Intra- and Inter-Slice Contrastive Learning for Point Supervised OCT Fluid Segmentation

Xingxin He, Leyuan Fang, *Senior Member, IEEE*, Mingkui Tan, *Member, IEEE*, and Xiangdong Chen

***Abstract*— OCT fluid segmentation is a crucial task for diagnosis and therapy in ophthalmology. The current convolutional neural networks (CNNs) supervised by pixel-wise annotated masks achieve great success in OCT fluid segmentation. However, requiring pixel-wise masks from OCT images is time-consuming, expensive and expertise needed. This paper proposes an Intra- and inter-Slice Contrastive Learning Network (ISCLNet) for OCT fluid segmentation with only point supervision. Our ISCLNet learns visual representation by designing contrastive tasks that exploit the inherent similarity or dissimilarity from unlabeled OCT data. Specifically, we propose an intra-slice contrastive learning strategy to leverage the fluid-background similarity and the retinal layer-background dissimilarity. Moreover, we construct an inter-slice contrastive learning architecture to learn the similarity of adjacent OCT slices from one OCT volume. Finally, an end-to-end model combining intra- and inter-slice contrastive learning processes learns to segment fluid under the point supervision. The experimental results on two public OCT fluid segmentation datasets (i.e., AI Challenger and RETOUCH) demonstrate that the ISCLNet bridges the gap between fully-supervised and weakly-supervised OCT fluid segmentation and outperforms other well-known point-supervised segmentation methods.**

***Index Terms*— Optical coherence tomography, fluid segmentation, convolutional neural network, weakly-supervised learning, contrastive learning.**

I. Introduction

THE macula is the area of the retina responsible for central vision. Macular edema caused by fluid accumulation is a common reason for visual impairment [1]. Frequent causes of macular edema include age-related macular degeneration and diabetic retinopathy [2]. Diagnosis and therapy of these retinal diseases depend on the accurate segmentation and quantitative analysis of macular fluid.

Optical coherence tomography (OCT) is a non-invasive and rapid retinal imaging technique [3], [4] that generates three dimensions cross-sectional images with high resolution. OCT has become a standard tool for accurate segmentation and quantitative analysis of macular fluid in the clinic [5]. However, the manual analysis of fluid is subjective, labor-intensive, and prone to errors. Therefore, computer-aided systems have been proposed to assist ophthalmologists in segmenting retinal fluid. Early OCT segmentation systems with conventional segmentation methods include threshold-based [6], graph-based [7], [8], or machine learning approaches [9] that use hand-crafted features. However, those conventional methods are sensitive to image quality, requiring domain knowledge, and lacking generalization.

Unlike conventional methods that use elaborate hand-crafted features, the convolutional neural network (CNN) automatically learns to extract image features. Several CNN-based methods have been proposed to perform segmentation tasks, e.g., FCN [10], UNet [11], SegNet [12], and DeepLab [13]. Also, some studies have been conducted to segment OCT fluid based on CNN in an end-to-end manner [14]–[16] and achieved great improvement compared with conventional methods. However, the success of CNN-based segmentation models heavily rely on large-scale training datasets with fully-annotated masks that are time-consuming and expensive to obtain.

Recently, weakly-supervised segmentation methods have been studied to reduce the human effort in training CNN-based semantic segmentation models [17]–[23]. Those works train a CNN-based segmentation model with various forms of weak supervision, e.g., tags of which object appearing in an image [17], [18], bounding boxes for each object [19], [20], and partial mask annotation, such as points or scribbles [21]–[23]. The idea of weakly-supervised segmentation also has been applied in OCT segmentation [24] with tag supervision to segment geographic atrophy lesions. Since the tags and bounding boxes cannot directly provide dense masks with accurate location information of objects, the typical tag and bounding box supervised segmentation model usually consider generating pseudo masks (proposals) and then train a regular segmentation model with the proposals. However, such proposal-based methods may introduce two sources of errors: 1. the error of generating proposals; 2. the error of learning a segmentation model with these proposals. Unlike the tags

Manuscript received May 27, 2021; revised October 10, 2021, January 9, 2022, and January 26, 2022; accepted January 26, 2022. Date of publication February 9, 2022; date of current version February 15, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61922029 and Grant 61771192, in part by the Science and Technology Plan Project Fund of Hunan Province under Grant 2019RS2016, and in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2018B010107001. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Abdesselam Bouzerdoum. *(Corresponding author: Leyuan Fang.)*

Xingxin He and Leyuan Fang are with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China (e-mail: xx_h@hnu.edu.cn; fangleyuan@gmail.com).

Mingkui Tan is with the School of Software Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: mingkuitan@scut.edu.cn).

Xiangdong Chen is with the Department of Ophthalmology, The First Affiliated Hospital of Hunan University of Chinese Medicine, Changsha 410082, China (e-mail: 564259166@qq.com).

Digital Object Identifier 10.1109/TIP.2022.3148814

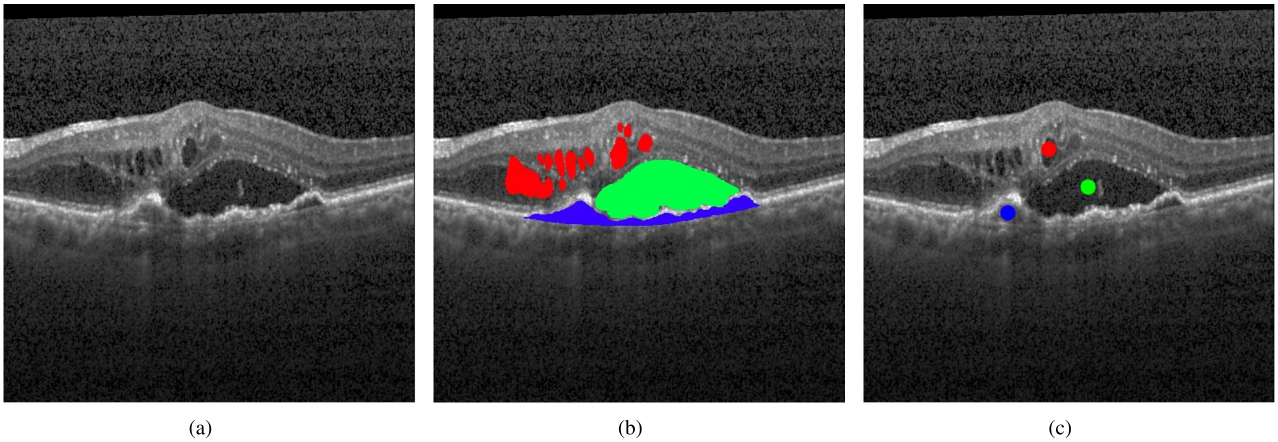


Fig. 1. Examples of retinal fluids in OCT slices with manual annotations. (a) The original OCT slice with fluids; (b) The full mask annotation; (c) The point annotations. In (b) and (c), the red, green, and blue contours denote IRF, SRF, and PED, respectively. The point annotations are enlarged in this figure for better visualization.

