

# CADA: Multi-scale Collaborative Adversarial Domain Adaptation for Unsupervised Optic Disc and Cup Segmentation

Peng Liu<sup>a</sup>, Charlie T. Tran<sup>b</sup>, Bin Kong<sup>d</sup>, Ruogu Fang<sup>a,b,c,\*</sup>

<sup>a</sup>*J. Crayton Pruitt Family Dept. of Biomedical Engineering, University of Florida, Gainesville, FL 32611, USA*

<sup>b</sup>*Dept. of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA*

<sup>c</sup>*Center for Cognitive Aging and Memory, University of Florida, Gainesville, FL 32611, USA*

<sup>d</sup>*Department of Computer Science, UNC Charlotte, Charlotte, NC 28223, USA*

---

## Abstract

Recently, deep neural networks have demonstrated comparable and even better performance than board-certified ophthalmologists in well-annotated datasets. However, the diversity of retinal imaging devices poses a significant challenge: domain shift, which leads to performance degradation when applying the deep learning models trained on one domain to new testing domains. In this paper, we propose a multi-scale input along with multiple domain adaptors applied hierarchically in both feature and output spaces. The proposed training strategy and novel unsupervised domain adaptation framework, called Collaborative Adversarial Domain Adaptation (CADA), can effectively overcome the challenge. Multi-scale inputs can reduce the information loss due to the pooling layers used in the network for feature extraction, while our proposed CADA is an interactive paradigm that presents an exquisite collaborative adaptation through both adversarial learning and ensembling weights at different network layers. In particular, to produce a better prediction for the unlabeled target domain data, we simultaneously achieve domain invariance and model generalizability via adversarial learning at multi-scale outputs from different levels

---

\*Corresponding author

Email address: [ruogu.fang@bme.ufl.edu](mailto:ruogu.fang@bme.ufl.edu) (Ruogu Fang)

of network layers and maintaining an exponential moving average (EMA) of the historical weights during training. Without annotating any sample from the target domain, multiple adversarial losses in encoder and decoder layers guide the extraction of domain-invariant features to confuse the domain classifier. Meanwhile, the ensembling of weights via EMA reduces the uncertainty of adapting multiple discriminator learning. Comprehensive experimental results demonstrate that our CADA model incorporating multi-scale input training can overcome performance degradation and outperform state-of-the-art domain adaptation methods in segmenting retinal optic disc and cup from fundus images stemming from the REFUGE, Drishti-GS, and Rim-One-r3 datasets. *The code will be available at <https://github.com/cswin/CADA>.*

**Keywords:** Domain adaptation, Unsupervised learning, Segmentation, Retinal fundus images

---

## 1. Introduction

Early diagnosis is vital for the treatment of various vision degradation diseases [1], such as glaucoma, Diabetic Retinopathy (DR), and age-related macular degeneration. Many eye diseases can be revealed by the morphology of Optic Disc (OD) and Optic Cup (OC) [2]. For instance, glaucoma is usually characterized by a large Cup to Disc Ratio (CDR), the ratio of the vertical diameter of the cup to the vertical diameter of the disc. Currently, determining CDR is mainly performed by pathology specialists. However, it is extremely expensive to accurately calculate CDR by human experts. Furthermore, manual delineation of these lesions also introduces subjectivity, intra- and inter-variability [3]. Therefore, it is essential to automate the process of calculating CDR. OD and OC segmentation are commonly adopted to automatically calculate the CDR. Nevertheless, OD segmentation is challenging because pathological lesions usually occur on OD boundaries, which affect the accurate identification of the OD region. Accurate OC segmentation is more challenging due to the region overlap between the cup and the blood vessels [4].

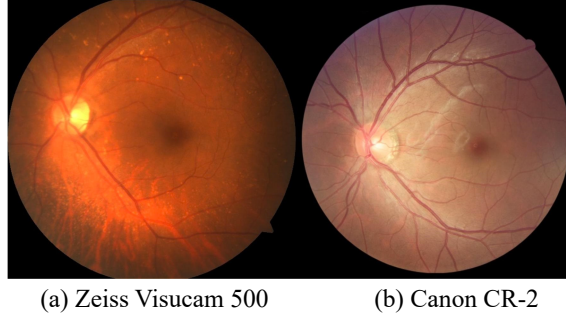


Figure 1: Retinal fundus images collected by different fundus cameras revealing a variation in color and intensity.

Recently, deep learning-based methods [5, 6, 7, 8] have been proposed to overcome these challenges and some of them, e.g., M-Net [5], have demonstrated impressive results. Although these methods tend to perform well when being applied to well-annotated datasets, the segmentation performance of a trained network may degrade severely on datasets with different distributions, particularly for retinal fundus images captured with different imaging devices (e.g., different cameras, as illustrated in Fig. 1). The variance among the diverse data domains limits deep learning’s deployment in reality and impedes us from building a robust application for retinal fundus image parsing. To recover the degraded performance, annotating the fundus images captured from every new domain and then re-training or fine-tuning a model is an easy way but extremely expensive and even impractical for the medical areas that require expertise.

To tackle this challenge, recent studies [9, 10, 11, 12, 13, 14, 15] have demonstrated the effectiveness of using deep learning for unsupervised domain adaptation to enhance the performance of applying models on unlabeled target domain data. Existing works have mainly focused on minimizing the distance between the source and target domains to align the latent feature distributions from different domains [12]. Several primary approaches can guide the alignment process, which includes image-to-image translation of the input images [16], adversarial training for the intermediate representations in the layers of the model

(encoder or decoder) [17], and applying adversarial learning to the output of the model [11]. However, adversarial discriminative learning usually suffers from the instability of its training. Numerous methods have been studied to  
40 tackle this challenge. Self-ensembling [18] is one of them recently applied to visual domain adaptation [19]. In particular, gradient descent is used to train a student network, and the exponential moving average of the weights of the student is transferred to a teacher network after applying each training sample. The mean square difference between the outputs of the student and the teacher  
45 is used as an unsupervised loss to train the student network. The paradigm of student-teacher has been a widely used strategy for unsupervised training of a deep neural network, feature extraction [20], and knowledge distillation [21]. This unsupervised training strategy allows the student network to capture more information about the data during training and achieve a better prediction.

50 Furthermore, most of the existing methods have not considered providing multi-scale information of the data to the deep neural networks for having a better understanding of the difference between target and source domain features. In addition, most deep neural networks employ pooling layers to reduce model parameters and extract important features. However, pooling layers lead  
55 to a significant loss of the information in the original input data. In addition, the optic cup and disc in fundus image have high variance in brightness, color, shape, and orientation, which makes single-scale and single-level adversarial adaptation insufficient. Capturing only information from the output space neglects the intuition that low-level features are similar across various domains,  
60 leading to poor domain adaptability and missed opportunities for incorporating hierarchical features in segmentation predictions. To overcome these problems, in this study, we propose a multi-scale input training strategy to integrate different scales of features into different levels of the network layers. On one hand, multi-scale inputs performing at different levels of network layers can reduce the  
65 information loss due to the pooling layers in the network; On the other hand, it can provide rich information for the network to easily distinguish the difference between the source and target domain features.

These multi-scale inputs are integrated into an encoder-decoder structure to form multiple sub-networks with multi-outputs. In this way, we can have multiple discriminators between the source and target domain input and leverage ensembling internally on the multiple distinctions at different levels of the network layers to have a better final segmentation prediction in target domain data. This proposed multi-domain adaptor approach can overcome the limitation of the current methods while comprehensively applying domain adaptation at hierarchical multi-scale in both feature and output space.

In this paper, we propose a novel unsupervised domain adaptation framework called Collaborative Adversarial Domain Adaptation (CADA) to further the state-of-the-art in overcoming the underlining domain shift problem. In particular, we take advantage of self-ensembling to stabilize the adversarial discriminative learning of the latent representations from domain shifting to prevent the network from getting stuck in degenerate solutions. Most importantly, we apply the unsupervised loss by adversarial learning not only to the output space but also to the input space and the intermediate representations of the network. Thus, from a complementary perspective, adversarial learning can consistently provide various model space and time-dependent weights to self-ensembling to accelerate the learning of the domain-invariant features and further enhance the stabilization of adversarial learning, forming a benign collaborative circulation and unified framework. The significant contributions of this paper are as follows:

- We propose CADA, a novel unsupervised domain adaptation framework, which exploits collaborative adversarial learning and weights self-ensembling for feature adaptation to tackle domain shift in a mutually beneficial and complementary manner at different network layers, resulting in a robust and accurate model.
- We propose a multi-scale input training strategy to overcome the information loss when applying pooling layers in the network and offer an opportunity to integrate various scales of low-level and high-level features for improved network learning.

