Structure-Guided Cross-Attention Network for Cross-Domain OCT Fluid Segmentation

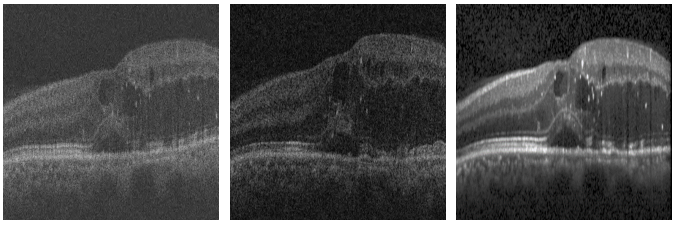
Xingxin He, Zhun Zhong, Leyuan Fang, *Senior Member, IEEE*, Min He, and Nicu Sebe, *Senior Member, IEEE*

***Abstract*—Accurate retinal fluid segmentation on Optical Coherence Tomography (OCT) images plays an important role in diagnosing and treating various eye diseases. The art deep models have shown promising performance on OCT image segmentation given pixel-wise annotated training data. However, the learned model will achieve poor performance on OCT images that are obtained from different devices (domains) due to the domain shift issue. This problem largely limits the real-world application of OCT image segmentation since the types of devices usually are different in each hospital. In this paper, we study the task of cross-domain OCT fluid segmentation, where we are given a labeled dataset of the source device (domain) and an unlabeled dataset of the target device (domain). The goal is to learn a model that can perform well on the target domain. To solve this problem, in this paper, we propose a novel Structure-guided Cross-Attention Network (SCAN), which leverages the retinal layer structure to facilitate domain alignment. Our SCAN is inspired by the fact that the retinal layer structure is robust to domains and can reflect regions that are important to fluid segmentation. In light of this, we build our SCAN in a multi-task manner by jointly learning the retinal structure prediction and fluid segmentation. To exploit the mutual benefit between layer structure and fluid segmentation, we further introduce a cross-attention module to measure the correlation between the layer-specific feature and the fluid-specific feature encouraging the model to concentrate on highly relative regions during domain alignment. Moreover, an adaptation difficulty map is evaluated based on the retinal structure predictions from different domains, which enforces the model focus on hard regions during structure-aware adversarial learning. Extensive experiments on the three domains of the RETOUCH dataset demonstrate the effectiveness of the proposed method and show that our approach produces state-of-the-art performance on cross-domain OCT fluid segmentation.**

***Index Terms*—Optical Coherence Tomography, Retinal Fluid Segmentation, Cross-Domain Segmentation, Retinal Structure.**

I. Introduction

OPTICAL coherence tomography (OCT) [1] generates 3-D images for living tissue at micrometer resolution, which is an indispensable part of quantitative analysis in clinical ophthalmology. Thanks to the development of semantic segmentation techniques [2]–[4], fluid segmentation on OCT images is introduced to aid doctors to diagnose and treat common retinal diseases, such as diabetic macular edema which is the most common cause of visual loss in diabetic retinopathy [5], [6].



(a)

Source/Target	Topcon	Cirrus	Spectralis
Topcon	56.77	9.70	16.31
Cirrus	10.02	45.89	22.09
Spectralis	16.79	7.27	57.54

(b)

Fig. 1. (a) Examples of OCT images from the RETOUCH dataset. The images in the three devices (domains) are very different, leading to a large domain shift. (b) Cross-domain performance without domain adaptation. A model trained on the source domain produces a poor performance on target domains.

Recently, several deep-based approaches [7]–[10] have been proposed to segment OCT fluid and achieve impressive improvement compared with the hand-crafted machine learning-based OCT image segmentation method [11]–[13]. However, existing methods commonly assume that the training and testing data share the same distribution, which largely limits their application in the real world. In clinical ophthalmology, the heterogeneity of OCT devices and acquisition protocols at each hospital may be very different, resulting in large differences in the distribution of data acquired by different scenes (see Fig. 1(a)). Given a model trained on one domain, it may produce a poor performance on the domain that has a large distribution gap to the trained domain (see Fig. 1(b)). This is caused by the domain shift, a well-known problem in the machine learning and deep learning community. One may suggest labeling data in the new domain and training a new model. However, labeling pixel-wise data is expensive and laborious, and requires adequate expert knowledge. As a consequence, labeling data in each new domain is not a permanent and efficient solution.

Unsupervised domain adaptation (UDA) [14], [15] is a promising direction to address the domain shift issue. UDA aims to transfer the knowledge of the labeled source domain

This work was supported in part by the National Natural Science Foundation of China under Grant 61922029, in part by the Science and Technology Plan Project Fund of Hunan Province under Grant 2022RSC3064.

X. He and L. Fang are with the College of Electrical and Information Engineering, Hunan University, Changsha, 410082, China (email: xx_h@hnu.edu.cn; fangleyuan@gmail.com).

M. He is with the College of Electrical and Information Engineering, Hunan University, Changsha, 410082, China, and with the Key Laboratory of Head & Neck Cancer Translational Research of Zhejiang Province, Zhejiang Cancer Hospital, Hangzhou, Zhejiang 310022, China (email: hemin@hnu.edu.cn).

Z. Zhong and N. Sebe are with the Department of Information Engineering and Computer Science (DISI), University of Trento, Trento, TN 38122, Italy (e-mail: zhunzhong007@gmail.com, sebe@disi.unitn.it).

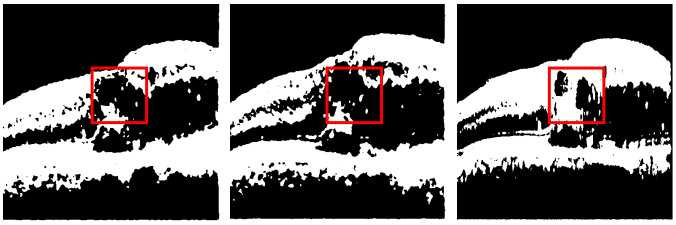


Fig. 2. Retinal layer maps obtained by the OTSU segmentation method on three devices. The regions highlighted in red boxed represent the region of intraretinal fluid.

to the unlabeled target domain so that the adapted model can perform well on the target domain. Although semi-supervised learning has proposed to transfer knowledge from labeled data to unlabeled data [16]–[18], they also ignore the fact that the labeled data and the unlabeled data may come from different domains. UDA has been extensively exploited in the literature, in which the most popular solution [19]–[21] is aligning source and target distributions by adversarial learning [22]. Recently, researchers have paid more attention to semantic segmentation and tried to bridge domain gaps by minimizing the difference in the aspects of image style [23], [24], intermediate latent features [25], [26], late output space [27]–[30], or the combinations of the aforementioned strategies [31]–[33]. Inspired by the success of UDA in semantic segmentation, in this paper, we study the task of UDA in OCT fluid segmentation to address the domain shift issue in a practical application.

Considering the high relation between semantic segmentation and OCT fluid segmentation, most existing UDA methods in semantic segmentation can be applied to OCT fluid segmentation. However, they mostly focus on natural scenes (*e.g.*, street) and design specific modules/algorithms based on corresponding knowledge (*e.g.*, class frequency in the street). Since OCT images are very different from natural images, we should consider medical professional knowledge during OCT image segmentation instead of directly using the ones adopted in natural tasks. In this paper, we aim to achieve this goal from the perspective of anatomical information. In OCT images, the accumulation of fluid will break the normal retinal structure, and the degree of retinal layer deformation is directly bound up with the fluid. Therefore, it is natural to leverage the mutual interaction between the fluid and retinal layer to improve the OCT segmentation result [34]. However, retinal layer segmentation is another challenging task in OCT, requiring expensive labeling efforts to learn an accurate model [10]. This limits the usage of retinal layer segmentation in UDA due to the absence of precise layer labels. Nevertheless, we can easily obtain a rough segmentation map of layer region by a discrepancy image segmentation method [35]. In Fig. 2, we show the rough retinal layer maps of samples across different devices and make the following observations. (1) The rough retinal layer maps can clearly show the retinal structures on different domains, and the overall retinal structure is relatively similar across domains. This indicates that the retinal map is robust to domains. (2) In some specific regions (highlighted in red boxes of Fig. 2), the discrepancies are significant across domains. These regions usually will occur fluids. This indicates that the retinal layer maps can reflect regions that are important to fluid segmentation.