and bounding boxes, partial labels can provide a few reliable groundtruth. Learning directly from partially annotated labels can avoid the error of generating proposals. Besides, the point label only needs a single click of each class in the image, which is easy to obtain as tags in OCT fluid segmentation, where only a few fluid types appeared in an OCT slice. Therefore, we choose point annotations as supervision for our retinal OCT fluid segmentation to minimize the human effort. Fig. 1 shows examples of B-scan slices, representing the appearance of three types of fluid, i.e., intraretinal fluid (IRF), subretinal fluid (SRF), and pigment epithelial detachment (PED) with the manually segmented annotations and the point labels.

The point annotation cannot provide sufficient information about the fluid for training. Compared with full masks, two folds of information are missing in point labels: 1. The point annotation only lies in a tiny region inside the fluid region without any extension, shape, or boundary information of the fluid; 2. A complete OCT volume is composed of a series of B-scan slices [25]. The fluid region is spatial continuous between adjacent OCT slices, but point annotation hardly provides inter-slice continuity information.

To exploit the fluid-related information from the OCT image itself, we employ the contrastive learning paradigm, which discovers visual features through learning the similarity and dissimilarity of unlabeled data [26]. Within an OCT slice, it displays retinal tissues and lesions through different optical reflections. The background region above the retinal layers is the vitreous body that contains 98%-99% of vitreous fluid with little solid matter [27]. Thus, the pathological fluid regions show high similarity to the background with low reflection in OCT images. In contrast, retinal layers have a relatively high reflection that is distinct from fluid and background (Fig. 2). Therefore, we conduct an intra-slice contrastive learning strategy to learn the intrinsic fluid-background similarity and retinal layer-background dissimilarity of an OCT slice. Besides, we construct an inter-slice contrastive learning architecture to learn the consistency between nearby slices. As a result, an Intra- and Inter-Slice Contrastive Learning Network (ISCLNet) is composed for retinal OCT fluid segmentation supervised by point annotations. The experimental results on two public OCT fluid segmentation datasets demonstrate that our proposed ISCLNet is an effective weakly-supervised retinal OCT fluid segmentation method that narrows the gap between weakly-supervised segmentation and fully-supervised segmentation, and outperforms other well-known weakly-supervised segmentation methods.

In summary, our contributions are as follows:

1. We propose a novel and effective point-supervised retinal OCT fluid segmentation method, namely Intra- and inter-Slice Contrastive Learning Network (ISCLNet) with point annotations.

2. We propose an intra-slice contrastive learning strategy to leverage the intrinsic local similarity and dissimilarity within an OCT slice.

3. We construct an inter-slice contrastive learning architecture to learn the inter-slice continuity of adjacent OCT slices.

The rest of the paper is organized as follows: Section II introduces the related works, including CNN-based methods for OCT fluid segmentation, weakly-supervised segmentation methods, and contrastive learning methods for weakly-supervised segmentation. Section III describes the proposed methodology, including inter-slice contrastive learning strategy, inter-slice contrastive learning architecture, and the entire end-to-end framework. Section IV exhibits the experimental results. Sections V and VI present the discussion, conclusion, and possible future works.

II. Related Works

A. CNN-Based OCT Fluid Segmentation

Many CNN-based segmentation methods have been successfully applied to OCT fluid segmentation. Most current OCT fluid segmentation methods are based on UNet [11]. Specifically, studies in [28] and [29] employed the UNet to achieve macular segmentation. To improve the performance of UNet for segmenting OCT fluid, Rashno *et al.* [30] implemented a graph shortest path technique as post-processing to refine the predictive results. Tennakoon *et al.* [31] proposed an adversarial training strategy to regularize the segmented results. Considering the structural relationship between retinal layers and fluids, Xu *et al.* [32] designed a two-stage fluid segmentation framework. They firstly trained a retinal layer segmentation network to extract retinal layer maps. Then, they utilized the layer maps as the constrain to train a UNet in the second stage. Similarly, other studies [33], [34] conducted a graph-cut-based method to obtain the retinal layers segmentation maps and then combine the maps to train a UNet. To further improve the ability of the OCT segmentation method, a study in [15] 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 [35], Mahapatra *et al.* introduced a GAN-based method to generate pseudo OCT images as data augmentation to improve OCT retinal fluid segmentation. In [36], Bekalo *et al.* adopted an encoder-decoder segmentation architecture combined with the skip-connect operation and atrous spatial pyramid pooling (ASPP) module [37] to improve retinal fluid segmentation. Besides, another work in [38] also introduced the ASPP module to a UNet-like segmentation method with residual and inception modules.

The excellent performance of current OCT fluid segmentation methods heavily relies on the dataset with pixel-wise annotations. However, it is hugely time-consuming and expensive to obtain such labels.

B. Weakly-Supervised Segmentation

Since weak supervisions cannot provide dense location information about the object, a popular learning strategy of weakly-supervised segmentation methods is to generate pseudo masks from the weak labels and then train a regular segmentation model with generated pseudo masks [22], [39], [40]. Some works employed standard segmentation techniques (e.g., graph-cut [22] or random-walk [39]) to access the pseudo masks. Pu *et al.* [40] trained a super-pixel-based graph neural network to obtain the pseudo masks from points or scribbles. However, the pseudo-label-based weakly-supervised segmentation method may produce two folds of error: 1. the error of pseudo labels generating labels; 2. the error of learning segmentation with the pseudo labels.

To avoid errors in proposal generation, another learning strategy for weakly-supervised segmentation trains the segmentation model directly on partial annotations [21], [23], [41]–[47]. Some trained segmentation models with additional parameters and tasks to enhance the CNN model [23]. The others used prior knowledge in the form of constraints to guide the segmentation during training [41]–[44]. Prior knowledge is beneficial for medical segmentation problems, where information about the target region is often known beforehand. The existing weakly-supervised methods have demonstrated the possibility of reducing manual efforts while enhancing the segmentation, but their adaptation in OCT fluid segmentation has not been explored.

C. Self-Supervised Learning for Medical Image Analysis

Self-supervised learning exploits representations from unlabeled data via self-supervised learning tasks. Therefore, self-supervised learning can boost medical image analysis with limited manual labels. The core problem of self-supervised learning is designing self-supervised tasks.

