments

The proposed method is compared to state-of-the-art (SOTA) unsupervised domain adaptation methods, as well as two recent UDA approaches particularly for few labeled source images. In supplementary material, extra SOTA fundus image segmentation methods [6, 20, 23, 35, 42] are also compared. CyCADA [11] translates the source images into the target style using cycle-consistent adversarial networks and trains the adversarial network with the translated images. AdvEnt [33] brings in entropy loss and adversarial loss respectively to address the domain shift problem. FDA [38] adopts frequency swap method for image stylization and evaluate the segmentation model with multi-band transfer. PixMatch [25] develops a new component to ensure that the model's predictions on a target image and a perturbed version of the same image are pixel-wise consistent. LTIR [16] learns texture invariant features from different domains using Style-Wrap to change the images' appearance. Consider from another perspective, the two recently-developed methods MT [43] and PCS [39] have a similar experimental setting, they both focus on domain adaptation with few source data. MT follows the mean teacher paradigm and adopts dual teacher models to provide both semantic and structural knowledge to the student model, whereas PCS performs in-domain and cross-domain learning using prototypes from feature memory banks. Some other latest fundus image segmentation baselines are also evaluated. BEAL [34] suggests boundary prediction and entropy-driven during adversarial training and achieves excellent results for cross-domain prediction. DPL [3] is a novel proposal for source-free domain adaptation in the field of fundus image segmentation, with a pseudo-label denoising technique. It utilizes a pre-trained source model to generate pseudo-labels. For comparison, we follow the same experimental settings in these segmentation models, i.e., only 10 randomly selected source images will be accessible throughout the experiment, even for the pre-trained model.

Quantitative analysis. As indicated in Table 1, the segmentation performance of all other comparison UDA approaches is at the same level. This indicates their adaptation abilities are limited due to the lack of sufficient supervision learning. For the MT [43] and PCS [39] which were originally designed for UDA with few labeled source data, we notice that their performance is suboptimal. For MT, the lack of consideration of the cross-domain category information makes the model learn insufficient semantic-level knowledge given the extremely limited labeled source data for segmentation supervision learning, which further incurs inferior performance on the target testing data. Although PCS proposes a class-aware

Methods	Dice Metric [%] ↑			ASD Metric [pixel] ↓		
	Cup	Disc	Average	Cup	Disc	Average
RIM-ONE-r3						
CyCADA [11]	69.94 (0.58)	72.50 (0.37)	71.32 (0.48)	19.55 (0.22)	37.00 (0.35)	28.28 (0.29)
AdvEnt [33]	67.73 (0.63)	78.54 (1.19)	73.34 (0.91)	30.96 (1.79)	32.48 (1.55)	31.72 (1.67)
FDA [38]	69.38 (0.20)	78.07 (1.10)	73.72 (0.65)	21.15 (1.47)	28.86 (1.41)	25.01 (1.44)
PixMatch [25]	64.91 (1.14)	75.88 (1.93)	70.39 (1.53)	18.60 (0.39)	30.50 (1.42)	24.55 (0.91)
LTIR [16]	65.84 (0.09)	78.01 (1.97)	71.92 (1.03)	24.72 (0.98)	29.51 (1.46)	27.11 (1.22)
MT [43]	63.50 (1.51)	67.92 (1.05)	65.71 (1.28)	20.72 (1.88)	39.26 (1.86)	29.99 (1.87)
PCS [39]	60.77 (1.19)	73.90 (1.50)	67.33 (1.35)	24.20 (1.86)	32.13 (1.51)	28.16 (1.69)
BEAL [34]	67.69 (1.49)	78.88 (1.23)	73.29 (1.36)	21.36 (1.93)	34.95 (1.79)	28.16 (1.86)
DPL [8]	68.58 (0.33)	87.61 (0.61)	78.02 (0.47)	12.46 (0.59)	19.01 (0.45)	15.74 (0.52)
Ours	**78.16 (0.96)**	**88.45 (0.36)**	**83.30 (0.66)**	**9.82 (0.65)**	**11.78 (0.75)**	**10.80 (0.70)**
Drishti-GS						
CyCADA [11]	78.85 (0.80)	92.15 (0.52)	85.50 (0.66)	15.33 (0.30)	12.84 (1.85)	14.09 (1.08)
AdvEnt [33]	79.17 (0.79)	91.47 (0.47)	85.32 (0.63)	15.42 (0.85)	15.08 (1.29)	15.25 (1.07)
FDA [38]	83.57 (0.40)	94.13 (0.67)	88.85 (0.54)	12.40 (0.25)	7.68 (0.50)	10.04 (0.38)
PixMatch [25]	76.45 (1.60)	91.97 (0.47)	84.21 (1.04)	17.06 (1.13)	10.92 (1.04)	13.99 (1.09)
LTIR [16]	80.82 (1.00)	92.86 (0.73)	86.84 (0.87)	13.29 (0.74)	8.83 (0.94)	11.06 (0.84)
MT [43]	72.82 (1.54)	90.37 (0.99)	81.59 (1.27)	19.43 (0.30)	13.11 (1.73)	16.39 (1.02)
PCS [39]	74.88 (1.19)	88.33 (0.03)	81.21 (0.61)	19.81 (1.45)	18.90 (0.78)	19.56 (1.12)
BEAL [34]	75.91 (1.29)	93.44 (0.36)	84.68 (0.83)	17.48 (1.47)	10.42 (0.36)	13.95 (0.92)
DPL [8]	78.60 (0.17)	95.28 (0.82)	86.94 (0.50)	19.99 (0.82)	6.07 (0.16)	13.03 (0.49)
Ours	**83.64(0.20)**	**95.47(0.23)**	**89.56(0.22)**	**11.04(0.51)**	**5.25(0.25)**	**8.14(0.38)**