Inspired by the above observations, this paper introduces the Structure-guided Cross-Attention Network (SCAN), which effectively exploits the retinal layer structure to boost domain adaptation. Specifically, in the preparation stage, we obtain the retinal layer maps for both source and target domains using an off-the-shelf segmentation method [35] to generate the rough retinal layer map. In the adaptation stage, we build a multi-task framework to jointly learn the fluid segmentation predictor and retinal layer structure predictor. To leverage the mutual benefit between fluid segmentation and retinal structure, we propose a cross-attention module to estimate the correlation between the layer-specific feature and the fluid-specific feature. This leads the model to focus on highly relative regions during domain alignment. Moreover, the discrepancy between the predicted retinal layer maps of source and target domains is used to guide the domain alignment, which can encourage the model to concentrate on the heterogeneous regions. In summary, the contributions of this paper are as follows:

- We consider the problem of unsupervised cross-domain OCT fluid segmentation, which promotes the study of image processing under multiple domains.
- We observe that within OCT images, the layer region is relatively stable across domains, and the fluid region is highly sensitive to different domains.
- We propose a novel method, SCAN, which leverages the retinal stable and unstable anatomy across domains to guide adversarial training focusing on the difficult regions, then conduct more efficient adaptation.
- Experiments on the RETOUCH dataset across three OCT domains show that our SCAN method outperforms the state-of-the-art UDA methods in all adaptation directions.

II. Related Works

A. Deep Learning for OCT Fluid Segmentation

Most of recent OCT fluid segmentation methods are based on UNet [4] or similar encoder-decoder architectures [36]. Studies in [37], [38] employed the plain UNet architecture to achieve macular fluid segmentation. Rashno *et al.* [39] implemented a graph shortest path technique as post-processing to refine the predictive results. Despite the success of existing methods, most of them mainly focus on supervised learning settings and ignore the domain shift problem. In this paper, we study the unsupervised domain adaptation in OCT fluid segmentation, aiming at advancing the practicability in clinical ophthalmology.

The close interaction of the fluid and the retinal layers is incorporated in many fluid segmentation approaches [7], [8], [11], [40]–[42]. Xu *et al.* [40] designed a two-stage fluid segmentation framework considering the structural relationship between retinal layers and fluids. They first trained a retinal layer segmentation network to extract retinal layer maps. Then, they utilized the layer maps as the constraint to train a UNet in the second stage. Similarly, other studies [41], [42] conducted a graph-cut-based method to obtain the

retinal layers segmentation maps and then combined the maps to train the fluid segmentation model. To further improve the ability of the OCT segmentation method, some works [7], [8], [11] constructed a UNet-based architecture to simultaneously segment retinal layers and fluid trained on a dataset with well-annotated pixel-wise retinal layer and fluid masks. In this paper, we also take advantage of the retinal layer map to solve the fluid segmentation. Different from existing methods, we leverage the retinal layer map to facilitate the process of domain alignment, which helps us obtain a better target model.

B. Unsupervised Domain Adaptation for Semantic Segmentation

UDA for semantic segmentation (UDASS) [43] aims to leverage pixel-wise annotated source data and unlabeled target data to learn a segmentation model that can perform well on the target data. Early UDASS methods were directly transferred from UDA for classification, which focused on aligning source and target data distributions at a latent feature space embedded from the input images. However, as for predicting pixel-wise classes, semantic segmentation is a more challenging task with much higher dimensionality and complexity. Applying such a feature alignment strategy directly to semantic segmentation models will lead to sub-optimization. Therefore, most UDASS methods adopt adversarial learning to multiple levels of the image/feature context [44]. Some works attempt to minimize the pixel dissimilarity directly (*i.e.*, performing style transfer) [23], [45]–[47]. For example, Li *et al.* [48] adopted an extra image translation model to translate a source domain image into a target-style image. Chen *et al.* [23] translated both from the source domain to the target domain and from the target domain to the source domain and then trained a pair of adaptation models with a cross-domain consistency loss. Others focus on minimizing the discrepancy at the latent feature level [43], [49]–[51]. For example, Zhang *et al.* [50] performed adversarial learning on latent features with multiple constraints to regularize the output on the target domain. Besides aligning latent features, performing aligning to the final output space is also valid in semantic segmentation [27]–[29]. Furthermore, combining alignment strategies at different levels also have been studied. For example, Hoffman *et al.* [45] and Luo *et al.* [52] combined feature and output-level alignments.

The aforementioned UDASS methods mainly focus on street scenes. Although they have achieved good performance on such natural images, it is still difficult to directly apply them to medical images since the medical and street images are very different, and the specific physiological structure important to analyze medical images is ignored. The common assumption and prior knowledge used for natural scenes (e.g., class frequency in the street) are unavailable in OCT images. For a specific task in medical imaging, the medical domain knowledge is usually used to improve deep model training [18], [53], [54]. Therefore, we should consider retina-specific knowledge to improve the training. We observe that OCT images captured from different devices have a relatively stable region (i.e., the layer region) and a domain-sensitive region (i.e., the fluid region). With the stable region, we can learn to extract domain-invariant features. Moreover, with the domain-sensitive region, we can locate where we should pay more attention to align the represented features. Besides, the layer region and the fluid region are highly coupled. Learning their correlation can further improve the representation for the final fluid segmentation task.

III. Methods

A. Problem Formulation

In this paper, we consider the problem of cross-domain OCT segmentation. Specifically, we are given the source OCT data $\mathbb{X}^s \in \mathbb{R}^{H\times W\times 1}$ with its corresponding pixel-wise manual annotation of the retinal fluid $\mathbb{Y}^s \in (1,...,C)^{H\times W}$, and the unlabeled target OCT data $\mathbb{X}^t \in \mathbb{R}^{H\times W\times 1}$. H, W, and C are the height, width, and number of fluid categories of OCT samples, respectively. The goal is to learn a semantic segmentation model $f(\mathbb{X}^s, \mathbb{Y}^s, \mathbb{X}^t|\theta)$ performing well on the target OCT data $\mathbb{X}^t$. θ denotes the learnable parameters of the model.

B. Overview

In this paper, we propose a novel method called Structure-guided Cross-Attention Network (SCAN) to solve the problem of cross-domain OCT fluid segmentation. The framework of our SCAN method is illustrated in Fig. 3. Our network comprises a shared feature encoder, two layer predictors, one fluid predictor, and one domain discriminator. In the training phase, given two OCT images from the source and target domains, we first use a shared encoder to extract the latent features. Then, two feature bottlenecks are used to convert the latent features to layer-related features and fluid-related features. Our training losses are computed on two feature-levels. In the first feature-level, we use the fluid-related features of the source domain to learn the fluid predictor based on fluid supervised loss. The layer-related features of both domains are used to optimize the layer predictor based on layer supervised loss and structure invariant loss. To align the two domains, we first obtain the outputs of the fluid predictor for both domains and train the domain discriminator with adversarial loss. To further leverage the mutual benefit between the retinal layer structure and the fluid, we propose a cross-attention module to calculate co-attention features. In this new feature-level, we adopt similar losses to the first feature-level to train the layer predictor, fluid predictor, and domain discriminator. The difference is that we estimate the discrepancy map between source and target layer predictions and use it to guide adversarial learning. In the inference phase, we focus on fluid segmentation and thus remove the layer predictors and domain discriminator.

C. Retinal Layer Prediction

Before introducing our SCAN, we first present how we get the rough layer maps for subsequent domain adaptation.