Jigsaw puzzle [48] pretext models are widely used for learning representation from unlabeled medical images [49]–[52]. To handle the 3D medical images (e.g., CT and MRI), Zhuang *et al.* [53] proposed a 3D version of the jigsaw puzzle task, named Rubik's cube recovery task, to extract representations from unlabeled 3D CT volumes. Similarly, Rubik's Cube+ [54] and Rubik's Cube++ [55] methods further improve the Rubik's cube recovery pretext task.

Besides, contrastive learning, as a special set of self-supervised learning, is another promising direction in medical image analysis, which exploits representations via learning the similarity and dissimilarity of unlabeled data. In [56], Jamaludin *et al.* adopted a siamese CNN with a contrastive loss function to recognize whether a pair of MRI scans belong to the same patient. In [57], Zhang *et al.* utilized two CNNs to encode features of CT images and textual reports, respectively. Then a contrastive loss learns to match the paired image and medical report.

In [58], Lei *et al.* leveraged the anatomical structure by predicting the relative position regression between any two patches from a CT volume. Chaitanya *et al.* [59] proposed a global contrastive learning strategy to extract the similarity across MRI volumes and a local contrastive learning strategy to exploit distinctive representations of local regions. Nguyen *et al.* [60] proposed two self-supervised learning tasks. One task aims to predict the spatial location of a patch in a CT scan. Another task aims to predict the order of slices in a CT volume. In general, the above methods do not consider the detailed biological structure inside medical images. Thus, those methods are challenging to implement to the specific OCT fluid segmentation problem.

D. Contrastive Learning for Weakly-Supervised Segmentation

The general idea of contrastive learning is to exploit representative features by learning the intrinsic similarity and dissimilarity of the unlabeled data [61]. Some works have employed contrastive learning for weakly-supervised semantic segmentation. For example, Qian *et al.* designed a contrastive learning process [45] to minimize the distance between features with the same class and maximize the distance between features with different classes to improve segmentation with point supervision. Araslanov *et al.* [62] proposed a contrastive learning strategy to tackle local inconsistency, semantic inaccuracy, and incompleteness of CAM [63]. Wang *et al.* proposed a contrastive learning task to force the scale equivariant of CAM [64] to enhance the pseudo label generation. Inspired by the contrastive learning and considering the biological characters of OCT images, we propose to learn the inherent intra- and inter-slice similarity and dissimilarity from unlabeled OCT images to bridge the gap between point-supervised OCT fluid segmentation and fully-supervised OCT fluid segmentation.

III. Proposed Methods

We propose to learn visual features through intra- and inter-slice contrastive learning from unlabeled OCT images

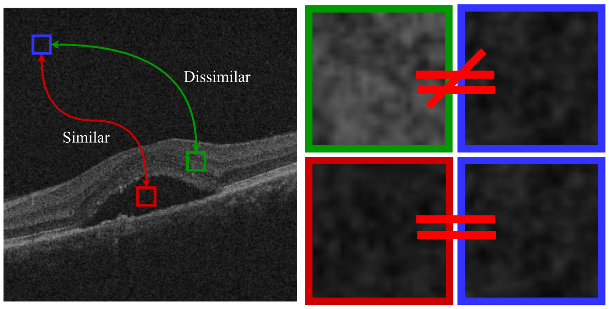


Fig. 2. The intra-slice similarity and dissimilarity.

to bridge the gap between fully-supervised and point-supervised retinal OCT fluid segmentation. Within an OCT slice, we propose an intra-slice contrastive learning strategy to learn the fluid-background similarity and retinal layer-background dissimilarity. Between OCT slices, we construct an inter-slice contrastive learning architecture to learn the inter-slice similarity that preserves the continuity between adjacent slices. Finally, an end-to-end framework including intra- and inter-slice contrastive learning is proposed to learn OCT fluid segmentation with point supervision. We detail each step of the proposed method in the following sections.

A. Problem Formulation

Given a training dataset $D = (\mathbf{x}_i^j, \mathbf{y}_i^j, \mathbf{p}_i^j)$, where $\mathbf{x}_i^j$ is the i^{th} slice of an OCT volume j, $\mathbf{y}_i^j$ denotes the full masks, and $\mathbf{p}_i^j$ denotes the corresponding point labels, we want to learn a segmentation model $f(.)$ with learnable parameters θ, such that $\mathbf{o}^{W\times H\times C} = f(\mathbf{x}_i^j|\theta) \in [0, 1]$ gives the label probabilities at each pixel for each predictive class C. The parameter updating of θ is learned from the point labels $\mathbf{p}_i^j$.

B. The Framework

The entire framework of ISCLNet is shown as Fig. 3. In the training phase, two OCT slices $\mathbf{x}_i^j$ and $\mathbf{x}_{i+t}^j$ from the same OCT volume j are inputted into two weight-shared segmentation backbones $f(.)$. The segmentation backbones output their segmentation predictive probability map $\mathbf{o}_i^j = f(\mathbf{x}_i^j)$ and $\mathbf{o}_{i+t}^j = f(\mathbf{x}_{i+t}^j)$, respectively. Then, we select the fluid region according to the probability map $\mathbf{o}$ by using a differentiable region selection operation (Equation 1, see details in Section III-C). Finally, intra-slice contrastive learning is performed with the selected fluid and background patches. Also, the similarity is enforced by learning the inter-slice similarity of both features and predictive probability maps. In the inference phase, since the two segmentation backbones are weight-shared, we use one of them to predict retinal fluid segmentation in OCT images.

C. Intra-Slice Contrastive Learning

Intra-slice contrastive learning aims to learn the fluid-background similarity and tissue-background dissimilarity within an OCT slice (Fig. 2). That is to construct a loss function $l_{intra}(\mathbf{x}_{flu}, \mathbf{x}_{lay}, \mathbf{x}_{bac})$, where $\mathbf{x}_{flu}, \mathbf{x}_{lay}, \mathbf{x}_{bac}$ denote the fluid region, the layer region, and the background region, respectively.