Table 1: Experimental results for comparison of different domain adaptation approaches in terms of Dice and ASD metrics on RIM-ONE-r3 and Drishti-GS target datasets. The numbers shown in the table are average values, taken from three sets of experiments under different groups of randomly selected 10 labeled source images. The numbers in parenthesis are the SD values.

UDA framework, it was particularly designed for UDA classification under the small domain gap. When validated on the UDA fundus image segmentation with a large domain bias, its segmentation results are limited by ignoring the appearance-level domain bias and the particular designs for segmentation. On the other hand, our method can tackle the aforementioned challenges by the SMSI for appearance-level adaption, CPC for cross-domain category-aware information processing, and the CSSL for further performance gain without auxiliary annotations. Overall, our method has outperformed others, achieving 6.22% Dice, 10.71 ASD pixel for RIM-ONE- r3 and 4.44% Dice, 5.92 ASD pixel for Drishti-Gs. We have also conducted a two-tailed paired t-test on the comparison studies, and given the p-value smaller than 0.01, our improvements are statistically significant.

Qualitative analysis. As presented in Figure 4, the segmentation results of some experiments show that focusing on content rather than appearance enables the network to better distinguish target objects from irrelevant backgrounds. The segmentation predictions from several comparison methods are significantly distracted by the background noise. Additionally, regardless of domain differences, the network faces difficulties when attempting to determine the spatial prior knowledge of the optic disc and optic cup. Our predictions alleviate these issues, have a much clear boundary between the cup and disc, and exhibit much fewer background segmentation error.

4.3 Ablation Studies

Ablation studies are conducted to evaluate the effectiveness of our proposed modules. In Table 2, the source-only and target-only experiments provide lower and upper bounds of this

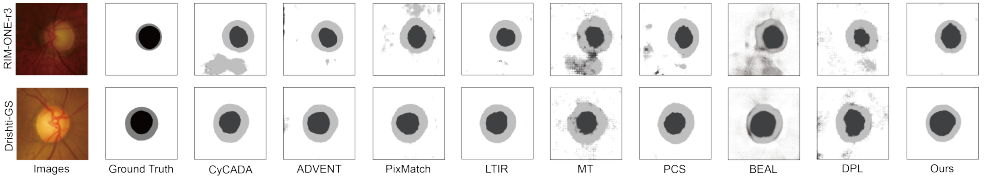

Figure 4: Segmentation results from some of the comparison experiments.

setting. The source-only experiment trains the segmentation network using only source images and directly adapts to the target domain. By contrast, a target-only experiment trains the network using annotated target data under a supervised learning setting. The first implemented module is standard adversarial training with additional discriminator loss. The improvement under two metrics suggests that the concept of adversarial training can be drawn on this task. Then SMSI further boosts the performance by diversifying the source styles and inducing the domain invariant feature generation, providing a large quantity of labeled data for few-shot learning. In addition, a novel class prototype consistency loss allows the network to particularly align the features at the category level. Both Dice and ASD metrics indicate that the proposed method significantly increases the adaptation ability of the segmentation network with limited labeled data.