Previous works involving retinal layers to enhance fluid segmentation usually train an extra layer prediction model with manually annotated retinal layer maps [7], [8], [11]. However,

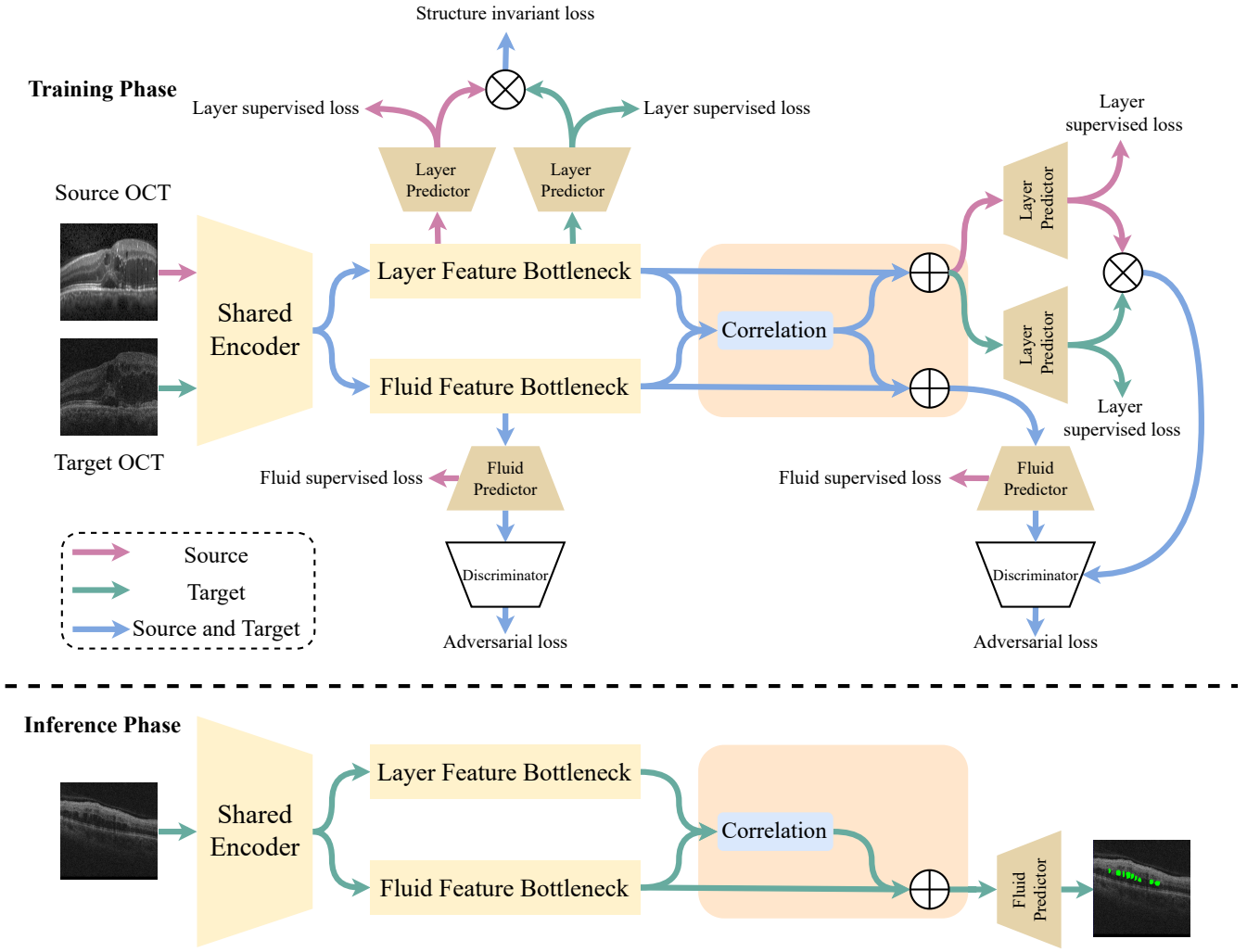


Fig. 3. The framework of our SCAN. We perform layer/fluid prediction and adversarial learning on two feature-levels (*before and after the correlation module*). The two feature-levels share the same layer predictors, the same fluid predictor, and the same domain discriminator. In the first feature level, A pair of source/target domain OCT images are fed into a shared encoder and converted to latent features. Then the layer/fluid feature bottleneck transforms the latent features into layer/fluid-specific features. The layer/fluid-specific features are further used to make retinal layer/fluid predictions via the layer/fluid predictors. Two domain-specific layers, supervised losses, and a structure-invariant loss across domains are used to train the source and target layer predictors. A fluid-supervised loss for the source domain is used to train the fluid predictor, and an adversarial loss is utilized to align the target/source features. Before the second feature-level, we introduce a cross-attention module to exploit the mutual interaction between layer/fluid-specific features. In the second feature-level, the layer/fluid attention features are fed again to the layer/fluid predictors to make the final layer/fluid prediction. We perform the same fluid/layer supervised loss to train the layer/fluid predictor. Besides, we evaluate an adaptation difficulty map based on the discrepancy between layer predictions of different domains. We perform structure-aware adversarial learning with the adaptation difficulty map, which focuses on the hard region. In the inference phase, we remove the layer predictor and the adversarial learning parts. The trained model takes a target domain OCT and directly outputs a fluid prediction map.

obtaining such additional layer annotation in the clinic is expensive and impractical. In addition, this will lead to another annotation issue in unsupervised domain adaptation. In this paper, instead of generating well-annotated retinal layer maps, we adopt a simple yet effective threshold segmentation method to generate rough layer maps. Specifically, we use OTSU [35] to distinguish an OCT image into two classes, *i.e.*, the layer region and the non-layer region, which can be regarded as a rough layer map for each OCT image. In the preparation stage, we generate rough layer maps for both source and target data, which are represented as $\mathbb{L}^s \in (0,1)^{W\times H} = OTSU(\mathbb{X}^s)$ and $\mathbb{L}^t \in (0,1)^{W\times H} = OTSU(\mathbb{X}^t)$, respectively.

D. Domain Adaptation with Multi-Task Learning

In our SCAN, we build a multi-task learning framework, where we simultaneously learn fluid segmentation, retinal layer structure prediction, and domain alignment. For fluid segmentation, we use the labeled source domain to train the fluid predictor with supervised loss. For layer prediction, we regard the generated rough layer maps $L \in \mathbb{L}$ as the pseudo layer labels. The layer predictors are trained with cross-entropy loss and structure invariant loss on both source and target samples. The domain alignment is conducted based on typical adversarial learning. Specifically, domain discriminator D tries to discriminate whether a segmentation map is from the source or target domain. The feature encoder and fluid predictor are regarded as the generator, aiming to fool the domain

discriminator. We will next introduce the loss functions in detail.

Given a source OCT image X^s and a target OCT image X^t, a weight-shared convolutional encoder $E(.)$ is used to obtain the latent feature representation $F = E(X)$, where F and X without superscript denote the latent features and input OCT images from either source or target OCT domain. Then, a layer feature bottleneck $f_l(.)$ and a fluid feature bottleneck $f_f(.)$ are applied to convert the latent feature F to layer-specific features $F_l = f_l(F)$ and fluid-specific features $F_f = f_f(F)$, respectively. Here, we call these two types of features the first-level feature. In the next section, we will introduce the co-attention feature, which will be regarded as the second-level feature.

For fluid segmentation, a domain-shared fluid predictor P_f is optimized with the cross-entropy loss. Since only the source fluid annotation Y^s is available, we only calculate the cross-entropy loss on the source data:

$$\mathcal{L}_f^s = -\frac{1}{|S^s|}\sum_{h,w}^{H,W}\sum_{c}^{C} Y^s_{(h,w,c)} \log S^s_{(h,w,c)}, \tag{1}$$

where $S_f = P_f(F_f)$ is the predicted fluid segmentation map. $|\cdot|$ denotes the cardinality of a set.

For domain alignment, we forward both the target and source fluid predicted maps S_f to the domain discriminator $D(.)$ and train the discriminator with a binary-classification loss:

$$\mathcal{L}_d = -\sum_{h,w}^{H,W}\sum_{c}^{C}(log(D(S_f^s)) - log(1 - D(S_f^t))). \tag{2}$$

This function enables the domain discriminator to distinguish whether the fluid segmentation map is from the source or target domain.

On the other hand, an adversarial loss is used to fool the domain discriminator:

$$\mathcal{L}_{adv} = -\sum_{h,w}^{H,W}\sum_{c}^{C} log(D(S_f^t)_{(h,w,1)}). \tag{3}$$

Note that this loss is used to optimize the shared encoder, fluid feature bottleneck, and fluid predictor instead of the domain discriminator.

For layer prediction, we have two individual layer predictors $P^s(.)$ and $P^t(.)$ composed of multiple deconvolution blocks for each domain. Given the layer-specific features, we first obtain the layer predicted maps $S_l^s = P_s(F_l^s)$ and $S_l^t = P_t(F_l^t)$. Then, the layer prediction loss is calculated with the cross-entropy loss between the predicted layer map and the generated pseudo layer label:

$$\mathcal{L}_l = -\frac{1}{|S_l|}\sum_{h,w}^{H,W}\sum_{c\in 0,1} L_{(h,w,c)} \log S_{l(w,h,c)}. \tag{4}$$

In addition to the supervised layer prediction loss, we also calculate the structure invariant loss by measuring the discrepancy between source and target layer predicted maps:

$$\mathcal{L}_{si} = \sum_{h,w}^{H,W}\sum_{c}^{C}\frac{S^s_{l(h,w,c)}\cdot S^t_{l(h,w,c)}}{max(\|S^s_{l(h,w,c)}\|\cdot\|S^t_{l(h,w,c)}\|,\epsilon)}, \tag{5}$$

where a small value $\epsilon = e^{-8}$ avoids the denominator being zero. $\|\cdot\|$ denotes the magnitude of the vector. This function ensures that the retinal structure is robust to domain variations.