To evaluate the intra-slice term, we extract the fluid region and the tissue region according to the prediction from the segmentation model. For simplicity, we use $\mathbf{o}^{W\times H\times C} = f(\mathbf{x})$ to denote the output (i.e., the predictive probability map) of the segmentation model with a single OCT slice $\mathbf{x}$ as input where H and W denote the height and width of input images. C denotes predictive classes with $C = 0$ being the background and tissue (non-fluid) class, $C = 1, 2, 3$ being three different fluid types. Specifically, the fluid region extraction process can be expressed as:

$$\mathbf{x}_{flu} = \sum_{h=1}^{C} (softmax(\beta\mathbf{o}))_h \odot \mathbf{x} \tag{1}$$

Equation 1 indicates the probability of selecting a location on an OCT slice $\mathbf{x}$ as representing class h. $\beta \in (0, +\infty)$ controls how much uncertainty is permitted, i.e., as $\beta \to +\infty$, the values returned become close to either 0 or 1.

Furthermore, as we have extracted the fluid region $\mathbf{x}_{flu}$ from an OCT slice $\mathbf{x}$, the residual regions in $\mathbf{x}$ should be non-fluid regions (i.e., regions without fluid, which including retinal layered tissue and vitreous body). We denote these non-fluid regions as $\mathbf{x}_{laybac} = \mathbf{x} - \mathbf{x}_{flu}$

To obtain the layer region from non-fluid regions $\mathbf{x}_{laybac}$, we apply a simple Otsu segmentation method [65] to the input slice $\mathbf{x}$ to obtain a rough mask $\mathbf{m}_{lay} \in \{0, 1\}$ of the layer region. Therefore, the layer region can be expressed as $\mathbf{x}_{lay} = \mathbf{m}_{lay} \odot \mathbf{x}_{laybac}$.

Until now, we have the retinal fluid region $\mathbf{x}_{flu}$ and the retinal layer region $\mathbf{x}_{lay}$ containing intensities of the predictive fluid and tissue region. Then, we randomly take fluid blocks $\mathbf{b}_{flu}$ from the fluid region $\mathbf{x}_{flu}$ and layer blocks $\mathbf{b}_{lay}$ from the layer region $\mathbf{x}_{lay}$. As for the background block $\mathbf{b}_{lay}$, since the region above the retinal layer is the vitreous body filling with vitreous fluid, we randomly crop blocks from the top of an OCT slice as reference background blocks $\mathbf{b}_{bac}$.

Consequently, we compute the similarity between the histograms of contrastive blocks (i.e., $\mathbf{b}_{flu}$ and $\mathbf{b}_{lay}$) as:

$$l_{intra} = D_{KL}(\mathbf{b}_{flu}||\mathbf{b}_{bac}) - D_{KL}(\mathbf{b}_{lay}||\mathbf{b}_{bac}), \tag{2}$$

where D_{KL} computes the Kullback–Leibler divergence [66] of two distributions.

With the contrastive learning loss function l_{intra}, a CNN model can learn visual representation by maximizing the similarity between fluids and background region and minimizing the similarity between layers and background regions. The learned representation enables the segmentation model to distinguish the fluid and the retinal layer.

D. Inter-Slice Contrastive Learning

The idea of inter-slice contrastive learning is based on the assumption that if a CNN received similar samples as input, their representations should be similar. Thus, we construct a contrastive learning architecture based on a siamese neural network [67] to learn the inter-slice continuity. The proposed architecture is shown as Fig. 3. Two weights shared backbones

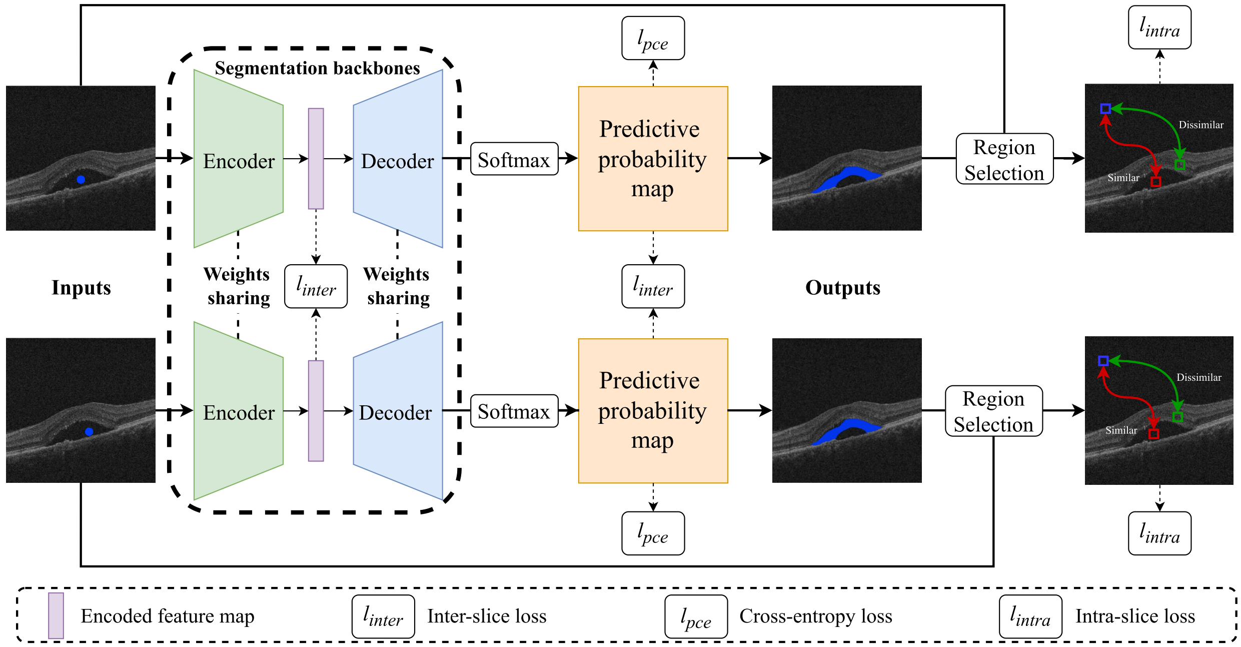


Fig. 3. Outline of our proposed ISCLNet method. Two OCT slices from the same OCT volume are inputted into two weight-shared segmentation backbones. The segmentation backbones output their segmentation predictive probability map, respectively. Then, we select the fluid region according to the predictive probability map by using a differentiable region selection operation (Equation 1). Finally, intra-slice contrastive learning is performed with the selected fluid and background patches. Also, the similarity is enforced by learning the inter-slice similarity of both features and predictive probability maps.