- We optimized feature adaptation by applying adversarial discriminative learning in two phases of the network, i.e., intermediate representation space and output space. More specifically, we apply adversarial learning at multiple layers of the network to learn the invariant features in both encoder and decoder phases simultaneously.
- We reduce the uncertainty of multiple discriminators collaborative learning for domain shift via the EMA to ensemble model weights dynamically during training.
- We evaluate the effectiveness of our CADA on the challenging task of unsupervised joint segmentation of the retinal OD and OC. Our CADA can overcome performance degradation to domain shift and outperform one of the state-of-the-art domain adaptation methods with a significant performance gain on various datasets.

This work is a substantial extension of our conference paper “CFEA: Collaborative Feature Ensembling Adaptation for Domain Adaptation in Unsupervised Optic Disc and Cup Segmentation” [22] published in Medical Image Computing and Computer Assisted Intervention (MICCAI) 2019. In this extension, we substantially expanded our framework’s reliability and scalability of overcoming domain shift issue in fundus images. In particular, we demonstrate the significant new contributions as below:

- We propose a novel multi-scale input layer to enhance the feature interaction between the encoder and the decoder where CFEA only uses a single scale. An input on each scale provides original image information to an encoder layer, which is followed by a decoder layer at the same network “pyramid” level. The rich original pixel-wise feature can infuse the interaction between encoder and decoder at the different feature-learning levels in the network. This infusion triggered by the multi-scale input can further guide the model learning and promote performance by reducing the significant information loss due to the pooling layers applied in the

network and reducing the high variance of the optical cup and disc images in brightness, color, shape, and orientation.

- Instead of a single domain adaptor (e.g., a discriminator network) at the end of the network in CFEA, we propose to apply multi-domain adaptors at hierarchical multi-scales in both feature and output space, which encourages the network to learn the domain-invariant features consistently. More importantly, they can collaboratively distinguish robust latent features in the scenarios that the optical cup and disc images have high variance in brightness, color, shape, and orientation, thus leading to a reliable and scalable domain adaptation framework.
- Comprehensive ablation studies are performed to investigate the effectiveness of the proposed framework. The ablation study investigates the importance of the encoder adversarial discriminative adaptation, the power of weights self-ensembling adaptation, the scalability of using multiple domain adaptors, and the choice of various combinations of the weights of loss functions.
- Evaluation on multiple public datasets is performed to show generalizability and stability of the proposed CADA framework compared to state-of-the-art methods.

## 2. Related Work

### 2.1. Optic Disc and Optic Cup Features

The features of the Optic Disc (OD) and Optic Cup (OC) are critical for the diagnosis of eye diseases [2]. For example, ophthalmic pathologies (e.g., glaucoma) can be indicated by the variations of the shape, color, or depth of OD. The Cup to Disc Ratio (CDR), the ratio of the vertical diameter of the cup to the vertical diameter of the disc, is considered a valuable feature for diagnosing eye diseases [23], such as glaucoma, because higher CDR is highly associated with detectable visual field damage [24]. Currently, determining CDR is mainly

155 performed by pathology specialists. However, it is expensive to calculate CDR  
by human experts accurately. Furthermore, the variance of determining the  
CDR among professionals is usually significant, which can be caused by both  
the diversity of retinal fundus images and different experiences of the profession-  
als [25]. Therefore, it is essential to automate the process of calculating CDR.  
160 On one hand, this automated process can reduce the cost of diagnosis. On the  
other hand, it can stabilize diagnostic accuracy and improve the efficiency of  
retinopathy screening procedures.

## 2.2. OD and OC Image Segmentation

Image segmentation is a long-term research topic in the field of computer  
165 vision and image analysis. It is the basis for feature recognition and quantitative  
feature understanding [26]. In medical imaging, image segmentation is partic-  
ularly important since it can help to locate related lesions/tumors and provide  
quantitatively analytical results of shapes/morphologies for clinicians. For ex-  
ample, image segmentation can automatically detect the OD and OC regions  
170 and calculate the CDR simultaneously, e.g., [4]. The task of OD segmentation is  
to detect the region between retinal and the rim. The presence of pathological  
lesions on the OD boundaries become problematic for OD detection. More to  
the point, OC detection is hindered by the region overlap between the cup and  
the blood vessels, as well as the color intensity changing between the cup and  
175 rim. It is critical to erase these challenges for reducing incorrect OD and OC  
segmentation that may cause a false diagnosis.

Recently, many deep learning-based studies [2, 4, 5] have been proposed  
to overcome these challenges. In general, there are several steps to achieve  
a decent result. Firstly, a pre-trained disc center localization method [27] is  
180 used to detect the OD and OC. The localization mainly acts as an attention  
mechanism so that the network can focus on essential regions and meanwhile, the  
polar transformation amplifies the relevant features to enable a more accessible  
learning process. Secondly, the localized areas are transformed (e.g., cropped,  
re-sized, and image coordinate system consistency) into the segmentation model



185 training stage. Lastly, these transformed image regions are fed into an encoder-decoder convolutional network to predict the actual OD and OC regions for an arbitrary fundus image. The encoder is utilized to extract rich image features; the decoder part is used to produce accurate segmentation results based on the encoded features. These combined techniques can reduce the negative effect  
190 on model performance caused by the variance in retinal images. However, the variation is only constrained within one image domain, in which the training and testing images usually have similar distributions, such as background color and intensity. In practice, the testing images can be acquired from different types of cameras and have varying background or image intensity (as illustrated  
195 in Fig. 1). The performance of a model trained on the dataset collected from one domain is severely degraded in another domain. This issue is referred to as “domain shift” [28]. It is critical to overcome this issue for a generalized and robust model in medical practice.

### 2.3. Unsupervised Domain Adaptation

200 Saenk et al. [29] originally introduced the unsupervised domain adaptation problem in tackling the performance degradation caused by the domain shift. In particular, unsupervised domain adaptation aims to tackle domain shift via adapting the training process of a model in an unsupervised manner. The model is adapted to improve the performance on the target domain. More  
205 importantly, leveraging unsupervised learning can reduce the tremendous and expensive data labeling work for the target domain. Therefore, unsupervised domain adaptation is a promising direction to solve domain shift problems, especially, in the medical field whose data is multi-modal and requires expensive and expertise data labeling.

210 Recently, many deep learning-based domain adaptation methods [4, 17, 30] have been proposed and achieved several encouraging results. Many of these methods tackle the domain shift issue by extracting invariant features across the source and target domains. A critical approach for reducing the domain discrepancy is adversarial learning [9], which has become a fundamental method

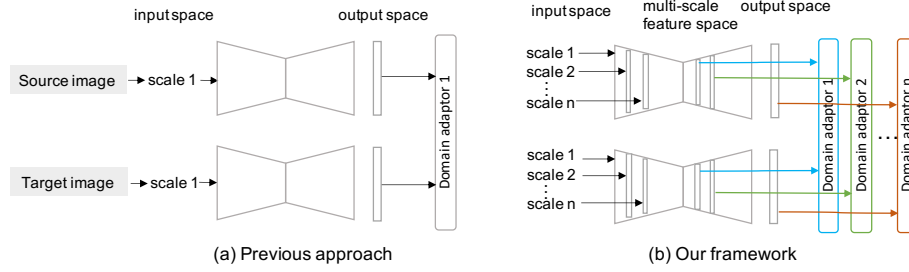


Figure 2: Comparison between prior domain adaptation methods and our framework. Each colored arrow represents a connection with each multi-scale output to a corresponding domain adaptor.

to obtain invariant information across multiple domains. In particular, it lever-  
ages the gradient discrepancy between learning the labeled and unlabeled data  
to minimize performance degradation. The implementation can either be image-  
to-image translation [31, 32, 33] in a convolutional neural network (CNN) input-  
end or multiple adversarial learning [11, 34] applied at the output-end of a CNN.  
Noticeably, the image-to-image translation usually introduces artifacts, which  
may be not a proper approach in the medical field. Therefore, in this work, we  
focus on gradient-based adversarial learning.