Settings	Dice Metric [%] ↑			ASD Metric [pixel] ↓		
	Cup	Disc	Average	Cup	Disc	Average
RIM-ONE-r3						
Source only	58.75	64.84	61.79	26.92	43.10	35.01
Adversarial baseline with X_s and X_t	66.95	69.46	68.20	26.65	46.94	36.80
+ SMSI with X_s and $X_{s\to t}$	67.37	77.61	72.49	21.07	27.72	24.40
+ Class-prototype consistency	68.93	82.52	75.72	21.81	19.86	20.84
+ Plain pseudo label learning with X_t	77.03	85.79	81.41	10.13	14.74	12.43
+ CSSL with X_t and $X_{t\to s}$	78.16	88.45	83.30	9.82	11.78	10.80
Target only	80.88	95.57	88.22	10.21	5.20	7.71
Drishti-GS						
Source only	76.55	84.53	80.54	18.77	23.08	20.93
Adversarial baseline with X_s and X_t	76.86	93.72	85.29	16.68	9.68	13.18
+ SMSI with X_s and $X_{s\to t}$	83.45	93.12	88.28	12.37	8.60	10.48
+ Class-prototype consistency	84.76	94.26	89.51	11.02	7.04	9.03
+ Plain pseudo label learning with X_t	84.24	94.88	89.56	10.52	5.93	8.23
+ CSSL with X_t and $X_{t\to s}$	83.64	95.47	89.56	11.04	5.25	8.14
Target only	84.19	97.07	90.63	11.11	3.85	7.48

Table 2: Ablation results with implemented domain adaptation modules in terms of Dice and ASD metrics on RIM-ONE-r3 and Drishti-GS target datasets.

On top of these modules at the first stage, our proposed Cross-Style Self-supervised Learning (CSSL) module brings an improvement of about 14% in dice value and 18 pixels in ASD over the non-adaptation model. We also conduct ablation experiments by conducting the self-supervised learning only on the target images, which introduces less performance gain than the CSSL. This further demonstrates the claim in Section 3.3 that our CSSL can alleviate the noises from the pseudo labels and lead to a better self-supervised segmentation performance. By jointly conducting our proposed strategies, the segmentation performance

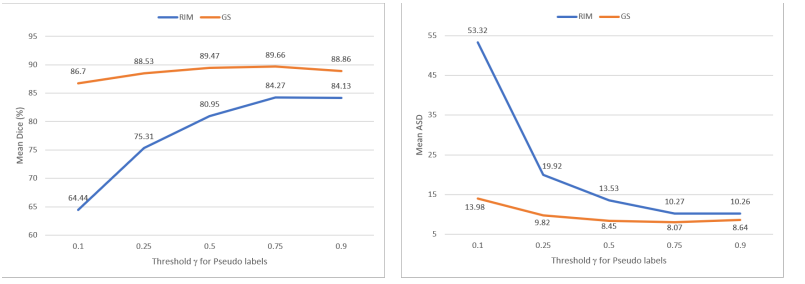

Figure 5: UDA Segmentation performance under different selections of the threshold γ for the pseudo label learning stage. RIM and GS indicate our experiment settings of using the RIM-ONE-r3 and Drishti-GS as the target domain, respectively.

of the source-only model can be improved to a level similar to that of the fully supervised model. In addition, we also explore the models' effectiveness under different thresholds γ for the pseudo label learning stage introduced in Section 3.3. As shown in Figure 5, the best segmentation performance under the Dice and ASD metrics is obtained under both settings when the threshold γ is 0.75.