TABLE I

SEGMENTATION RESULTS COMPARED WITH STATE-OF-THE-ART METHODS ON AI CHALLENGER AND RETOUCH DATASET: MEAN (STANDARD DEVIATION) OF 10-FOLD CROSS-VALIDATION. THE BEST RESULTS IN THIS TABLE ARE LABELED IN BOLD. (UNIT: %)

Dataset	AI Challenger					RETOUCH				
Model	Label	DSC	IRF	PED	AUC	DSC	IRF	SRF	PED	AUC
Lower Baseline	point	59.72(2.40)	70.05(1.62)	49.38(3.17)	88.05(4.58)	39.41(2.14)	20.34(1.21)	57.40(4.09)	40.48(1.13)	95.06(0.53)
CRF loss [43]	point	65.80(0.73)	72.58(1.47)	59.03(0.02)	88.85(6.06)	40.76(2.90)	16.94(2.77)	**63.17(1.80)**	42.16(4.14)	96.44(0.34)
SCRL [47]	point	65.30(1.45)	72.63(2.73)	57.98(5.63)	91.31(4.50)	41.51(6.46)	23.08(6.81)	61.55(4.61)	39.91(7.97)	97.50(0.77)
Chaitanya et al. [59]	point	67.52(3.09)	68.10(2.67)	66.95(3.51)	96.89(2.01)	40.07(2.61)	22.25(2.89)	58.02(2.83)	39.95(2.11)	94.68(1.91)
ISCLNet	point	**71.29(0.94)**	**73.16(0.05)**	**69.42(1.88)**	**99.51(0.02)**	**44.38(2.89)**	**25.70(2.69)**	57.98(3.19)	**49.46(2.79)**	**97.65(1.33)**
Upper Baseline	full	75.68(2.39)	78.76(2.39)	72.60(2.41)	96.64 (3.31)	53.56(2.78)	43.72(3.92)	71.81(2.66)	45.13(1.77)	97.88(0.74)

receive two adjacent slices $\mathbf{x}_i^j$ and $\mathbf{x}_{i+t}^j$ from the same patient j and predict the segmented maps, respectively, where t is the distance of two adjacent slices. We enforce the similarity of the central features between the Encoder and the Decoder and the final respond map. Thus, we have two loss functions to enforce the similarity:

$$l_{inter} = \sum_{i=1}^{N} \lambda(l_f + l_s), \lambda = \frac{N-t}{t \times (N-1)}, \tag{3}$$

$$l_f = (E(\mathbf{x}_i^j) - E(\mathbf{x}_{i+t}^j))^2, \tag{4}$$

where E denotes the encoder of backbone model. N denotes the number of slices of a OCT volume. The l_f enforces the continuity between inter-slice encoded feature maps.

$$l_s = ||\mathbf{o}_i^j - \mathbf{o}_{i+t}^j||_1. \tag{5}$$

The l_s constrains the inter-slice predictive probability maps to be similar. Here λ controls the weight of the similarity. When the spatial distance of two slices is far, t is large. λ is small, and l_{inter} does not constrain the far away slices. When two slices are nearby, t is small and λ large, encouraging the slices to be similar.

TABLE II

WILCOXON RANK-SIGN TEST OF COMPARED METHODS

Comparison	Median Difference of DSC	p-value
Lower Baseline vs. ISCLNet	11.70%	0.006
CRF loss vs. ISCLNet	5.36%	0.006
SCRL vs. ISCLNet	5.97%	0.006
Chaitanya et al. vs. ISCLNet	3.97%	0.006

E. Loss Functions

The whole framework of ISCLNet is shown as Fig. 3. We optimize the learnable parameter θ and impose three requirements on the network output: (1) It should respect labeled pixels in point labels $\mathbf{p}$; (2) The segmented fluid patch must be similar with the background patch and dissimilar with the retinal layer patch (i.e., the intra-slice contrastive learning); (3) The inter-slice feature and output should be similar (i.e., the inter-slice contrastive learning). Considering these requirements, we formulate the task as the following loss function:

$$l = l_{pce}(\mathbf{o}, \mathbf{p}) + l_{intra} + l_{inter}, \tag{6}$$

where l_{pce} denotes the cross-entropy between the predictive probability map $\mathbf{o}$ and the point label $\mathbf{p}$. Thus, an end-to-

TABLE III

ABLATION EXPERIMENTAL RESULTS: MEAN (STANDARD DEVIATION) OF 10-FOLD CROSS-VALIDATION. THE BEST RESULTS IN THIS TABLE ARE LABELED IN BOLD. (UNIT: %)

Dataset		AI Challenger				RETOUCH				
Method		DSC	IRF	PED	AUC	DSC	IRF	SRF	PED	AUC
Intra-Slice										
Sim	Dissim									
✓	✗	66.40(1.76)	74.07(0.31)	58.73(3.21)	91.49(4.67)	40.78(2.00)	24.21(1.74)	58.78(0.58)	39.36(3.68)	95.81(0.81)
✗	✓	67.02(1.78)	70.59(3.25)	63.46(0.32)	97.63(4.83)	41.08(1.25)	20.80(1.61)	57.91(0.77)	44.54(1.37)	94.72(1.12)
✓	✓	68.29(2.27)	69.68(3.33)	66.90(5.58)	95.93(3.32)	42.85(2.00)	23.07(0.80)	59.00(0.16)	46.48(5.03)	96.30(0.34)
Inter-Slice										
Fea	Seg									
✓	✗	64.31(3.48)	69.62(3.08)	59.00(4.70)	98.51(1.01)	41.27(2.75)	22.26(1.20)	**60.53(2.58)**	41.02(4.47)	96.23(0.35)
✗	✓	66.64(2.55)	**75.60(3.26)**	57.68(8.37)	97.35(0.75)	41.50(6.46)	23.78(6.81)	51.91(4.61)	48.79(7.97)	96.78(0.77)
✓	✓	67.82(1.47)	71.52(1.82)	64.12(0.07)	99.38(1.39)	42.57(4.00)	21.62(1.03)	56.11(3.85)	**49.97(7.12)**	97.20(0.21)
ISCLNet		**71.29(0.94)**	73.16(0.05)	**69.42(1.88)**	**99.51(6.62)**	**44.38(2.89)**	**25.70(2.69)**	57.98(3.19)	49.46(2.79)	**97.65(1.33)**

TABLE IV

SEGMENTATION RESULTS OF ALTERING BACKBONES: MEAN (STANDARD DEVIATION) OF 10-FOLD CROSS-VALIDATION. THE BEST RESULTS IN THIS TABLE ARE LABELED IN BOLD. (UNIT: %)