E. Cross-Attention between Layer Structure and Fluid

In the last section, we have introduced a multi-task learning based domain adaptation framework. Next, we aim to fully exploit the mutual benefit between layer structure and fluid during the domain adaptation process. To achieve this goal, in this paper, we adopt an attention module [55] to measure the correlation between the layer-specific feature and the fluid-specific feature, which is used to produce co-attention features.

Specifically, a correlation module is applied to the layer-specific feature F_l and the fluid-specific feature F_f:

$$A_l = softmax(F_l F_f^T)F_l + F_l, \tag{6}$$

and

$$A_f = softmax(F_f F_l^T)F_f + F_f, \tag{7}$$

where the uppercase superscript T is the transpose operation. A_l and A_f are the obtained co-attention features. The co-attention features emphasize the highly relative region (*e.g.*, the deformation of retinal layers or the accumulation of fluid) between the layer-specific and fluid-specific features.

The co-attention features are regarded as the second-level features, which are also used to calculate the fluid segmentation loss Eq. 1 and layer prediction loss Eq. 4. These two losses are represented as $\mathcal{L}_f^{s'}$ and $\mathcal{L}_l^{'}$, respectively. We will next present the structure-aware adversarial learning process, which differs from the adversarial learning on the first feature-level.

F. Structure-Aware Adversarial Learning

In structure-aware adversarial learning, we aim to encourage the model to focus on aligning the regions with significant domain discrepancies while paying less attention to similar regions. For this purpose, we first calculate the discrepancy map between the source and the target layer predictions, formulated as,

$$M_{dis} = \frac{S_l^{s'}\cdot S_l^{t'} + \|S_l^{s'}\|\cdot\|S_l^{t'}\|}{2max(\|S_l^{s'}\|\cdot\|S_l^{t'}\|,\epsilon)}. \tag{8}$$

The layer discrepancy map $M_{dis} \in (0,1)$ can be treated as a domain discrepancy map, where smaller values represent low similarities (*i.e.*, regions are difficult to align) and vice versa. We utilize M_{dis} to highlight the difficult-to-align regions and suppress the remaining easy regions. Given M_{dis}, structure-aware adversarial learning can be formulated as:

$$\begin{aligned}\mathcal{L}_d^{'} = -\sum_{h,w}^{H,W}\sum_{c}^{C}(&log(D((2-M_{dis})S_f^{s'}))- \\ &log(1 - D((2-M_{dis})S_f^{t'}))),\end{aligned} \tag{9}$$

and

$$\mathcal{L}_{adv}^{'} = -\sum_{h,w}^{H,W}\sum_{c}^{C} log(D((2-M_{dis}S_f^{t'})_{(h,w,1)})), \tag{10}$$

where $\mathcal{L}_d^{'}$ and $\mathcal{L}_{adv}^{'}$ are used to optimize the domain discriminator and the generator (shared encoder, fluid feature bottleneck, and fluid predictor), respectively.

G. Overall Training Losses

In the training phase, we adopt a two-step optimization strategy. We first minimize the fluid segmentation losses ($\mathcal{L}_f^s$ and $\mathcal{L}_f^{s^{'}}$), layer prediction losses ($\mathcal{L}_l$ and $\mathcal{L}_l^{'}$), structure invariant loss ($\mathcal{L}_{si}$), and the adversarial losses ($\mathcal{L}_{adv}$ and $\mathcal{L}_{adv}^{'}$), which can be summarized as the generating loss,

$$\begin{aligned}\mathcal{L}_G =&\mathcal{L}_l + \mathcal{L}_f^s + \mathcal{L}_{si} + \mathcal{L}_{adv} \\ &+\mathcal{L}_l^{'} + \mathcal{L}_f^{s^{'}} + \mathcal{L}_{adv}^{'}.\end{aligned} \tag{11}$$

We update the learnable parameters of all modules except for the domain predictor. Then, we freeze all other modules and update the domain predictor by minimizing the domain discrimination losses, formulated as,

$$\mathcal{L}_D = \mathcal{L}_d + \mathcal{L}_d^{'}. \tag{12}$$

$\mathcal{L}_G$ and $\mathcal{L}_D$ are alternatively optimized in each training step. We also provide an algorithmic way for the SCAN's training process in Algorithm 1.

In the inference phase, we discard the layer predictors and the domain discriminator and directly generate the fluid segmentation map of the second feature-level.

IV. Experimental Results

A. Experimental Setup

Dataset We use the public RETOUCH dataset [56] to conduct experiments on cross-domain OCT fluid segmentation. RETOUCH includes OCT images from three different OCT devices, which are regarded as three domains. This dataset contains a total of 70 OCT volumes, where 24 volumes are acquired with the Cirrus device (Zeiss), 24 volumes are acquired with the Spectralis device (Heidelberg), and 22 volumes are acquired with the Topcon device (T-1000 and T-2000). For each volume, there are 128, 49, and 128 B-scans with resolutions of 512×1024, 512×496, and 512×885 for Cirrus, Spectralis and Topcon, respectively. In this dataset, three different fluid types, *i.e.*, the intraretinal fluid (IRF), subretinal (SRF), and PED (pigment epithelial detachments), are manually annotated. The RETOUCH challenge evaluates the results upon submission and the ground truth of the RETOUCH test data is unknown to the public. Therefore, we split the annotated 70 OCT volumes into training, validation, and testing subset as shown in Table I in our experiments (*i.e.*, 60% for training, 20%, and 20% of the 70 OCT volumes for validation and testing, respectively). To evaluate our SCAN method, we perform 5-fold cross-validation based on the data splitting above. In other words, we randomly split the dataset into five equal groups. Then, we take one group as the testing set, the remaining three groups as the training set, and one group as the validation set. We repeat the above process until every group is taken as testing set exactly once.

Evaluation Protocol For each experiment, we select the training data from two domains and regard them as the labeled source domain and an unlabeled target domain, respectively. The model trained on these two domains is used to evaluate the testing set of the target domain. For example, in our experiments, "C→T" indicates using Cirrus as the source domain and Topcon as the target domain. The Dice Similarity Coefficient (DSC) score is applied to evaluate the segmentation performance:

$$DSC = 2 \times \frac{|S_p \cap S_g|}{|S_p| + |S_g|}, \tag{13}$$

where S_p is the predicted segmentation map and the S_g is the ground-truth map.

TABLE I
THE SPLITTING DETAILS OF THE RETOUCH DATASET.

Device	Training Volumes (Slices)	Validation Volumes (Slices)	Testing Volumes (Slices)
Cirrus	14 (1792)	5 (640)	5 (640)
Spectralis	14 (686)	5 (245)	5 (245)
Topcon	13 (1664)	4 (512)	5 (640)

Implementation Details During training, we use the Adam optimizer [57] with a weight decay of 1e-4 to optimize the parameters of the model. The initial learning rate is set to 1e-3, which is divided by 10 after 100 epochs. The batch size is set to 16, which contains 8 source images and 8 target images. All images are resized to 496×496. Random horizontal flipping is used as the data augmentation strategy. The input images are standardized and normalized in a domain-wise manner before being fed into the model. The statistics used for the standardization of each domain are as follows:

Domain	μ	σ
Topcon	64.929860	19.604365
Cirrus	36.009738	23.599870
Spectralis	41.310338	39.703057

We train the model for 200 epochs. All experiments are conducted under an Ubuntu 20.04.1 LTS operating system with CPU Intel® Xeon(R) E5-2678 v3 2.50GHz, GPU NVIDIA GeForce RTX 3090, and RAM of 128 GB. The proposed method is built on PyTorch 1.7.0 [58]. Our SCAN has a total of 5.32 MB trainable parameters. Under this environment, the training time of our method is about 7 hours for each adaptation direction. As for inference time, our method removes the layer/fluid prediction heads in the first-feature level and only reserve the segmentation backbone and the fluid prediction head in the second-feature level. Therefore, the adapted model will be as efficient as a regular segmentation model. The inference takes about 4 ms for a single OCT image fluid segmentation on our device. For all methods, we use UNet [4] as the backbone for a fair comparison. During the inference phase, we forward the target samples into the model and produce the corresponding segmentation maps, without using any post-processing operations and model ensemble techniques.