5 Conclusion

In this work, we propose a novel framework for domain adaptive optic disc and cup segmentation given only a few labeled source data. To alleviate the domain bias issue under the data-scarce setting, the SMSI, CCP, and CSSL modules are designed. In comparison to alternative domain adaptation strategies and even fully supervised networks, the model has been trained to reach competitive outcomes. In this work, we notice the principle bias between the two domains results from the different image styles due to the device variation, and there are no severe distinctions between the morphological structures for the foreground objects in the two domains. As such, future studies are suggested on the cross-domain segmentation problems with larger distinctions in the labeling space.

References

[1] Ahmed Almazroa, Ritambhar Burman, Kaamran Raahemifar, and Vasudevan Lakshminarayanan. Optic disc and optic cup segmentation methodologies for glaucoma image detection: a survey. *Journal of ophthalmology*, 2015.

[2] James Bergstra, Daniel Yamins, and David Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[3] Cheng Chen, Quande Liu, Yueming Jin, Qi Dou, and Pheng-Ann Heng. Source-free domain adaptive fundus image segmentation with denoised pseudo-labeling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021.

[4] Ekin Dogus Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. In *Conference on Computer Vision and Pattern Recognition*, 2019.

[5] Wei Feng, Lin Wang, Lie Ju, Xin Zhao, Xin Wang, Xiaoyu Shi, and Zongyuan Ge. Unsupervised domain adaptive fundus image segmentation with category-level regularization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2022.

[6] Huazhu Fu, Jun Cheng, Yanwu Xu, Damon Wing Kee Wong, Jiang Liu, and Xiaochun Cao. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. *IEEE transactions on medical imaging*, 37(7):1597–1605, 2018.

[7] Francisco Fumero, Jose Sigut, Silvia Alayón, Marta González-Hernández, and Marta González de la Rosa. Interactive tool and database for optic disc and cup segmentation of stereo and monocular retinal fundus images. *International Conference in Central Europe on Computer Graphics*, 2015.

[8] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, 2015.

[9] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[10] Amir Gholami, Shashank Subramanian, Varun Shenoy, Naveen Himthani, Xiangyu Yue, Sicheng Zhao, Peter Jin, George Biros, and Kurt Keutzer. A novel domain adaptation framework for medical image segmentation. In *International MICCAI Brainlesion Workshop*, 2018.

[11] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998, 2018.

[12] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6891–6902, 2021.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[14] Mehran Javanmardi and Tolga Tasdizen. Domain adaptation for biomedical image segmentation using adversarial training. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 554–558, 2018.

[15] Meirui Jiang, Zirui Wang, and Qi Dou. Harmofl: Harmonizing local and global drifts in federated learning on heterogeneous medical images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1087–1095, 2022.

[16] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12975–12984, 2020.

[17] Canran Li, Dongnan Liu, Haoran Li, Zheng Zhang, Guangming Lu, Xiaojun Chang, and Weidong Cai. Domain adaptive nuclei instance segmentation and classification via category-aware feature alignment and pseudo-labelling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 715–724, 2022.

[18] Shaohua Li, Xiuchao Sui, Jie Fu, Huazhu Fu, Xiangde Luo, Yangqin Feng, Xinxing Xu, Yong Liu, Daniel Ting, and Rick Siow Mong Goh. Few-shot domain adaptation with polymorphic transformers. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021.

[19] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. Fast autoaugment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[20] Bingyan Liu, Daru Pan, Zhenbin Shuai, and Hui Song. Ecsd-net: A joint optic disc and cup segmentation and glaucoma classification network based on unsupervised domain adaptation. *Computer Methods and Programs in Biomedicine*, 213:106530, 2022.

[21] Dongnan Liu, Donghao Zhang, Yang Song, Fan Zhang, Lauren O'Donnell, Heng Huang, Mei Chen, and Weidong Cai. Unsupervised instance segmentation in microscopy images via panoptic domain adaptation and task re-weighting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4243–4252, 2020.

[22] Dongnan Liu, Donghao Zhang, Yang Song, Fan Zhang, Lauren O'Donnell, Heng Huang, Mei Chen, and Weidong Cai. Pdam: A panoptic-level feature alignment framework for unsupervised domain adaptive instance segmentation in microscopy images. *IEEE Transactions on Medical Imaging*, 40(1):154–165, 2020.