Dataset	AI Challenger				RETOUCH				
Method	DSC	IRF	PED	AUC	DSC	IRF	SRF	PED	AUC
Plain UNet	59.72(2.40)	70.05(1.62)	49.38(3.17)	88.05(4.58)	39.41(2.14)	20.34(1.21)	57.40(4.09)	40.48(1.13)	95.06(0.53)
ISCLNet-UNet	71.29(0.97)	73.16(0.05)	**69.42(1.88)**	**99.51(6.62)**	**44.38(2.89)**	**25.70(2.69)**	57.98(3.19)	**49.46(2.79)**	**97.65(1.33)**
Plain UNet++ [71]	65.80(2.23)	72.58(2.27)	59.03(2.19)	88.85(1.32)	40.76(2.96)	16.94(3.30)	**63.17(2.51)**	42.16(3.06)	96.30(0.27)
ISCLNet-UNet++	**71.81(1.68)**	**81.68(3.01)**	61.94(0.35)	98.53(4.43)	42.85(2.00)	23.07(0.80)	59.00(0.16)	46.48(5.03)	96.44(0.34)
DeepLab [13]	58.90(3.93)	70.07(3.69)	47.73(4.16)	96.22(2.67)	33.94(2.53)	19.65(2.42)	42.63(2.57)	39.55(2.61)	92.35(2.27)
ISCLNet-DeepLab	68.13(2.54)	76.89(2.22)	59.38(2.86)	98.30(1.06)	36.94(2.62)	21.86(1.49)	45.27(2.57)	43.59(3.80)	93.62(1.47)

end OCT fluid segmentation model is built, which learns to segment OCT fluid directly from the point supervision.

IV. EXPERIMENTAL RESULTS

A. Dataset

The AI Challenger[1] dataset and the RETOUCH dataset [68] are used in the experiments. The AI Challenger dataset is well annotated at the pixel level for SRF and PED fluids. The dataset comprises training, validation, and test parts containing 70, 15, and 15 cases, respectively. Each case contains 128 OCT slices, with a resolution of 512 × 1024. It should be noted that only the training and validation datasets' annotations have been released so that the following experiments are evaluated on the validation dataset.

The RETOUCH dataset includes three training sets from different OCT devices, including a total of 70 volumes, with 24 volumes acquired with Cirrus (Zeiss), 24 volumes acquired with Spectralis (Heidelberg), and 22 volumes acquired with T-1000 and T-2000 (Topcon). For each volume from these three devices, the numbers of B-scans were 128, 49, and 128 with resolutions of 512×1024, 512×496, and 512×885, respectively. Three different fluid types, i.e., the IRF, SRF, and PED, are manually labeled and provided as ground truth. Then, three testing datasets with 14 volumes acquired with three devices were released to validate the proposed method. The RETOUCH organizers evaluated the results upon submission by the research teams, and hence, the ground truth of the RETOUCH test data remains unknown to the public. Therefore, we implement 10-fold cross-validation on the RETOUCH training dataset split on the case level in our experiments.

TABLE V

SEGMENTATION RESULTS ON AI CHALLENGE DATASET OF DIFFERENT SUPERVISION STRATEGIES: MEAN. (UNIT: %)

Method	supervision	DSC
CAM [63]	image	44.08
UNet	point	59.72
ISCLNet	point	71.29

To generate point labels for training. We randomly select a pixel from the ground-truth masks. Then the selected pixel combined with its 4-connected pixels are treated as point labels. Each fluid category has a point label in our experiments

B. Experimental Setting

Since the various resolutions of raw slices from different OCT devices, the input size is resized to 512 × 496. We perform standardization and normalization to input OCT images. Also, we conduct data augmentation strategies on each OCT image by random horizontal flipping. We optimized the network with Adam optimizer [69] on randomly drawn OCT samples from the dataset. The initial learning rate is set to 10^{-3}, and the weight decay factor is set to 10^{-4}. We train the network for 60 epochs, and the batch size is set to 8 due to hardware limitations. The whole framework is built on PyTorch 1.7.0 [70]. All experiments are performed under an Ubuntu 20.04.1 LTS operating system with CPU Intel Core i7-8700K 3.70 GHz, GPU NVIDIA GeForce GTX 1080 Ti, and RAM of 32 GB. The network has 5.32 MB trainable parameters.