B. Ablation Study

There are three main components in the proposed SCAN, *i.e.*, multi-task adversarial learning, cross-attention learning,

Algorithm 1 The training process of SCAN

Input: Source domain OCT samples X^s, Target domain OCT samples X^t, Source domain labels Y^s
Output: Adapted model

Initialization random weights
Iteration $n = 1$;**Begin**
for $n < maxiteration$ **do**
  Obtain Layer Predictions $L^s \in (0,1)^{W\times H} = OTSU(X^s)$ $L^t \in (0,1)^{W\times H} = OTSU(X^t)$
  Latent feature representation $F = E(X)$
  Convert F *into layer* $-$ *specific features* $F_l = f_l(F)$ *and fluid* $-$ *specific features* $F_f = f_f(F)$
  Get the fluid segmentation map $S_f = P_f(F_f)$
  Cauculate source domain fluid segmentation loss Eq 1
  Cauculate discriminative loss Eq 2
  Cauculate adverserial loss Eq 3
  Cauculate sturcutre invariant loss Eq 5
  Compute the cross $-$ *attention feautres by* Eq 6 *and* Eq 7
  Estimate the discrepency map M_{dis} *by* Eq 8
  Cauculate sturcture guided discriminative loss Eq 9
  Caucalate structure guided adverserial loss Eq 10
  Compute backpropogation of Eq 11
  Update weights of $E()$, $f_f()$, $f_l()$, $P_l()$, $P_f^s()$ $P_l'()$, *and* $P_f^{s'}()$
  Update weights of $D()$, *and* $D'()$
end for

TABLE II
SEGMENTATION RESULTS OF ABLATION EXPERIMENTS . "T", "C", AND "S" INDICATE TOPCON, CIRRUS, AND SPECTRALIS DOMAIN OF OCT IMAGES. WE REPORT THE MEAN ± STANDARD DEVIATION DSC SCORE OF 5-FOLD CROSS VALIDATION. THE BEST RESULT OF MEAN DSC SCORE IN EACH COLUMN IS HIGHLIGHT IN **BOLD** (UNIT: %). ADV: ADVERSARIAL LEARNING, MTL: MULTI-TASK LEARNING, CA: CROSS-ATTENTION LEARNING, S-ADV: STRUCTURE-AWARE ADVERSARIAL LEARNING.

Adv	MTL	CA	S-Adv	S→T	C→T	T→C	S→C	T→S	C→S	Average
✗	✗	✗	✗	16.79 ± 4.13	10.02 ± 8.96	9.70 ± 4.81	7.27 ± 5.45	16.31 ± 2.54	22.09 ± 4.82	13.70
✓	✗	✗	✗	19.80 ± 2.22	7.96 ± 1.36	29.82 ± 3.75	20.64 ± 1.90	46.61 ± 0.82	41.82 ± 3.23	27.77
✓	✓	✗	✗	24.55 ± 0.28	5.92 ± 3.13	**31.67** ± 1.11	24.82 ± 2.49	46.74 ± 0.85	40.53 ± 1.25	29.04
✓	✓	✓	✗	21.02 ± 1.70	15.55 ± 7.63	28.78 ± 5.66	24.53 ± 4.73	43.66 ± 1.34	**46.18** ± 3.89	29.95
✓	✓	✗	✓	24.81 ± 1.72	8.48 ± 6.80	29.60 ± 2.88	**28.58** ± 2.77	47.64 ± 1.24	37.73 ± 4.16	29.47
✓	✓	✓	✓	**39.63** ± 1.30	**17.25** ± 4.44	30.61 ± 1.59	24.88 ± 1.55	**49.76** ± 2.35	42.11 ± 1.22	**34.04**

and the structure-aware adversarial leaning. In Table II, we conduct ablation experiments to investigate the effectiveness of each component.

Effectiveness of Vanilla Adversarial Learning. Our method is constructed based on traditional adversarial learning. We first verify its effect on domain adaptation. As shown in the first and second rows in Table II, vanilla adversarial learning can significantly improve the results in all adaptation directions. The method of the first row indicates the source training. Specifically, the average DSC in all adaptation directions is improved from 13.70% to 27.77% when using the vanilla adversarial learning. In the following ablations, the model with vanilla adversarial learning is regarded as the baseline of our method. We next focus on investigating the effectiveness of the proposed three components.

Effect of Multi-Task Learning. To verify the effectiveness of the multi-task learning, we directly add the layer prediction on the first feature-level into the baseline. As the comparison between the #2 and #3 rows in Table II, multi-task learning improves the baseline method in most of adaptation cases and obtains an improvement of 1.27% in average DSC. As observed from the visualization results (Fig. II), multi-task learning, to a certain extent, can rectify the fluid segmentation result. These experiments indicate that learning layer prediction can help the mode to learn a representation that is more suitable for fluid segmentation.

Since our proposed cross-attention learning and structure-aware adversarial are designed based on multi-task learning, in the following, we study their effectiveness with multi-task learning.

Effect of Cross-Attention Learning. We further add the proposed cross-attention learning into the method with multi-task learning and perform the prediction on the second feature-level but without adversarial learning of the second feature-level. As reported in #3 and #4 rows in Table II, cross-attention learning improves the average DSC from 29.04% to 29.49%. It should be noticed that the cross-attention learning significantly improves the adaptation case of C→T, from mean DSC 5.92% to 15.55%. The visualization results in Fig. II also demonstrate that the cross-attention learning can clearly reduce the errors in the case of C→T. These experiments show that leveraging the co-relation between layer structure and fluid can further activate the benefit of the retinal layer structure in improving the fluid segmentation.