In addition, previous adversarial learning based approaches [35] are mainly  
applied domain adaptation at the output space of a deep neural network. How-  
ever, accurate image segmentation requires the model to capture both low-  
and high-level image representations; thus, it would be ideal for applying do-  
main adaptation at various feature spaces with multi-scale representations. In  
this study, we comprehensively investigated whether employing multiple domain  
adaptors at a different level of layers of the network can benefit the represen-  
tation learning across domains compared to the approaches only focusing on  
adaptation at the output space. We showed the design difference between pre-  
vious methods and our framework and illustrated the major novelty in Fig. 2.

Furthermore, although adversarial learning can align the latent feature dis-

tribution of the source and target domain and have achieved encouraging results [9], the results of multiple adversarial learning-based methods are easily suffering from sub-optimal performance due to the difficulty of stabilizing the training process of multiple adversarial modules. Thus, in this work, we propose to leverage the Exponential Moving Average (EMA) [36] computing method to dynamically ensemble learning weights as embedding multiple adversarial modules in a network. Meanwhile, this stabilization can bring not only a more robust but also an accurate model to effectively overcome the domain shift issue in fundus image segmentation problems.

### 3. Multi-scale Collaborative Feature Ensembling Adaptation

#### 3.1. Problem Formulation

Unsupervised domain adaptation typically refers to the scenario: given a labeled source domain dataset with distribution  $P(X_s)$  and the corresponding label  $Y_s$  with distribution  $P(Y_s|X_s)$ , as well as a target dataset with distribution  $P(X_t)$  and unknown label with distribution  $P(Y_t|X_t)$ , where  $P(X_s) \neq P(X_t)$ , the goal is to train a model from both labeled data  $X_s$  and unlabeled data  $X_t$ , with which the expected model distribution  $P(\hat{Y}_t|X_t)$  is close to  $P(Y_t|X_t)$ .

#### 3.2. Overview of the Proposed Method

Fig. 3 illustrates the design of the proposed domain adaptation framework. Our network includes three primary networks, i.e. the Source-domain Network (SN, in blue), the Target-domain Student Network (TSN, in gray), and the Target-domain Teacher Network (TTN, in orange). We utilize multi-scale inputs and outputs to each of the primary networks to adapt various levels of features hierarchically. During training, at each iteration, the source images are fed into SN to generate the Source-encoder Feature (SF)  $P_{sf}$  and source decoder output  $P_{so}$ . The source domain segmentation loss is obtained by comparing the  $P_{so}$  with the source domain ground truth. TSN shares the same weights with SN, and the weights of TTN are the Exponential Moving Average (EMA) weights of

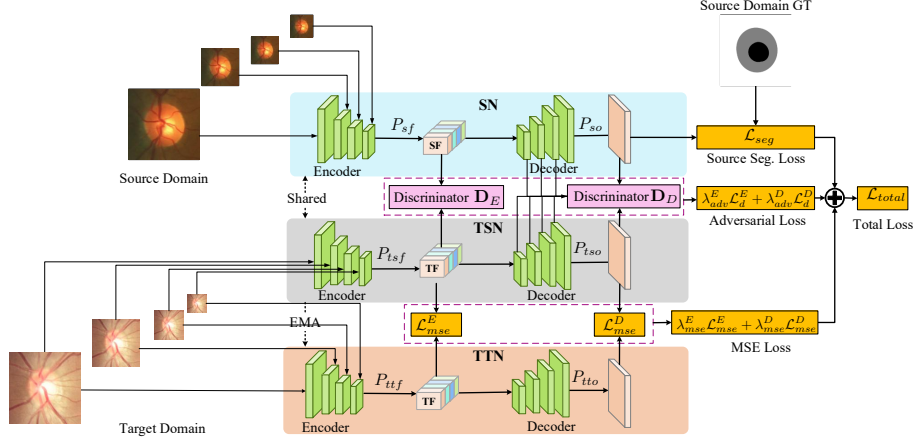


Figure 3: Overview of the proposed model architecture with each primary network block colored: the Source-Domain Network (SN, in blue), the Target-domain Student Network (TSN, in gray) and the Target-domain Teacher Network (TTN, in orange). Note that discriminators can be added between all the intermediate decoder layers of SN and TSN. However, we only add the discriminators among the input ( $P_{sf}$  and  $P_{tsf}$ ) and output ( $P_{so}$  and  $P_{tso}$ ) of the decoders in this figure for simplicity.

TSN. Adversarial losses for domain confusion are added for both encoder and decoder outputs of SN and TSN. Moreover, Mean Square Error (MSE) losses are added to both the encoder and decoder outputs of TSN and TTN. To reduce the difficulty of high-dimensional feature calculations, the outputs of all encoders

Although each of the networks plays a distinctive role in guiding networks to learn domain invariant representations, all of them can interact with each other, benefit one another, and work collaboratively as a unified framework during an end-to-end training process. SN and TSN focus on supervised learning for labeled samples from the source domain ( $X_s$ ) and adversarial discriminative learning for unlabeled samples from the target domain ( $X_t$ ), separately. More importantly, we allow SN and TSN to share the weights that are sequentially learned from both labeled and unlabeled samples. This technique is commonly adopted in unsupervised domain adaptation to reduce the number of learnable

parameters [11, 37]. The labeled samples enable the network to learn accurate segmentation predictions while the unlabeled samples bring unsupervised learning and further present a type of perturbation to regularize the model training [38]. Furthermore, TTN conducts the weight self-ensembling [19] part by replicating the average weights of the TSN instead of predictions. TTN solely takes unlabeled target images as input and then the mean square difference between TSN and TTN is computed for the same target sample. Different data augmentations (e.g., adding Gaussian noise and random intensity or brightness scaling) are applied to TSN and TTN to avoid the loss vanishing issue.

In this work, the U-Net [39] with encoder-decoder structure is employed as the backbone of each network, since U-Net is one of the most successful segmentation frameworks in medical image segmentation. With the adaptability and flexibility of our framework, we expect that the results can generalize to other backbone networks and medical image analysis tasks.

### 3.3. Multi-scale Input Sub-networks

We allow multi-scale inputs to provide rich original pixel-wise features that can infuse the interaction between encoder and decoder at the different feature-learning levels of the network. Each level of the network is considered as a sub-network that primarily processes one scale of the original input for segmentation prediction. Fig. 4 shows the paradigm of the multi-scale input networks. To simplify the re-scaling procedure, we apply a  $2 \times 2$ ,  $4 \times 4$ , and  $8 \times 8$  size of pooling followed by a  $3 \times 3$  convolutional layer to reduce the scale of the original input. A level scale of an input passes through several convolutional layers across both encoder and decoder of the network. Four scales of inputs are built and each of them is corresponding to a level of the neural network layers, which is considered a sub-network. The output of each sub-network uses a domain adaptor for domain adaptation. Lastly, another domain adaptor is employed as the average of the four sub-networks outputs for domain adaptation. The whole network uses multiple domain adaptors to adjust itself for domain adaptation learning at both low-level and high-level features and perform an accurate

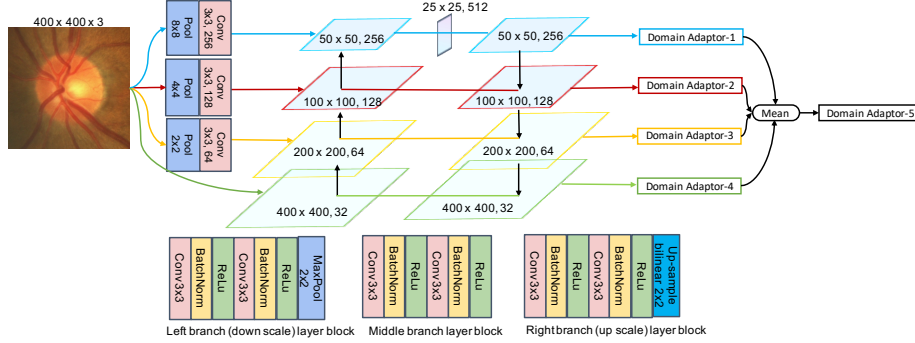


Figure 4: The network design details illustrating the multi-scale input in the combination of multiple domain adaptors applied at different hierarchical layers. Each color represents a level scale of the input that passes through several convolutional layers across both encoder and decoder of the network. The label on each layer includes the size of the input to the layer-block and the number of channels in all convolutional layers in that block.

segmentation prediction, especially those from the fundus optic cup and disc. There are 5 adaptors and each of them is a combination of supervised segmentation learning, the mean square difference between the prediction of the teacher and student network, and adversarial learning through deep convolutional networks as discriminators. We show the paradigm of the multi-scale input in the combination of multiple domain adaptors applied at different hierarchical layers and the details of network design in Fig. 4. Notably, we include the multi-scale inputs to both the student and teacher networks of our CADA framework.