[1] https://www.challenger.ai

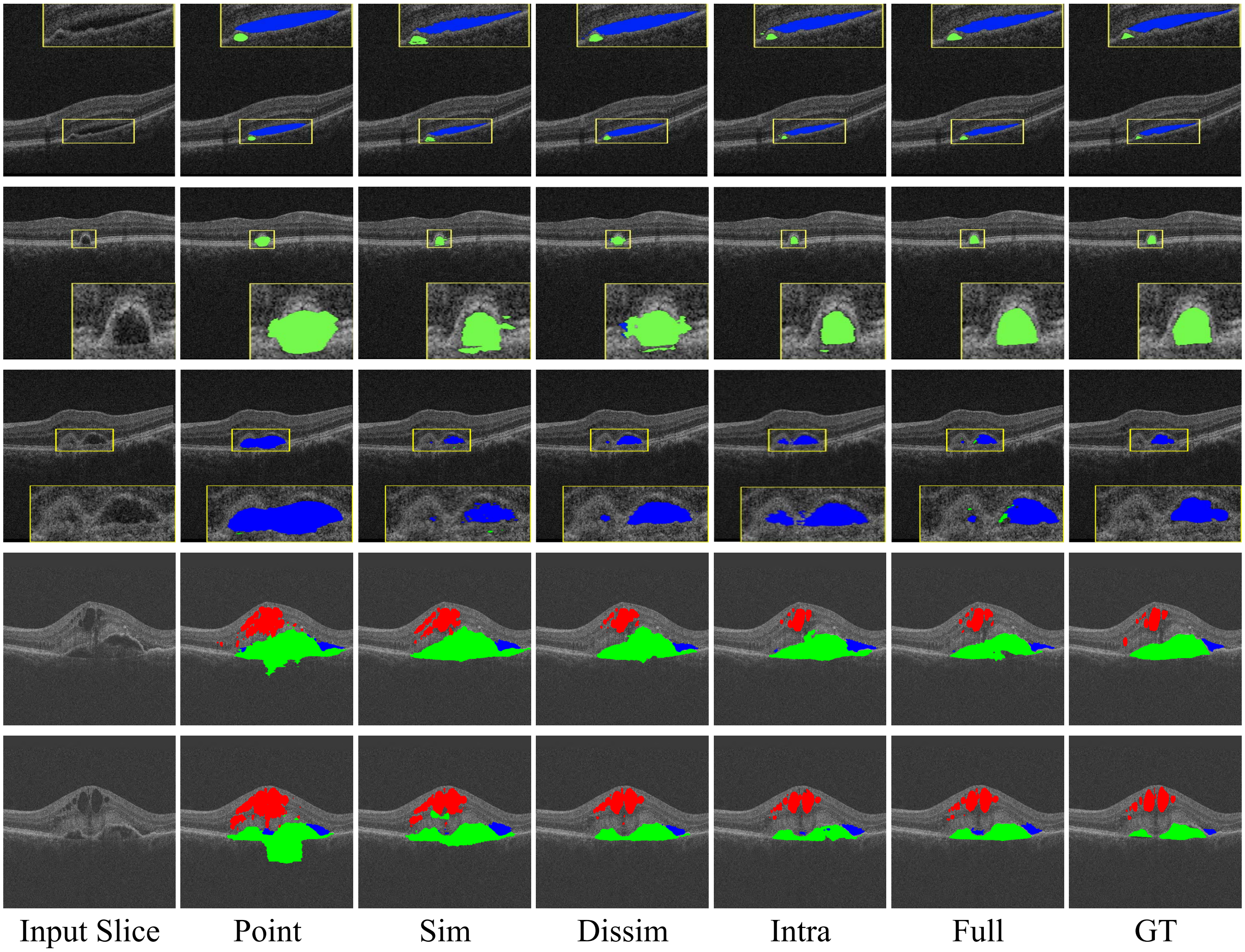


Fig. 4. Visualized segmentation samples of with and without intra-slice learning. The five lines exhibit the visualized segmentation results of five individual OCT cases with different segmentation methods. Point, the lower baseline method that utilize point supervision to train the segmentation model directly; Sim, backbone method combined with fluid-background similarity learning; Dissim, backbone method combined with fluid-retinal layer dissimilarity learning; Intra, backbone method combined with bot similarity and dissimilarity learning; Full, the upper base line method that train the backbone model with full masks; GT, the groundtruth masks. The segmentation results are resized for better visualization. Since fluid regions in the first three lines are too small to be observed, we zoom in on those fluid regions to yellow boxes. Colors indicate different fluid types (i.e., blue refers to SRF, red refers to IRF and green refers to PED, respectively).

TABLE VI

Segmentation Results on AI Challenge Dataset of Two Point Selection Strategies: Mean. (Unit: %)

Method	supervision	DSC
ISCLNet	Central Point	42.62
ISCLNet	Random Point	71.29

The network's training time was 6 hours, and the inference time was 5.37 seconds per OCT volume with 128 slices. During inference, we did not use any post-processing operations and model ensemble techniques. The Dice Similarity Coefficient (DSC) score is used to evaluate the segmentation. The source code of our proposed ISCLNet will be released at https://github.com/lphxx6222712/ISCLNet.

C. Comparison of the State-of-The-Art Methods

We compare our ISCLNet with the state-of-the-art point-supervised segmentation methods. We choose the UNet as the segmentation backbone. The UNet trained with point annotations is set as the lower baseline method and the UNet trained with full masks is set as the upper baseline method. We compare our ISCLNet with state-of-the-art point-supervised segmentation methods [43], [47]. All compared methods are trained with the same backbone and training setting. Table I exhibits the segmentation results on average DSC scores of all fluid types and DSC scores for each fluid type (IRF, SRF, and PED). Also, the average classification AUC of all fluid types is exhibited. The experimental results demonstrate that all compared methods improve the point-supervised segmentation baseline. Our ISCLNet achieves the best results (71.29% on average DSC score on the AI Challenger dataset and 44.38% on the RETOUCH dataset, respectively).

We further performed the Wilcoxon rank-sign test of our method between each compared method to demonstrate whether the improvement of our method is statically significant. The test results are reported in Table II. Here, we set the

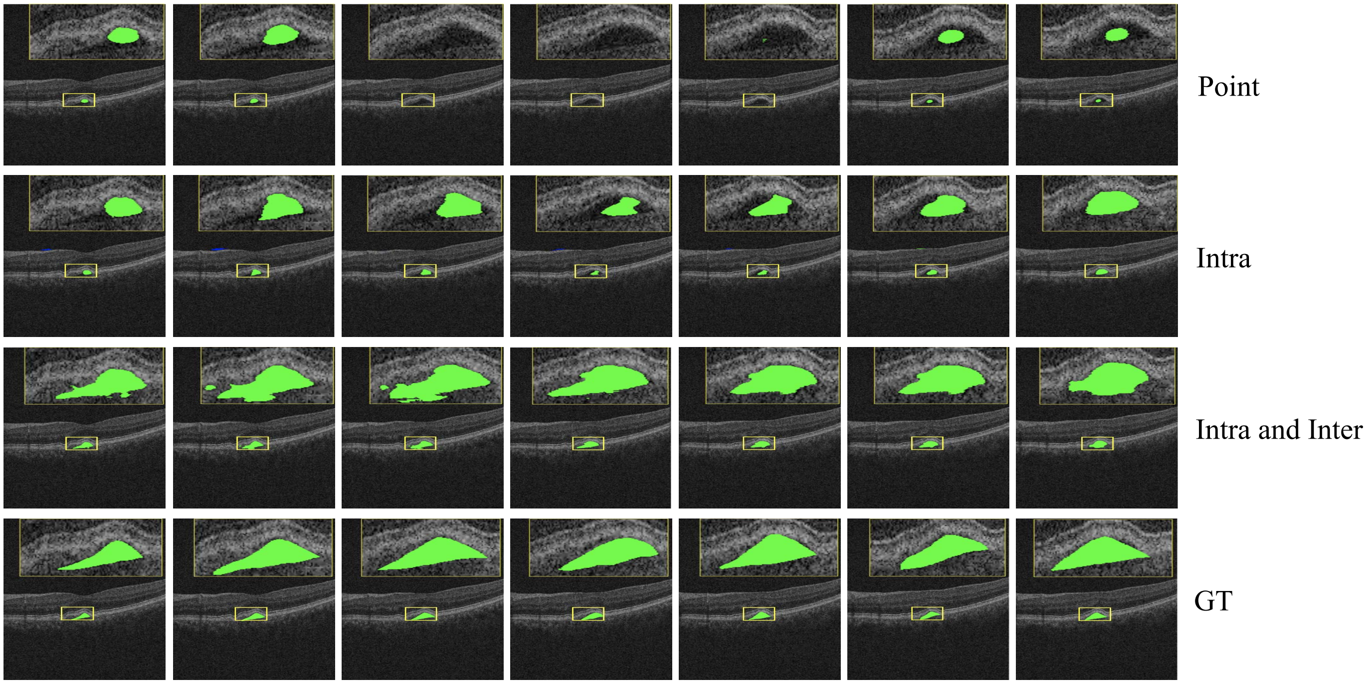


Fig. 5. Visualized segmentation results of adjacent OCT slices from the same case. Each line exhibits the different methods' visualized segmentation results. Point, the lower baseline method that train segmentation model with point label; Intra the combination of baseline method and intra-slice contrastive learning; Intra and Inter, the combination of intra-slice and inter-slice contrastive learning; GT, the groundtruth masks.