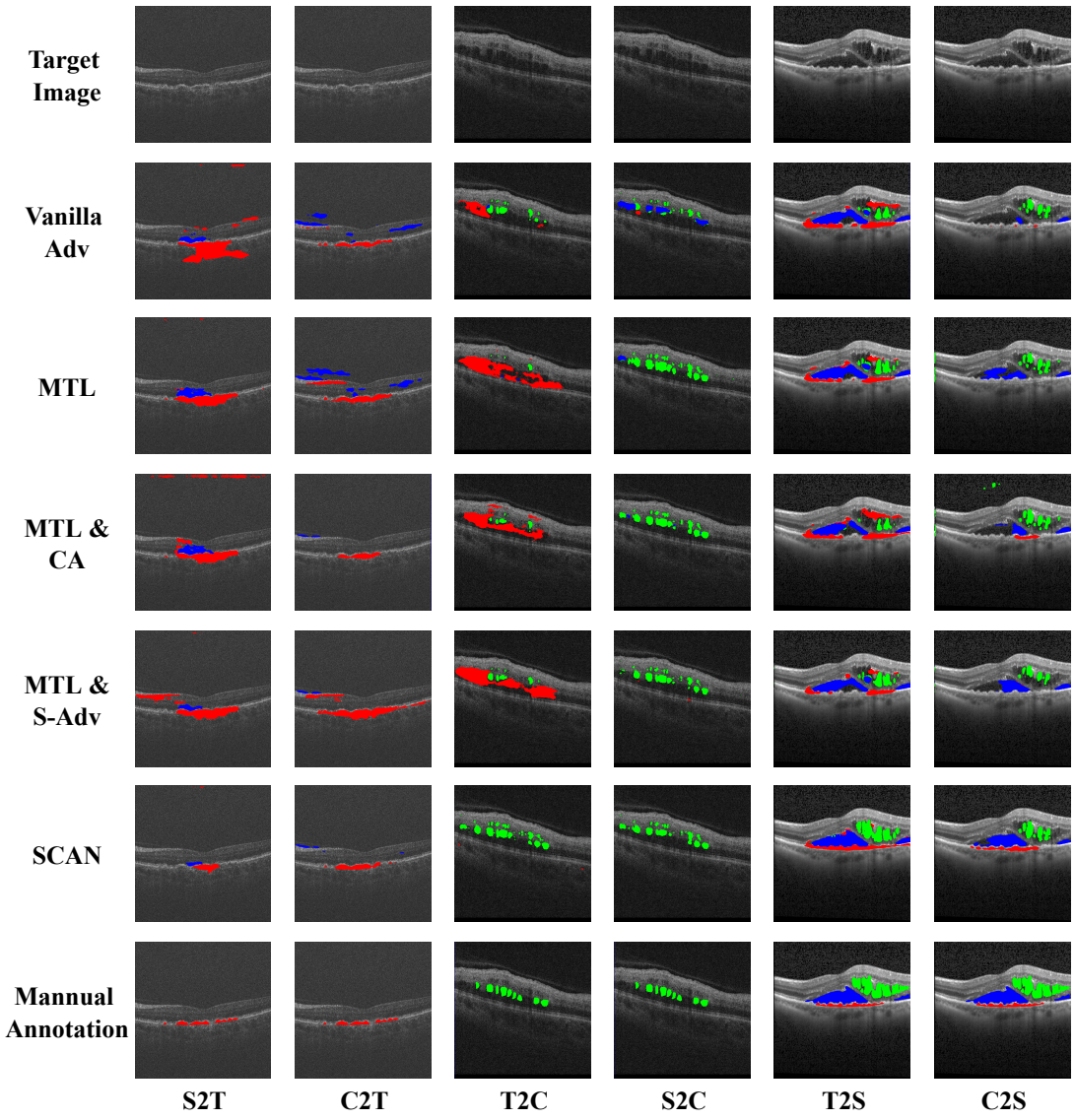


Fig. 4. Visualization Examples of ablation experimental results. Each Column exhibits a condition of adaptation (*e.g.*, S2T denotes adapting from the Spectralis domain to the Topcon domain), and each row exhibits visualized result of different compared methods.

TABLE III
SEGMENTATION RESULTS COMPARED WITH STATE-OF-THE-ART METHODS. "T", "C", AND "S" PRESENT TOPCON, CIRRUS, AND SPECTRALIS DOMAIN OF OCT IMAGES. WE REPORT THE MEAN ± STANDARD DEVIATION DSC SCORE OF 5-FOLD CROSS VALIDATION. THE BEST RESULT OF MEAN DSC SCORE IN EACH COLUMN IS HIGHLIGHT IN **BOLD** (UNIT: %)

Method	S→T	C→T	T→C	S→C	T→S	C→S	Average
Source Training	16.79 ± 4.13	10.02 ± 8.96	9.70 ± 4.81	7.27 ± 5.45	16.31 ± 2.54	22.09 ± 4.82	13.70
ADSegNet [27]	22.57 ± 1.21	5.58 ± 3.33	22.57 ± 7.57	**31.15** ± 0.86	46.87 ± 4.04	**42.38** ± 4.49	28.52
CLAN [52]	32.44 ± 6.20	15.62 ± 10.60	26.70 ± 2.44	23.49 ± 4.00	41.15 ± 2.79	33.89 ± 7.36	28.88
ADVENT [28]	29.22 ± 0.38	13.88 ± 5.45	25.26 ± 0.60	29.86 ± 0.83	45.24 ± 4.58	34.90 ± 0.30	29.73
SCAN (ours)	**39.63** ± 1.30	**17.25** ± 4.44	**30.61** ± 1.59	24.88 ± 1.55	**49.76** ± 2.35	42.11 ± 1.22	**34.04**
Target Training (Oracle)	56.77	56.77	45.89	45.89	57.54	57.54	53.40

Effect of Structure-Aware Adversarial Learning. Finally, we study the impact of the proposed structure-aware adversarial learning. We implement it in two ways, (1) directly adding it into the multi-task learning model; (2) adding it into the model with multi-task learning and cross-attention learning. In the first manner, we obtain a slight improvement for the average DSC score. In the second manner, the average DSC score is significantly improved. Specifically, compared with the model trained with multi-task learning and cross-attention learning, adding structure-aware adversarial learning improves the average DSC from 29.95% to 34.04%. These results demonstrate the effectiveness of structure-aware adversarial learning. In addition, the cross-attention learning is an essential step in our full method. As the visualized results

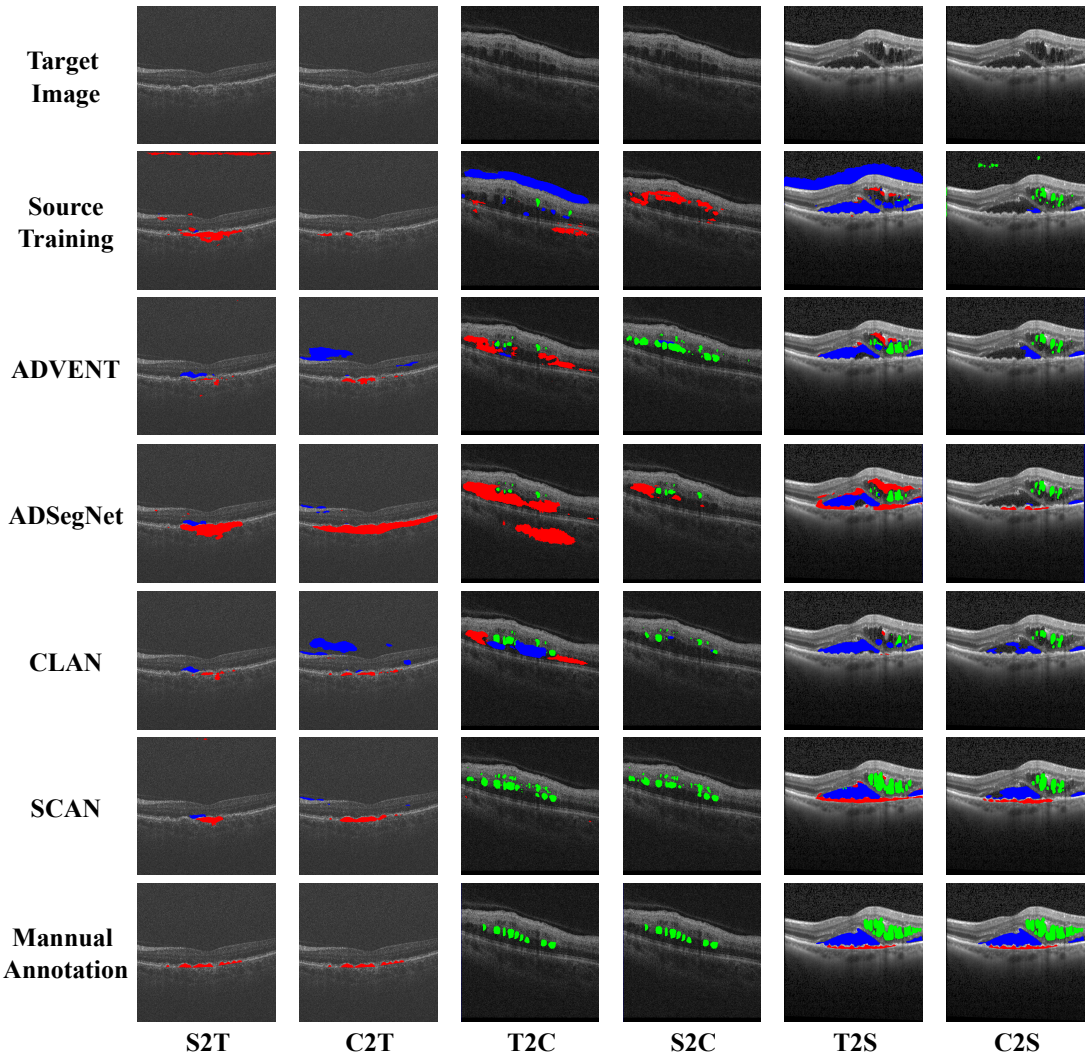


Fig. 5. Visualized results of comparative methods. Each column exhibits a condition of adaptation (e.g., S2T denotes adapting from the Spectralis domain to the Topcon domain), and each row exhibits visualized result of different compared methods.