### 3.4. Multiple Adversarial Discriminative Learning

We apply five discriminators at the encoder and decoder of the networks, separately, to achieve adversarial discriminative learning. To simplify our method's explanation, the following section is using two types of discriminators for discussion: one discriminator is applied for encoder features and the other four discriminators are used for the outputs of the decoder. The adversarial loss functions are calculated between SN and TSN. Each of the loss calculations is performed by two steps in each training iteration: (1) train the target domain

segmentation network to maximize the adversarial loss  $\mathcal{L}_{adv}$ , thereby fooling the domain discriminator  $\mathbf{D}$  to maximize the probability of the source domain feature  $P_s$  being classified as target features:

$$\mathcal{L}_{adv}(X_s) = \mathbb{E}_{x_s \sim X_s} \log(1 - \mathbf{D}(P_s)), \quad (1)$$

325 and 2) minimize the discriminator loss  $\mathcal{L}_d$ :

$$\mathcal{L}_d(X_s, X_t) = \mathbb{E}_{x_t \sim X_t} \log(\mathbf{D}(P_t)) + \mathbb{E}_{x_s \sim X_s} \log(1 - \mathbf{D}(P_s)), \quad (2)$$

where  $P_t$  is the target domain feature.

Note that discriminators can be added between all the intermediate decoder layers of SN and TSN. However, we only add the discriminators among the input ( $P_{sf}$  and  $P_{tsf}$ ) and output ( $P_{so}$  and  $P_{tso}$ ) of the decoders in this section  
330 for simplicity.

### 3.5. Self-ensembling

In self-ensembling for domain adaptation, the training of the student model is iteratively improved by the task-specific loss upon an Exponential Moving Average (EMA) model (teacher) of the student model, which can be illustrated  
335 as:

$$\Phi'_t = \alpha \Phi'_{t-1} + (1 - \alpha) \Phi_t \quad (3)$$

where  $\Phi_t$  and  $\Phi'_t$  denote the parameters of the student network and the teacher network, respectively. EMA transfers a smooth version of the weights of the student to the teacher network. Thus, the teacher network is more stable and robust than the student.

340 More specifically, at each iteration, a mini-batch of the labeled source domain and unlabeled target samples are drawn from the target domain  $T$ . Then, the EMA predictions and the base predictions are generated by the teacher model and the student model respectively with different augmentations applied to the target samples. Afterward, a Mean-Squared Error (MSE) loss between the EMA  
345 and target predictions is calculated. Finally, the MSE loss together with the task-specific loss on the labeled source domain data is minimized to update the

parameters of the student network. Since the teacher model is an improved model at each iteration, the MSE loss helps the student model to learn from the unlabeled target domain images. Therefore, the student model and teacher  
 350 model can work collaboratively to achieve robust and accurate predictions.

### 3.6. Unsupervised Domain Adaptation

Unlike existing methods, our method appropriately integrates adversarial domain confusion and self-ensembling with an encoder-decoder architecture.

#### 3.6.1. Adversarial Feature Adaptation

Adversarial domain confusion is applied to both the encoded features and decoded predictions between Source domain Network (SN) and Target domain Student Network (TSN) to reduce the distribution differences. According to Eq. 1 and 2, this corresponds to the adversarial loss function  $\mathcal{L}_{adv}^E$  for the encoder output of SN and TSN, and the adversarial loss function  $\mathcal{L}_{adv}^D$  for the decoder output of SN and TSN:

$$\mathcal{L}_{adv}^E(X_s) = \mathbb{E}_{x_s \sim X_s} \log(1 - \mathbf{D}_E(P_{sf})), \quad (4)$$

$$\mathcal{L}_{adv}^D(X_s) = \mathbb{E}_{x_s \sim X_s} \log(1 - \mathbf{D}_D(P_{so})), \quad (5)$$

355 where  $P_{sf} \in \mathbb{R}^{W_e \times H_e \times C_e}$  is the encoder output and  $P_{so} \in \mathbb{R}^{W_d \times H_d \times C_d}$  is the decoder output.  $H_d$  and  $W_d$  are the width and height of the decoders' output;  $C_d$  refers to pixel categories of the segmentation result, which is three in our cases.  $H_e$ ,  $W_e$ , and  $C_e$  are the width, height, channel of the encoders' output.  $\mathbf{D}_E$  and  $\mathbf{D}_D$  are the discriminator networks for the encoder and decoder outputs,  
 360 respectively.

The discriminator loss  $\mathcal{L}_d^E$  for the encoder feature and the discriminator loss  $\mathcal{L}_d^D$  for the decoder feature are as follows:

$$\begin{aligned} \mathcal{L}_d^E(X_s, X_t) = & \mathbb{E}_{x_t \sim X_t} \log(\mathbf{D}_E(P_{tsf})) + \\ & \mathbb{E}_{x_s \sim X_s} \log(1 - \mathbf{D}_E(P_{sf})) \end{aligned} \quad (6)$$



$$\mathcal{L}_d^D(X_s, X_t) = \mathbb{E}_{x_t \sim X_t} \log(\mathbf{D}_D(P_{tso})) + \mathbb{E}_{x_s \sim X_s} \log(1 - \mathbf{D}_D(P_{so})) \quad (7)$$

where  $P_{tsf} \in \mathbb{R}^{W_e \times H_e \times C_e}$  is the encoder output and  $P_{tso} \in \mathbb{R}^{W_d \times H_d \times C_d}$  is the decoder output of TSN.

### 3.6.2. Collaborative Adaptation with Self-ensembling

Self-ensembling is also applied to both the encoded features and decoded predictions between TSN and Target domain Teacher Network (TTN). In this work, the MSE is used for self-ensembling. The MSE loss  $\mathcal{L}_{mse}^E$  between encoder outputs of TSN and TTN, and the MSE loss  $\mathcal{L}_{mse}^D$  between decoder outputs of TSN and TTN can be formulated as:

$$\mathcal{L}_{mse}^E(X_t) = \mathbb{E}_{x_t \sim X_t} \left[ \frac{1}{M} \sum_{i=1}^M (p_i^{tsf} - p_i^{ttf})^2 \right], \quad (8)$$

$$\mathcal{L}_{mse}^D(X_t) = \mathbb{E}_{x_t \sim X_t} \left[ \frac{1}{N} \sum_{i=1}^N (p_i^{tso} - p_i^{tto})^2 \right]. \quad (9)$$

where  $p_i^{tsf}$ ,  $p_i^{ttf}$ ,  $p_i^{tso}$ , and  $p_i^{tto}$  denote the  $i^{th}$  element of the flattened predictions ( $P_{tsf}$ ,  $P_{ttf}$ ,  $P_{tso}$ , and  $P_{tto}$ ) of the student encoder, student decoder, teacher encoder, teacher decoder, respectively.  $M$  and  $N$  are the number of elements in the encoder feature and decoder output, respectively.

### 3.6.3. Patch-based Discriminator Learning in Multi-scale Output Space

Unlike the domain adaption for classification problems, we need to adapt both low-level and high-level features for pixel-wise image segmentation tasks. To have a better segmentation on the image across the source domain, the invariant features from both low-level and high-level layers are considered in this study. Particularly, like the study [35], the geometry structure of the predicted segmentation masks in the output space is used for domain adaptation. However, the dimension of the feature space from different levels of the layers varies from 256, 128, 64, to 32. Thus, we perform a 2D convolutional layer with  $1 \times 1$  kernels on each scale of features (see Fig. 4) to reduce the feature dimensions

consistently to be the number of pixel classes in the segmentation mask, which is 3 classes, including background, optic disc, and optic cup. In other words, we convert the features in high-dimensional feature space to a low-dimensional output space. We achieve domain adaptation at each level of layers of the network by applying adversarial learning on the converted low-dimensional features in the output space. We apply a patch-based discriminator on each output of the segmentation network. In the adversarial learning, the segmentation network is to fool each discriminator by producing a  $m \times n \times 3$  size of output having a similar distribution either from source or target domain, where  $m$  and  $n$  is the width and height of the output, respectively. A patch-based discriminator [32] is used to perform adversarial learning for capturing the local statistical similarity. Basically, this type of discriminator tries to classify whether each patch in a predicted mask image is following the distribution of that from the source or target domain. More particularly, each discriminator network is composed of five convolutional layers. They have 64, 128, 256, and 512 channels, respectively. A kernel size  $4 \times 4$  and a stride size  $2 \times 2$  are implemented in each layer. A Leaky ReLU [40] layer with an alpha value of 0.2 is used after each convolutional layer. Each discriminator produces a  $16 \times 16$  size output.