[23] Peng Liu, Charlie T Tran, Bin Kong, and Ruogu Fang. Cada: Multi-scale collaborative adversarial domain adaptation for unsupervised optic disc and cup segmentation. *Neurocomputing*, 469:209–220, 2022.

[24] Ali Madani, Mehdi Moradi, Alexandros Karargyris, and Tanveer Syeda-Mahmood. Semi-supervised learning with generative adversarial networks for chest x-ray classification with ability of data domain adaptation. In *IEEE 15th International symposium on biomedical imaging*, pages 1038–1042, 2018.

[25] Luke Melas-Kyriazi and Arjun K Manrai. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12435–12445, 2021.

[26] Yanda Meng, Hongrun Zhang, Dongxu Gao, Yitian Zhao, Xiaoyun Yang, Xuesheng Qian, Xiaowei Huang, and Yalin Zheng. Bi-gcn: boundary-aware input-dependent graph convolution network for biomedical image segmentation. *British Machine Vision Conference*, 2021.

[27] José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel van Keer, Deepti R Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis*, 59: 101570, 2020.

[28] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015.

[29] Cheng Peng, Andriy Myronenko, Ali Hatamizadeh, Vishwesh Nath, Md Mahfuzur Rahman Siddiquee, Yufan He, Daguang Xu, Rama Chellappa, and Dong Yang. Hypersegnas: Bridging one-shot neural architecture search with 3d medical image segmentation using hypernet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20741–20751, 2022.

[30] Artem Sevastopolsky. Optic disc and cup segmentation methods for glaucoma detection with modification of u-net convolutional neural network. *Pattern Recognition and Image Analysis*, 27(3):618–624, 2017.

[31] Jayanthi Sivaswamy, SR Krishnadas, Gopal Datt Joshi, Madhulika Jain, and A Ujjwaft Syed Tabish. Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation. In *2014 IEEE 11th international symposium on biomedical imaging (ISBI)*, pages 53–56, 2014.

[32] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[33] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019.

[34] Shujun Wang, Lequan Yu, Kang Li, Xin Yang, Chi-Wing Fu, and Pheng-Ann Heng. Boundary and entropy-driven adversarial learning for fundus image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 102–110, 2019.

[35] Shujun Wang, Lequan Yu, Xin Yang, Chi-Wing Fu, and Pheng-Ann Heng. Patch-based output space adversarial learning for joint optic disc and cup segmentation. *IEEE transactions on medical imaging*, 38(11):2485–2495, 2019.

[36] Xinyi Wang, Tiange Xiang, Chaoyi Zhang, Yang Song, Dongnan Liu, Heng Huang, and Weidong Cai. Bix-nas: Searching efficient bi-directional architecture for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 229–238, 2021.

[37] Tiange Xiang, Chaoyi Zhang, Xinyi Wang, Yang Song, Dongnan Liu, Heng Huang, and Weidong Cai. Towards bi-directional skip connections in encoder-decoder architectures and beyond. *Medical Image Analysis*, 78:102420, 2022.

[38] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020.

[39] Xiangyu Yue, Zangwei Zheng, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, and Alberto Sangiovanni Vincentelli. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13834–13844, 2021.

[40] Jingyi Zhang, Jiaxing Huang, Zichen Tian, and Shijian Lu. Spectral unsupervised domain adaptation for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9829–9840, 2022.

[41] Youshan Zhang and Brian D Davison. Deep least squares alignment for unsupervised domain adaptation. *British Machine Vision Conference*, 2021.

[42] Yuhao Zhang, Xiangrui Cai, Ying Zhang, Hong Kang, Xin Ji, and Xiaojie Yuan. Tau: Transferable attention u-net for optic disc and cup segmentation. *Knowledge-Based Systems*, 213:106668, 2021.

[43] Ziyuan Zhao, Kaixin Xu, Shumeng Li, Zeng Zeng, and Cuntai Guan. Mt-uda: Towards unsupervised cross-modality medical image segmentation with limited source labels. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 293–303, 2021.

[44] Yangtao Zheng, Di Huang, Songtao Liu, and Yunhong Wang. Cross-domain object detection through coarse-to-fine feature adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13766–13775, 2020.

[45] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.