(Fig. II), our full model greatly reduces the prediction errors, and can correctly locate the location as well as the accurate boundaries and shapes of fluids. When using Spectralis as the source domain, the cross-attention module combined with the structure-aware adversarial learning leads to a significant improvement on S→T from mean DSC 24.81% to 39.63% (see TABLE II). However, it impairs the adaptation direction of S→C from mean DSC 28.58% to 24.88%. We can find an explanation from the visualized results. On S→C , structure-aware adversarial learning without the cross-attention module makes incorrect class predictions (see Fig.4, the 4th row). After adding the cross-attention module, our SCAN eliminates the errors and generates some predictions extending to unlabeled regions. Those regions, however, are highly possible fluid regions but might be ignored due to unclear imaging. Therefore, we demonstrate that our SCAN method is more robust, producing consistent predictions across domains and datasets. In addition, our SCAN method achieves the most stable performance over cross-validation with a standard deviation of 1.55% on S→C (see TABLE II). That also demonstrates the benefits of our SCAN method, making the adapted model more robust across the training/testing data variance.

C. Comparisons with State-of-the-Art

We compare our method with three popular UDA methods in semantic segmentation on the RETOUCH dataset, including ADSegNet [27], ADVENT [28], and CLAN [52]. All three methods are constructed based on the adversarial learning algorithm. Specifically, ADSegNet presents a multi-level framework, which applies adversarial learning on both the feature space and the output space. ADVENT [28], and CLAN [52] are both inspired by this work. ADVENT incorporates an entropy-minimizing loss with the output space adversarial learning. CLAN utilizes two separated classifiers to obtain a local alignment score map, which is applied with the output space adversarial training. We reproduce these three methods with the official source codes. For a fair comparison, all methods use the UNet [4] as the backbone. In addition to the results of these three methods and our SCAN, we also

TABLE IV
COMPARISON OF DIFFERENT LAYER MAP GENERATION METHODS.

Method	Average DSC
SCAN-OTSU	34.04
SCAN-ReLayNet	29.30

show the results of source training and target training. The model with target training is learned with the labeled target training data, which can be regarded as the upper bound of each direction.

Result. The comparisons are reported in Table III. We make the following observations. First, there is a large gap between the source training model and the target training model. For example, the difference in average DSC is about 40% between source training and target training models. This indicates the significance of addressing the domain shift in fluid segmentation. Second, all the four domain adaptation methods (ADSegNet, ADVENT, CLAN, and our SCAN) significantly improve the performance of the source training model, demonstrating the effectiveness of adversarial learning in domain adaptation. Specifically, ADSegNet, ADVENT, CLAN and our SCAN improve the average DSC score by 14.82%, 16.03%, 15.18%, 20.34%, respectively. Third, our SCAN significantly outperforms the three compared methods. For example, our SCAN produces a higher average DSC score than ADSegNet, ADVENT and CLAN by 5.52%, 4.31% and 5.16%, respectively. This verifies the advantage of learning with the retinal structure.

Visualization. To better understand the benefit of our SCAN, we compare the visualization results of different methods in Fig. 5. We can find that the source training model is difficult to locate the correct position of fluids on the target OCT domain or predict the category and shape of the fluid. The three compared domain adaptation methods clearly improve fluid segmentation but still generate serious errors. For example, predicting a fluid region that is out of the retinal layers or that has an incorrect fluid shape. One reason for such failure is that these domain adaptation methods do not consider the specific structure of medical images. In contrast, benefiting from the retinal layer structure, our SCAN significantly enhances fluid segmentation and provides practical quantitative analysis advice with precious fluid boundaries and shapes. This further demonstrates the effectiveness of the proposed SCAN.

D. Comparison with Deep Learning based Layer Segmentation Method

In our method, we use a simple off-the-shelf segmentation method (OTSU [35]) to produce the rough retinal layer map. One may ask how about replacing it with a deep learning based layer segmentation method. To answer this question, we adopt the well-known ReLayNet [7] to predict the layer maps. Due to the lack of layer annotations, it is not possible to train the ReLayNet from scratch on the RETOUCH dataset. Instead, we use the publicly released ReLayNet model, which is pretrained on the Duke OCT dataset [59]. The comparison of generating layer maps with two different methods is reported in Table IV. We can find that SCAN w/ OTSU significantly outperforms the SCAN w/ ReLayNet. The main failure reason of SCAN w/ ReLayNet is that ReLayNet is trained on another dataset and also suffers from the domain shift problem. As such, ReLayNet produces bad layer maps for the RETOUCH dataset and thus fails to provide accurate supervision during domain adaptation. In contrast, OTSU is robust to different OCT domains and produces adequate layer maps on RETOUCH. The comparison of ReLayNet and OTSU is shown in Fig. 6.

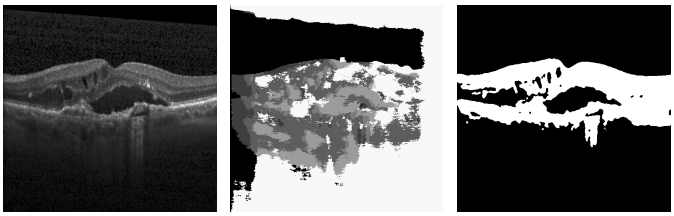


Fig. 6. Layer maps of different methods. Left: input image, Middle: ReLayNet, Right: OTSU.

V. CONCLUSION

In this paper, we studied a challenging yet practical problem in fluid segmentation, named cross-domain fluid segmentation. To solve this problem, we proposed a Structure-guided Cross-Attention Network (SCAN), which explicitly exploits the retinal layer structure to facilitate domain alignment. Specifically, we built a multi-task learning framework to jointly learn layer prediction and fluid segmentation. To explicitly leverage the interaction between the layer and fluid, a cross-attention module was proposed to extract co-attention features, which was used to learn the fluid predictor and layer predictor. Moreover, a structure-aware adversarial learning approach was introduced to guide adversarial learning focusing on the region with large discrepancies. Experiments on the RETOUCH dataset showed the advantage of the proposed method. Our SCAN also clearly outperformed the popular domain adaptation methods by a large margin. In this paper, we investigated the feasibility of using "structure knowledge" to assist the OCT fluid segmentation task. However, we believe that leveraging structure knowledge does not limit to this specific task and has the potential to apply to other medical imaging tasks (such as prostate segmentation or lesions segmentation in the brain) and other modalities (such as CT or MRI). We hope this study can motivate more researchers to address medical imaging tasks in the perspective of "structure knowledge".

REFERENCES

[1] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, C. A. Puliafito *et al.*, "Optical coherence tomography," *Science*, vol. 254, no. 5035, pp. 1178–1181, 1991.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[3] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. intervent.* Springer, 2015, pp. 234–241.

[5] G. Trichonas and P. K. Kaiser, "Optical coherence tomography imaging of macular oedema," *Br. J. Ophthalmol.*, vol. 98, no. Suppl 2, pp. 24–29, 2014.

[6] R. Varma, N. M. Bressler, Q. V. Doan, M. Gleeson, M. Danese, J. K. Bower, E. Selvin, C. Dolan, J. Fine, S. Colman *et al.*, "Prevalence of and risk factors for diabetic macular edema in the united states," *JAMA Ophthalmol.*, vol. 132, no. 11, pp. 1334–1340, 2014.

[7] A. G. Roy, S. Conjeti, S. P. K. Karri, D. Sheet, A. Katouzian, C. Wachinger, and N. Navab, "Relaynet: Retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks," *Biomed. Opt. Express*, vol. 8, no. 8, pp. 3627–3642, 2017.

[8] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nat. Med.*, vol. 24, no. 9, pp. 1342–1350, 2018.

[9] J. Hu, Y. Chen, and Z. Yi, "Automated segmentation of macular edema in oct using deep neural networks," *Med. Image Anal.*, vol. 55, pp. 216–227, 2019.

[10] L. Ngo, J. Cha, and J.-H. Han, "Deep neural network regression for automated retinal layer segmentation in optical coherence tomography images," *IEEE Trans. Image Process.*, vol. 29, pp. 303–312, 2019.

[11] A. Montuoro, S. M. Waldstein, B. S. Gerendas, U. Schmidt-Erfurth, and H. Bogunović, "Joint retinal layer and fluid segmentation in oct scans of eyes with severe macular edema using unsupervised representation and auto-context," *Biomed. Opt. Express*, vol. 8, no. 3, pp. 1874–1888, 2017.

[12] J. Guo, W. Zhu, F. Shi, D. Xiang, H. Chen, and X. Chen, "A framework for classification and segmentation of branch retinal artery occlusion in sd-oct," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3518–3527, 2017.

[13] D. Xiang, H. Tian, X. Yang, F. Shi, W. Zhu, H. Chen, and X. Chen, "Automatic segmentation of retinal layer in oct images with choroidal neovascularization," *IEEE Trans. Image Process.*, vol. 27, no. 12, pp. 5880–5891, 2018.