#### 3.6.4. Total Objective Function

Finally, we use cross-entropy as the segmentation loss for labeled images from the source domain. Combing Eq. 4, 5, 6, 7, 8, and 9, the total loss is obtained, which can be formulated as below.

$$\begin{aligned} \mathcal{L}_{total}(X_s, X_t) = & \mathcal{L}_{seg}(X_s) + \lambda_{adv}^E \mathcal{L}_d^E(X_s, X_t) + \\ & \lambda_{adv}^D \mathcal{L}_d^D(X_s, X_t) + \lambda_{mse}^E \mathcal{L}_{mse}^E(X_t) + \lambda_{mse}^D \mathcal{L}_{mse}^D(X_t), \end{aligned} \quad (10)$$

where  $\lambda_{adv}^E$ ,  $\lambda_{adv}^D$ ,  $\lambda_{mse}^E$ , and  $\lambda_{mse}^D$  balance the weights of the losses. The choice of each weight component  $\lambda$  was accomplished by cross-validation in our experiments.  $\mathcal{L}_{seg}(X_s)$  is the segmentation loss. Based on Eq. 10, we optimize the following min-max problem:

$$\min_{f_\phi, f_{\bar{\phi}}} \max_{\mathbf{D}_E, \mathbf{D}_D} \mathcal{L}_{total}(X_s, X_t), \quad (11)$$

where  $f_{\tilde{\phi}}$  and  $f_{\phi}$  are the source domain network with trainable weights  $\tilde{\phi}$  and target domain network with trainable weights  $\phi$ .

## 4. Experiments and Results

### 4.1. Dataset

Extensive experiments are conducted on three public datasets, including REFUGE [41], Drishti-GS [42], and RIM-ONE-r3 [43], to validate the effectiveness of the proposed method. The dataset REFUGE includes 400 source domain retinal fundus images (supervised training dataset) with size  $2124 \times 2056$ ,  
 415 acquired by a Zeiss Visucam 500 camera, 400 labeled (testing dataset), and 400 additional unlabeled (unsupervised training dataset) target domain retinal fundus images with size  $1634 \times 1634$  collected by a Canon CR-2 camera. As different cameras are used, the source and target domain images have totally distinct appearances (e.g., color and texture). For other experiments, we utilize  
 420 the REFUGE training dataset as our source domain for training the segmentation network and consider the Drishti-GS and RIM-ONE-r3 datasets as our target domain datasets for adaptation. The dataset Drishti-GS includes 50 images for training and 51 images for testing. The dataset RIM-ONE-r3 is split to include 99 images for training and 60 images for testing. More details of the  
 425 three datasets are shown in Table 1. The optic disc and optical cup regions were carefully delineated by the experts. All of the methods in this section are supervised by the annotations of the source domain and evaluated by the disc and cup dice indices (DI), and the cup-to-disc ratio (CDR) on the target domain.

### 4.2. Data Preprocessing

Firstly, we detect the center of the optic disc by a pre-trained disc-aware ensemble network [5] and then center and crop the optic disc regions. The REFUGE source domain was preprocessed to a size of  $600 \times 600$  while the

Table 1: The details of the three datasets used for the evaluation of the proposed method.

Domain	Dataset	Number of images	Image size	Cameras
Source	REFUGE Train	400	$2124 \times 2056$	Zeiss Visucam 500
	REFUGE Train(w/o label)/Test	400/400	$1634 \times 1634$	Canon CR-2
Target	Drishti-GS Train/Test	50/51	$2047 \times 1759$	-
	RIM-ONE-r3 Train/Test	99/60	$2144 \times 1424$	-

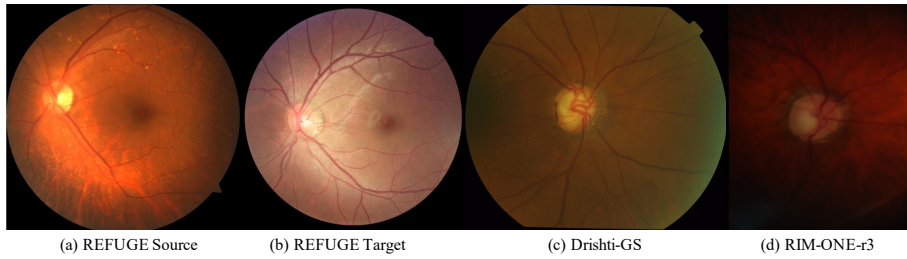


Figure 5: Samples from three datasets (REFUGE, Drishti-GS, and RIM-ONE-r3). The variations in color, field of vision, and textures demonstrate the challenge of domain shift. In particular, the RIM-ONE-r3 dataset carries the strongest deviation from the REFUGE and Drishti-GS datasets and thus poses a greater domain shift.

REFUGE target domains were pre-processed to a size of  $500 \times 500$ . The Drishti-GS and Rim-One datasets were preprocessed to a size of  $700 \times 700$ . This is due to the different sizes of images acquired by the various cameras.

During training, all images are resized to a small size of  $400 \times 400$  to adapt the network’s receptive field. In addition, we used this pre-trained model to understand the image quality of the datasets. For example, some images’ optic discs are not detectable because the image contrast is significantly low. In this case, we removed two low-quality images from the dataset RIM-ONE-r3 to evaluate our method, and the other two datasets remained as the original.

#### 4.3. Implementation Details

The U-Net is used for both student and teacher networks. Four scales of layer blocks are used for network design. Each block includes two groups of

the combination with one convolutional layer, one batch normalization layer, and one ReLU layer. Each layer block is followed by either one max-pooling layer or one up-sampling layer. The details can be found in Fig. 4. To train our network on the REFUGE dataset, we started from a randomly initialized U-Net and used both source and target domain images. However, for another two smaller datasets (e.g, RIM-ONE-r3 and Drishti-GS), to avoid overfitting, we first trained our network on REFUGE using the source domain images with their annotations and then applied our multi-scale domain adaptors to further train the network on the smaller datasets for domain adaptation.

We used the stochastic gradient descent (SGD) optimizer for both student and teacher networks and used the Adam [44] optimizer for all discriminators. We set the initial learning rate as  $1e - 4$  for both student and teacher networks and adjusted the learning rate during training by  $Lr_{initial} \times (1 - Iter_{current}/Iter_{max})^{0.9}$ , where  $Lr_{initial}$  means the initial learning rate,  $Iter_{current}$  means the current iteration number, and  $Iter_{max}$  means the maximum iteration number. We also applied polynomial decay with a power of 0.9 in a total of 200 epochs for network training. Lastly, we used morphological operations (i.e., hole filling) to post-process the results. For the discriminators, we set the initial learning rate as  $2.5e - 5$  and applied the same learning rate adjusting function. All experiments are processed on Python v3.6, and PyTorch v1.0.0 with NVIDIA TITAN Xp GPUs. The detailed implementations and code will be available at <https://github.com/cswin/CADA>.

#### 4.4. Evaluation Metrics

We followed the REFUGE challenge [41] and used a spatial overlap index, dice coefficients (DI), to evaluate the segmentation performance for both OD and OC. We also used the optic cup to disc ratio (CDR) to understand the overall model performance for clinical glaucoma screening convention. The DI

and CDR can be formulated as the below:

$$DI = \frac{2N^{TP}}{2N^{TP} + N^{FP} + N^{FN}} \quad (12)$$

$$CDR = \frac{V^{cup}}{V^{disc}}, \quad \gamma_{CDR} = |CDR^p - CDR^g|$$

where  $N^{TP}$ ,  $N^{FP}$ , and  $N^{FN}$  are the number of true positive, false positive, and  
475 false negative overlapped pixels, respectively; while  $V^{cup}$  and  $V^{disc}$  represents  
the vertical diameters of optic cup and disc, respectively. An absolute difference  
 $\gamma_{CDR}$  between the predicted CDR ( $CDR^p$ ) and the true CDR ( $CDR^g$ ) is used  
for evaluation where a lower  $\gamma_{CDR}$  indicates a better model performance.

#### 4.5. Adaptation to Datasets using Different Fundus Cameras

480 Tables 2-4 lay out the segmentation results on each of the three datasets:  
REFUGE, Drishti-GS, and Rim-One-r3. As mentioned in the prior section, we  
use three metrics to evaluate the model performance, the mean dice coefficient  
for the optic cup (OC), the mean dice coefficient for the optic disc (OD), and  
the mean absolute error for the vertical cup to disc ratio ( $\gamma_{CDR}$ ). A larger value  
485 tending to one for OD and OC means better segmentation results; for  $\gamma_{CDR}$ , a  
smaller value tending to zero represents better results. To test the effectiveness  
of our model, a “Source only” model is trained only over the REFUGE training-  
source domain in a supervised manner and tested upon each respective target-  
testing domain. AdaptSegNet [11] is one of the state-of-the-art unsupervised  
490 domain adaptation methods for image segmentation. The domain adaptation  
algorithm used in AdaptSegNet is also adopted similarly in the solution proposed  
by POSAL [35], the winner of the REFUGE challenge for fundus image OD and  
OC segmentation. CFEA [22] is our previous work published in MICCAI 2019.

In particular, it is observed that training a network solely over the source  
495 domain is insufficient to generalize onto the new domain, thus demonstrating a  
key problem of domain shift. The baseline model AdaptSegNet is outperformed  
by CADA consistently for OD, OC, and  $\gamma_{CDR}$ . Furthermore, the improvement  
of CADA over our prior model CFEA, indicates a new paradigm for model  
development through multi-scale inputs and multi-discriminators. A sample

Table 2: Results of adapting the source REFUGE domain (Zeiss Visucam 500 camera) to the REFUGE target domain (Canon CR-2 camera).

Evaluation-Index	Source only	AdaptSegNet [11]	CFEA [22]	CADA
Optic Cup	0.7317	0.8198	0.8627	<b>0.8714</b>
Optic Disk	0.8532	0.9315	0.9416	<b>0.9498</b>
$\gamma_{CDR}$	0.0676	0.0588	0.0481	<b>0.0447</b>

500 visualization of the segmentation results is presented Fig. 6. Overall, these results indicate that the proposed framework has the capability of overcoming domain shifts, thus allowing us to build a robust and accurate model.

Table 3: Results of adapting the REFUGE source domain to the Drishti-GS target domain.

Evaluation-Index	Source only	AdaptSegNet [11]	CFEA [22]	CADA
Optic Cup	0.8183	0.8267	0.8271	<b>0.8400</b>
Optic Disk	0.8800	0.8814	0.8875	<b>0.8900</b>
$\gamma_{CDR}$	0.1238	0.1176	0.1133	<b>0.1110</b>

#### 4.6. Comparison with State-of-the-art Domain Adaptation Techniques

We compared with AdaptSegNet [11], which is one of the state-of-the-art  
505 domain adaptation methods for image segmentation. AdaptSegNet adopts adversarial learning with two discriminator networks at the last two output layers. POSAL [35], the winner of the REFUGE challenge, applied a similar technique into their framework for optic disc and cup segmentation at the final layer. In this study, we compared their domain adaptation technique with ours, which  
510 applies multiple domain adaptors with adversarial learning at multi-scales in both feature and output spaces. Moreover, compared to POSAL, our approach applied domain adaptation ensembles inside our framework rather than outside by averaging the results of multiple trained sub-networks. From this perspective, the primary goal of this study is to investigate a novel domain adaptation

Table 4: Results of adapting the REFUGE source domain to the Rim-One-r3 target domain. The weaker performance relative to the domain adaptation over the REFUGE and Drishti-GS target domains is in part, considered to be due to the poor image qualities of the Rim-ONE-r3 dataset and larger domain shift.

Evaluation-Index	Source only	AdaptSegNet [11]	CFEA [22]	CADA
Optic Cup	0.5916	0.6271	0.6351	<b>0.6404</b>
Optic Disk	0.7285	0.7365	0.7506	<b>0.7664</b>
$\gamma_{CDR}$	0.1069	0.1017	0.0947	<b>0.0869</b>

515 framework rather than solely pursuing higher scores by experimenting with different preprocessing techniques and backbone architectures. In particular, we find a significant improvement by extending adversarial learning not only at one scale (i.e., solely the encoder or decoder level), but rather at multiple scales, as well as our other aforementioned contributions.

520 To study the computational performance, we compared both CADA and POSAL on the REFUGE source and target domain experiment, the results of which are shown in Table 5. Due to the variability of factors that affect model running time [45], i.e. the coding framework, structure, model depth, and GPU utilization, a holistic account is described including model procedures, model complexity, training time, and inference time. Two NVIDIA Titan Xp GPU's  
525 were utilized with a mini-batch size of 4 for 100 epochs. In particular, the CADA model was run in a PyTorch framework while the open-access POSAL model was run in a Keras-Tensorflow framework. The discriminator architectures are practically identical and as the discriminators are removed during testing, the testing time is influenced only by the network backbone. It is observed in one  
530 setting, that despite POSAL having fewer parameters than our U-NET based CADA model, our model's training and testing time is more than twice as fast. This computational discrepancy is conjectured to be influenced by the computational burden of the residual network structure [46], the depth of the CNN's, as well as the coding framework differences (differences in parallel computing sup-  
535



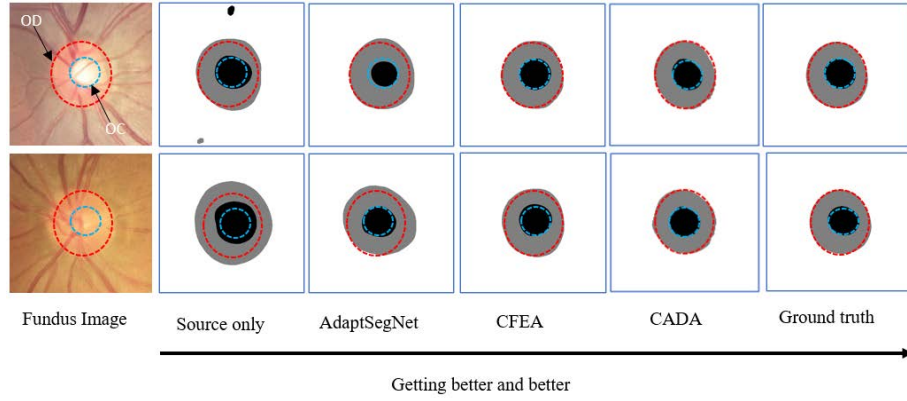


Figure 6: The visual examples of the optic disc (OD) and Optic Cup (OC). The red and blue circles are the position of OD and OC in the ground truth respectively. The contours are indicative of the gap between the model results and the ground truth.

port, memory utilization, etc.). Nevertheless, the inference time suggests that our CADA model is an invaluable tool to practitioners for diagnostic purposes, being substantially faster than manual practice.

Table 5: Computational comparison with other framework on REFUGE challenge [35]

	CADA	POSAL [35]
Average Training Time Per Epoch (s)	148.5	306.5
Backbone Parameters	9.7M	5.8M
Discriminator Parameters	2.8M	2.8M
Average Testing Time (s)	0.023	0.053

#### 4.7. Ablation Study

540 An ablation study is conducted solely over the REFUGE source and target domains respectively to demonstrate the influence of various components of our proposed framework.

#### 4.7.1. The Importance of the Encoder Adaptation (No-Enc-Ada)

To demonstrate the importance of encoder adaptation modules, we remove  
545 the adversarial discriminator  $D_E$  and the MSE module  $mse_E$  from the encoders  
and then retrain the model. Fig. 7 shows the performance comparison of the  
models with modifications on the test dataset. As one can see, without the  
encoder adaptation, the performance drops apparently. This comparison result  
may indicate that the encoder discriminative adaptation module is a crucial  
550 component for learning the domain-invariant representation.

#### 4.7.2. Reduce Uncertainty via Weights Self-ensembling Adaptation (No-SE-Ada)

We also investigate how self-ensembling adaptation affects domain adapta-  
tion performance. For this, we retrain our framework after removing the teacher  
network. The performance comparison of the models with modifications shows  
555 in Fig. 7. As one can see, the average performance on the test dataset is much  
worse than using both adversarial domain confusion and self-ensembling adap-  
tation. Especially, for predicting the CDR, in Fig. 7-c, we can see that without  
weights ensembling, the CDR prediction drops down significantly. This com-  
parison result shows that self-ensembling can significantly improve the model’s  
560 robustness and the generalizability for domain shift. More importantly, weight  
ensembling can reduce the model uncertainty of learning domain-invariant latent  
features when incorporating multiple discriminators in different feature learning  
space. Meanwhile, this is able to enforce all discriminators to maximize their  
ability to discriminate the deeper latent space features.