[14] B. Gong, K. Grauman, and F. Sha, "Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation," in *Int. Conf. Mach. Learn. ICML*. PMLR, 2013, pp. 222–230.

[15] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Int. Conf. Mach. Learn. ICML*. PMLR, 2015, pp. 1180–1189.

[16] M. Luo, X. Chang, L. Nie, Y. Yang, A. G. Hauptmann, and Q. Zheng, "An adaptive semisupervised feature analysis for video semantic recognition," *IEEE Trans. Cybern.*, vol. 48, no. 2, pp. 648–660, 2017.

[17] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[18] D. Zhang, L. Yao, K. Chen, S. Wang, X. Chang, and Y. Liu, "Making sense of spatio-temporal preserving representations for eeg-based human intention recognition," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3033–3044, 2019.

[19] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, 2016.

[20] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7167–7176.

[21] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Adv. Neural Inf. Process. Syst.*, vol. 27. Curran Associates, Inc., 2014.

[23] Y.-C. Chen, Y.-Y. Lin, M.-H. Yang, and J.-B. Huang, "Crdoco: Pixel-level domain transfer with cross-domain consistency," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1791–1800.

[24] C. Yang, T. Kim, R. Wang, H. Peng, and C.-C. J. Kuo, "Show, attend, and translate: Unsupervised image translation with self-regularization and attention," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 4845–4856, 2019.

[25] X. Guo, C. Yang, B. Li, and Y. Yuan, "Metacorrection: Domain-aware meta loss correction for unsupervised domain adaptation in semantic segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, June 2021, pp. 3927–3936.

[26] H. Tian, S. Qu, and P. Payeur, "A prototypical knowledge oriented adaptation framework for semantic segmentation," *IEEE Trans. Image Process.*, vol. 31, pp. 149–163, 2021.

[27] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7472–7481.

[28] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2517–2526.

[29] J. Huang, S. Lu, D. Guan, and X. Zhang, "Contextual-relation consistent domain adaptation for semantic segmentation," in *European conference on computer vision*. Springer, 2020, pp. 705–722.

[30] W. Zhou, Y. Wang, J. Chu, J. Yang, X. Bai, and Y. Xu, "Affinity space adaptation for semantic segmentation across domains," *IEEE Trans. Image Process.*, vol. 30, pp. 2549–2561, 2020.

[31] Y.-H. Tsai, K. Sohn, S. Schulter, and M. Chandraker, "Domain adaptation for structured output via discriminative patch representations," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1456–1465.

[32] F. Yu, M. Zhang, H. Dong, S. Hu, B. Dong, and L. Zhang, "Dast: Unsupervised domain adaptation in semantic segmentation based on discriminator attention and self-training," in *Proc. Innov. Appl. Artif. Intell. Conf.*, vol. 35, 2021, p. 10, issue: 12.

[33] C. Chen, W. Xie, Y. Wen, Y. Huang, and X. Ding, "Multiple-source domain adaptation with generative adversarial nets," *Knowl. Based Syst.*, vol. 199, p. 105962, 2020.

[34] A. Montuoro, S. M. Waldstein, B. S. Gerendas, U. Schmidt-Erfurth, and H. Bogunović, "Joint retinal layer and fluid segmentation in oct scans of eyes with severe macular edema using unsupervised representation and auto-context," *Biomed. Opt. Express*, vol. 8, no. 3, pp. 1874–1888, 2017.

[35] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man Cybern. Syst*, vol. 9, no. 1, pp. 62–66, 1979.

[36] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. intervent*. Springer, 2018, pp. 3–11.

[37] C. S. Lee, A. J. Tyring, N. P. Deruyter, Y. Wu, A. Rokem, and A. Y. Lee, "Deep-learning based, automated segmentation of macular edema in optical coherence tomography," *Biomed. Opt. Express*, vol. 8, no. 7, pp. 3440–3448, 2017.

[38] T. Schlegl, S. M. Waldstein, H. Bogunovic, F. Endstraßer, A. Sadeghipour, A.-M. Philip, D. Podkowinski, B. S. Gerendas, G. Langs, and U. Schmidt-Erfurth, "Fully automated detection and quantification of macular fluid in oct using deep learning," *Ophthalmology*, vol. 125, no. 4, pp. 549–558, 2018.

[39] A. Rashno, D. D. Koozekanani, and K. K. Parhi, "Oct fluid segmentation using graph shortest path and convolutional neural network," in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2018, pp. 3426–3429.

[40] Y. Xu, K. Yan, J. Kim, X. Wang, C. Li, L. Su, S. Yu, X. Xu, and D. D. Feng, "Dual-stage deep learning framework for pigment epithelium detachment segmentation in polypoidal choroidal vasculopathy," *Biomed. Opt. Express*, vol. 8, no. 9, pp. 4061–4076, 2017.

[41] T. Hassan, M. U. Akram, M. F. Masood, and U. Yasin, "Deep structure tensor graph search framework for automated extraction and characterization of retinal layers and fluid pathology in retinal sd-oct scans," *Comput. Biol. Med.*, vol. 105, pp. 112–124, 2019.

[42] D. Lu, M. Heisler, S. Lee, G. W. Ding, E. Navajas, M. V. Sarunic, and M. F. Beg, "Deep-learning based multiclass retinal fluid segmentation and detection in optical coherence tomography images using a fully convolutional neural network," *Med. Image Anal.*, vol. 54, pp. 100–110, 2019.

[43] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "Fcns in the wild: Pixel-level adversarial and constraint-based adaptation," *arXiv preprint arXiv:1612.02649*, 2016.

[44] G. Csurka, R. Volpi, and B. Chidlovskii, "Unsupervised domain adaptation for semantic image segmentation: a comprehensive survey," *arXiv preprint arXiv:2112.03241*, 2021.

[45] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *Int. Conf. Mach. Learn. ICML*. PMLR, 2018, pp. 1989–1998.

[46] P. Li, X. Liang, D. Jia, and E. P. Xing, "Semantic-aware grad-gan for virtual-to-real urban scene adaption," *arXiv preprint arXiv:1801.01726*, 2018.

[47] K. Wang, C. Yang, and M. Betke, "Consistency regularization with high-dimensional nonadversarial source-guided perturbation for unsupervised

domain adaptation in segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 11, 2021, pp. 10 138–10 146.
[48] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6936–6945.
[49] Y. Luo, P. Liu, T. Guan, J. Yu, and Y. Yang, "Significance-aware information bottleneck for domain adaptive semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis*, 2019, pp. 6778–6787.
[50] Y. Zhang, Z. Qiu, T. Yao, C.-W. Ngo, D. Liu, and T. Mei, "Transferring and regularizing prediction for semantic segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9621–9630.
[51] Z. Wang, M. Yu, Y. Wei, R. Feris, J. Xiong, W.-m. Hwu, T. S. Huang, and H. Shi, "Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12 635–12 644.
[52] Y. Luo, P. Liu, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Category-level adversarial adaptation for semantic segmentation using purified features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 3940–3956, 2021.
[53] X. Xie, J. Niu, X. Liu, Z. Chen, S. Tang, and S. Yu, "A survey on incorporating domain knowledge into deep learning for medical image analysis," *Medical Image Analysis*, vol. 69, p. 101985, 2021.
[54] L. Sun, C. Li, X. Ding, Y. Huang, Z. Chen, G. Wang, Y. Yu, and J. Paisley, "Few-shot medical image segmentation using a global correlation network with discriminative embedding," *Computers in biology and medicine*, vol. 140, p. 105067, 2022.
[55] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.
[56] H. Bogunović, F. Venhuizen, S. Klimscha, S. Apostolopoulos, A. Bab-Hadiashar, U. Bagci, M. F. Beg, L. Bekalo, Q. Chen, C. Ciller *et al.*, "Retouch: The retinal oct fluid detection and segmentation benchmark and challenge," *IEEE Trans. Med. Imaging*, vol. 38, no. 8, pp. 1858–1874, 2019.
[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[58] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS 2017 Workshop on Autodiff*, 2017.
[59] S. J. Chiu, M. J. Allingham, P. S. Mettu, S. W. Cousins, J. A. Izatt, and S. Farsiu, "Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema," *Biomed. Opt. Express*, vol. 6, no. 4, pp. 1172–1194, 2015.