#### 565 4.7.3. Multiple Discriminators Adaptation Study (CADA-2,3,4D)

We exploit multiple discriminators at the decoder to further investigate the  
maximum power of collaborative feature learning. We compare the results of  
applying different numbers of discriminators to different decoder layers. As one  
can see from CADA-2D, CADA-3D, and CADA-4D in Fig. 7, the more discrim-  
570 inators we use, the better result we obtain. When we apply discriminators to  
all decoder layers (one is at the end of the encoder, and another four are at

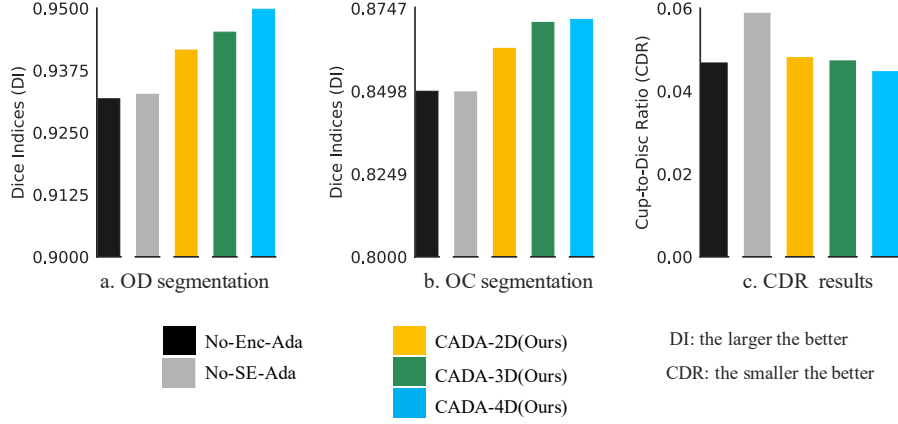


Figure 7: Ablation study: performance comparison of the models with modifications on the REFUGE dataset. No-Enc-Ada means removing the discriminator from the encoder and only applying a discriminator on the decoder. No-SE-Ada means removing self-ensembling (the teacher network) from the proposed CADA.

each layer of the decoder), we obtain the best result. Notably, CADA-2D is the method proposed in our previous work [22]. More importantly, this comparison result further indicates that collaborative feature learning between adversarial adaptation and dynamic weight ensembling can overcome domain shift.

#### 4.7.4. Evaluation of $\lambda$

We evaluate the various combinations of  $\lambda$  for balancing the segmentation, adversarial, and self-ensembling loss. Due to the tremendous combinations, it is impossible to study all of them. We follow the existing studies [47] and use cross-validation to investigate the most effective  $\lambda$  combinations. We find the following combination is the most effective one that can stabilize our framework training:  $\lambda_{seg} = 1, \lambda_{adv}^E = 0.002, \lambda_{adv}^D = 0.018, \lambda_{mse}^E = 0.057, \lambda_{mse}^D = 0.79$ . We also show qualitative results in Fig. 8 to demonstrate the effectiveness of the proposed domain adaptation model. As one can see, these qualitative results are consistent with Fig. 7. This result can further support that collaboration between adversarial learning and dynamic weight ensembling is an effective

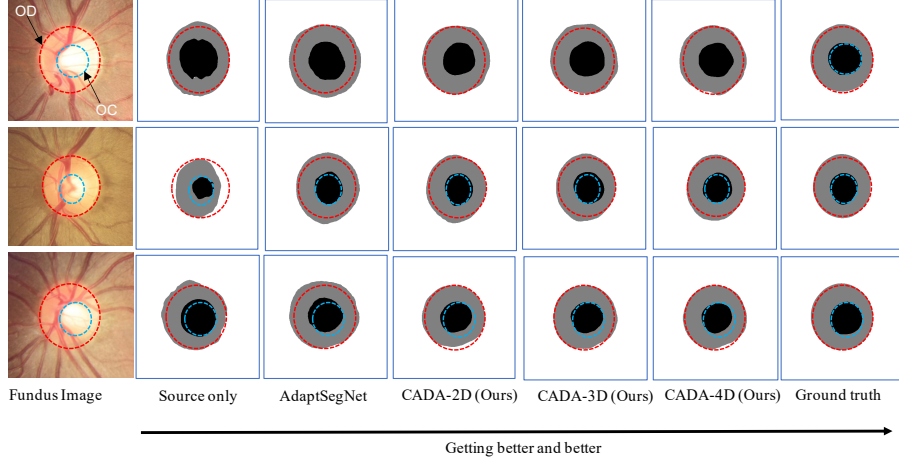


Figure 8: Ablation study comparison: the qualitative examples of the Optic Disc (OD) and Optic Cup (OC) segmentation, where the black and gray region denote the cup and disc segmentation, respectively. The red and blue dash circle are the positions of the OD and OC in the ground truth, respectively. The contours are indicative of the gap between the model results and the ground truth.

strategy to overcome domain shift in fundus images.

## 5. Discussions and Conclusion

In this work, we propose a novel method, CADA, for unsupervised domain  
 590 adaptation across different retinal fundus imaging cameras, specifically over the  
 REFUGE, Drishti-GS, and Rim-ONE-r3 datasets. Our CADA framework col-  
 laboratively combines multiple adversarial discriminative learning and weights  
 self-ensembling to obtain domain-invariant features from both feature represen-  
 tation (encoder) and output space (decoder) in different feature scale levels.  
 595 Multi-scale inputs provide hierarchical features to the collaborative learning  
 process, while multiple domain adaptors collaboratively offer a comprehensive  
 solution for out of distribution (OOD) samples. Weights self-ensembling sta-  
 bilizes adversarial learning and prevents the network from getting stuck in a  
 sub-optimal solution. From a complementary perspective, adversarial discrimi-

600 native learning can consistently provide various model space and time-dependent  
 weights to self-ensembling. With which, we can accelerate the learning of the  
 domain-invariant features and conversely enhance the stabilization of adversarial  
 discriminative learning, forming a fine collaborative circulation and generalized  
 framework. Moreover, we apply multiple discriminators to the multi-scale out-  
 605 put from each layer of the decoder. These adversarial discriminative modules  
 collaboratively encourage the encoder to extract the latent domain-invariant  
 features. Therefore, the collaborative mutual benefits from multi-scale inputs,  
 adversarial discriminative feature learning, weights self-ensembling, and multi-  
 scale outputs during an end-to-end learning process, resulting in a robust and  
 610 accurate model.

Notably, the proposed framework also suggests a generalizable unsupervised  
 learning approach. For example, we could replace the discriminator with the  
 contrastive learning objective functions [20, 48]. With which, the encoder can  
 learn the rich representations rather than the invariant features. Then, we can  
 615 fine-tune the encoder with limited labeled data for specific tasks, such as image  
 classification and segmentation. Simultaneously transferring weights with EMA  
 from both encoder and decoder during model training is a significant novelty  
 compared to existing representation learning methods.

In terms of the running time, due to the multiple discriminator architecture,  
 620 our framework needs relatively more computational during the training stage,  
 compared to AdaptSegNet [11], to help the segmentation network to adapt to  
 the target domain. However, in the testing stage, the computational costs are  
 the same as a standard U-Net network, as the images only need to go through the  
 TTN network. Experimental results demonstrate the superiority of our domain  
 625 adaptation method over other methods either by a significant performance gain  
 or computational efficiency. Our approach potentiates a general and extendable  
 framework to other semi-supervised and unsupervised representation learning  
 problems.

Lastly, although we have shown marked advantages of our method, the cur-  
 630 rent study has some limitations that we hope to address in the future. First, it

is hard to balance the contributions (e.g., learning losses) from multiple domain  
 adaptors during training. To find the optimal weight for each domain adaptor,  
 we currently have to apply a massive grid-search, which can be time-consuming.  
 Second, our framework solely investigates one type of encoder-decoder network  
 635 architecture. It would be interesting to understand how our framework can im-  
 prove other architectures for domain adaptation and performance across various  
 tasks. The current work sheds light on the underlying superiority of applying  
 multiple domain adaptors at hierarchical multi-scale feature and output space.

## Acknowledgment

640 Research reported in this publication is partially supported by the National  
 Science Foundation under Grant No. IIS-1564892 and IIS-1908299, the Uni-  
 versity of Florida Informatics Institute Junior SEED Program (00129436), and  
 the UF Clinical and Translational Science Institute, which is supported in part  
 by the NIH National Center for Advancing Translational Sciences under award  
 645 number UL1 TR001427. The content is solely the responsibility of the authors  
 and does not necessarily represent the official views of the National Institutes  
 of Health and the National Science Foundation.

## References

- [1] D. Regan, D. Neima, Low-contrast letter charts in early diabetic retinopa-  
 650 thy, ocular hypertension, glaucoma, and parkinson’s disease., British Jour-  
 nal of Ophthalmology 68 (12) (1984) 885–889.
- [2] A. Almazroa, R. Burman, K. Raahemifar, V. Lakshminarayanan, Optic  
 disc and optic cup segmentation methodologies for glaucoma image detec-  
 tion: a survey, Journal of Ophthalmology 2015.
- 655 [3] M. E. Danielewska, P. Krzyżanowska-Berkowska, D. R. Iskander, Glau-  
 comatous and age-related changes in corneal pulsation shape. the ocular  
 dicotism, PloS One 9 (7) (2014) e102814.

- [4] A. Almazroa, A novel automatic optic disc and cup image segmentation system for diagnosing glaucoma using riga dataset, Ph.D. thesis, University of Waterloo (2016).
- [5] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, X. Cao, Joint optic disc and cup segmentation based on multi-label deep network and polar transformation, *IEEE Transactions on Medical Imaging* 37 (7) (2018) 1597–1605.
- [6] A. 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) (2017) 618–624.
- [7] G. Lim, Y. Cheng, W. Hsu, M. L. Lee, Integrated optic disc and cup segmentation with deep learning, in: *International Conference on Tools with Artificial Intelligence (ICTAI)*, IEEE, 2015, pp. 162–169.
- [8] J. H. Tan, U. R. Acharya, S. V. Bhandary, K. C. Chua, S. Sivaprasad, Segmentation of optic disc, fovea and retinal vasculature using a single convolutional neural network, *Journal of Computational Science* 20 (2017) 70–79.
- [9] Y. Ganin, V. Lempitsky, Unsupervised domain adaptation by backpropagation, *arXiv preprint arXiv:1409.7495*.
- [10] K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert, et al., Unsupervised domain adaptation in brain lesion segmentation with adversarial networks, in: *International Conference on Information Processing in Medical Imaging (IPMI)*, Springer, 2017, pp. 597–609.
- [11] Y.-H. Tsai, W.-C. Hung, S. Schuster, K. Sohn, M.-H. Yang, M. Chandraker, Learning to adapt structured output space for semantic segmentation, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7472–7481.

- [12] E. Tzeng, J. Hoffman, K. Saenko, T. Darrell, Adversarial discriminative domain adaptation, in: Proceedings of the IEEE International Conference on Computer Vision (CVPR), Vol. 1, 2017, p. 4.
- [13] D. Liu, D. Zhang, Y. Song, F. Zhang, L. O'Donnell, H. Huang, M. Chen, W. Cai, Pdam: A panoptic-level feature alignment framework for unsupervised domain adaptive instance segmentation in microscopy images, IEEE Transactions on Medical Imaging.
- [14] D. Liu, D. Zhang, Y. Song, F. Zhang, L. O'Donnell, H. Huang, M. Chen, W. 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, 2020, pp. 4243–4252.
- [15] Y. Wu, Y. Xia, Y. Song, D. Zhang, D. Liu, C. Zhang, W. Cai, Vessel-net: retinal vessel segmentation under multi-path supervision, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 264–272.
- [16] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, T. Darrell, Cycada: Cycle-consistent adversarial domain adaptation, arXiv preprint arXiv:1711.03213.
- [17] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks, The Journal of Machine Learning Research 17 (1) (2016) 2096–2030.
- [18] S. Laine, T. Aila, Temporal ensembling for semi-supervised learning, arXiv preprint arXiv:1610.02242.
- [19] G. French, M. Mackiewicz, M. Fisher, Self-ensembling for visual domain adaptation, arXiv preprint arXiv:1706.05208.



- [20] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.
- [21] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531.
- [22] P. Liu, B. Kong, Z. Li, S. Zhang, R. Fang, Cfea: Collaborative feature ensembling adaptation for domain adaptation in unsupervised optic disc and cup segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 521–529.
- [23] R. W. Beck, G. E. Servais, S. S. Hayreh, Anterior ischemic optic neuropathy: Ix. cup-to-disc ratio and its role in pathogenesis, *Ophthalmology* 94 (11) (1987) 1503–1508.
- [24] A. Le, B. N. Mukesh, C. A. McCarty, H. R. Taylor, Risk factors associated with the incidence of open-angle glaucoma: the visual impairment project, *Investigative Ophthalmology & Visual Science* 44 (9) (2003) 3783–3789.
- [25] E. Trucco, A. Ruggeri, T. Karnowski, L. Giancardo, E. Chaum, J. P. Hubschman, B. Al-Diri, C. Y. Cheung, D. Wong, M. Abramoff, et al., Validating retinal fundus image analysis algorithms: issues and a proposal, *Investigative Ophthalmology & Visual Science* 54 (5) (2013) 3546–3559.
- [26] S. Yuheng, Y. Hao, Image segmentation algorithms overview, arXiv preprint arXiv:1707.02051.
- [27] H. Fu, J. Cheng, Y. Xu, C. Zhang, D. W. K. Wong, J. Liu, X. Cao, Disc-aware ensemble network for glaucoma screening from fundus image, *IEEE Transactions on Medical Imaging (TMI)* 37 (11) (2018) 2493–2501.
- [28] M. Wang, W. Deng, Deep visual domain adaptation: A survey, *Neurocomputing* 312 (2018) 135–153.

- [29] K. Saenko, B. Kulis, M. Fritz, T. Darrell, Adapting visual category models to new domains, in: European Conference on Computer Vision (ECCV), Springer, 2010, pp. 213–226.
- [30] Y. Taigman, A. Polyak, L. Wolf, Unsupervised cross-domain image generation, arXiv preprint arXiv:1611.02200.
- [31] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, arXiv preprint.
- [32] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1125–1134.
- [33] M.-Y. Liu, T. Breuel, J. Kautz, Unsupervised image-to-image translation networks, in: Advances in Neural Information Processing Systems (NIPS), 2017, pp. 700–708.
- [34] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv preprint arXiv:1511.06434.
- [35] S. Wang, L. Yu, X. Yang, C.-W. Fu, P.-A. Heng, Patch-based output space adversarial learning for joint optic disc and cup segmentation, IEEE transactions on medical imaging 38 (11) (2019) 2485–2495.
- [36] G. French, M. Mackiewicz, M. Fisher, Self-ensembling for visual domain adaptation, in: International Conference on Learning Representations (ICLR), 2018.
- [37] X. Zhu, J. Pang, C. Yang, J. Shi, D. Lin, Adapting object detectors via selective cross-domain alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 687–696.

- [38] A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, in: *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 1195–1204.
- 770 [39] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Springer, 2015, pp. 234–241.
- 775 [40] B. Xu, N. Wang, T. Chen, M. Li, Empirical evaluation of rectified activations in convolutional network, *arXiv preprint arXiv:1505.00853*.
- [41] J. I. Orlando, H. Fu, J. B. Breda, K. van Keer, D. R. Bathula, A. Diaz-Pinto, R. Fang, P.-A. Heng, J. Kim, J. Lee, et al., Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs, *Medical image analysis* 59 (2020) 101570.
- 780 [42] J. Sivaswamy, S. Krishnadas, A. Chakravarty, et al., A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis.
- [43] F. Fumero, S. Alayón, J. L. Sanchez, J. Sigut, M. Gonzalez-Hernandez, Rim-one: An open retinal image database for optic nerve evaluation, in: *2011 24th international symposium on computer-based medical systems (CBMS)*, IEEE, 2011, pp. 1–6.
- 785 [44] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980*.
- [45] A. Canziani, A. Paszke, E. Culurciello, An analysis of deep neural network models for practical applications, *CoRR abs/1605.07678*.
- 790 [46] Y. Yi, Z. Zhang, W. Zhang, C. Zhang, W. Li, T. Zhao, Semantic segmentation of urban buildings from vhr remote sensing imagery using a deep convolutional neural network, *Remote Sensing* 11 (15).

- 795 [47] C. S. Perone, P. Ballester, R. C. Barros, J. Cohen-Adad, Unsupervised domain adaptation for medical imaging segmentation with self-ensembling, arXiv preprint arXiv:1811.06042.
- [48] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR, 2020, pp. 1597–1607.