significance level $\alpha = 0.05$. All p-values from comparisons allow us to determine the significance of the improvement of our method. Furthermore, since we performed significant tests four times, the Holm-Bonferroni correction is needed to further determine the significance of multiple comparisons to avoid the type I error. After correction, we have the adjusted significance level $\alpha^* = \alpha \times (1/N) = 0.05 \times (1/4) = 0.0125$. The p-values of our experimental results are still smaller than α^*. Therefore, we can accept the hypothesis that the improvement of our proposed method is statically significant (i.e., we reject the hypothesis H_0: median difference of DSC $= 0$).

D. Ablation Experiments

To explore the contribution of each part of the ISCLNet, we investigate the impact of intra- and inter-slice contrastive learning. To this end, we compare the OCT image fluid segmentation results with and without each part of ISCLNet.

1) Effect of Intra-Slice Contrastive Learning: As shown in Fig. 4, the point-supervised model hardly segments the fluid with appropriate extension, shape, and boundary. With the learning of fluid-background similarity, the segmented fluids have better extension and shape but cannot fit the boundary well. Besides, we can obtain results with the fluid-retinal layer dissimilarity to fit the boundary, but the extension and shape are not aware of the fluid region. Then we combine similarity and dissimilarity learning as intra-slice contrastive learning. We can achieve nearly the same results as fully-supervised segmentation with appropriate extension while fitting the boundary well. As a result, intra-slice learning improves the average DSC score from 59.72% to 68.29% (Table III). Also, intra-slice learning improves the average DSC score from 39.41% to 42.85% on the RETOUCH dataset.

2) Effect of Inter-Slice Contrastive Learning: Although the intra-slice information has been learned, the inter-slice continuity can not be ensured (Fig. 5). We try to learn the inter-slice similarity. As a result, inter-slice contrastive learning enhances inter-slice continuity, thus improving fluid segmentation. Combining intra- and inter-slice learning, i.e., the ISCLNet achieves higher performance on average DSC score and DSC score of SRF and PED from 59.72%, 70.05% and 49.38% to 71.29%, 73.16% and 69.42%, respectively on the AI Challenger dataset. The segmentation performance on average DSC score and DSC score of IRF, SRF, and PED from 39.41%, 20.34%, 57.40%, and 40.48% to 25.70%, 57.98%, 49.46%, respectively, on the RETOUCH dataset. (Table III). It can be observed from Fig. 5 that our inter-slice contrastive learning further improves the point-supervised method compared with the intra-slice contrastive learning alone. However, when combining intra- and inter-slice contrastive learning, the visualization results show some anti-regularizing effects (e.g., line 3, row 3 of Fig. 5), i.e., the segmentation model is unaware of the fluid boundary. These results are expected. The point label only provides supervision information of a tiny region in the fluid. Therefore, a plain point-supervised segmentation model only focuses on the central region of fluid and has no chance to fit the boundary. Compared with plain point-supervised learning, our contrastive learning method encourages the model to learn comprehensive and effective representations of fluid and non-fluid regions. Those representations improve the point-supervised model to achieve a complete segmentation and try to fit the boundary. However, boundaries in OCT images are difficult to recognize, even for an ophthalmologist. As a result, with the point label only, our proposed method may not achieve satisfactory results around the boundary region.

E. Compared With Other Supervision Strategies

The image-level label is the easiest annotation to obtain. Many works have demonstrated the possibility of training segmentation methods with the supervision of image labels. Proposal-based methods have been widely studied in image tags supervised segmentation. Most of them adopt a feature maps activation technique (i.e., class activation map, CAM) [63]. CAM allows inspecting class-related pixels of images that have contributed more to the final output of the model. Usually, those regions with high contributions can be highlighted and used to generate pseudo segmentation maps to train weakly-supervised segmentation methods. In our setting, the burden of obtaining a single point for each fluid class is almost the same as the image label. At the same time, the point label can improve the segmentation baseline dramatically (from 44.08% to 59.29%) (Table V) on the AI Challenger dataset. Therefore we chose the point label as the weak supervision for our method.

F. Compared With Different Point Selection Strategies

In the above experiments, we train segmentation models with point labels generated by randomly selecting pixels from the ground-truth masks. It is interesting to explore other point selection strategies. In the clinic, an intuitive point selection strategy is to mark the central region of fluids since central fluid regions usually have a clean texture and can be easily recognized. Hence, to imitate the annotation in a clinical context, we experimented with point labels located in the center of fluid regions. Note that the size of central point labels is the same as random point labels. The experimental result (Table VI) demonstrates that central point labels have a huge performance drop compared with random point labels. One reason is that all fluid central areas have highly similar contexts, and the center point label leads the model to be sensitive to the central region of fluid and ignore other non-central and complicated fluid regions. As a result, the central point label is worse for learning a segmentation model with the same annotation cost.

G. Effect of Annotation Density

To understand how our method helps the deep learning method reduce dependencies on manual annotations. We further perform our proposed ISCLNet with different annotation densities. The graph (Fig. 6) demonstrates that a plain OCT fluid segmentation model can be improved with the increase of label density. Meanwhile, our proposed method boosts the plain OCT fluid segmentation model in all settings. However, with higher label densities, our proposed method yields more marginal improvements. One reason is that our proposed contrastive learning method tries to supplement the missing information of point labels (e.g., the extension, shape, and boundary of fluids). Therefore, our method may not be effective for cases where manual annotations are fine enough to provide that information. In addition, the previous section illustrates the importance of annotating non-central regions, which may also explain why higher annotation densities, including more of these regions, produce better results.

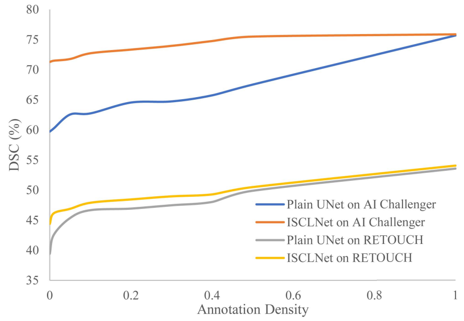


Fig. 6. Segmentation performance with different annotation densities of our proposed method.

H. Combine With Different Backbones

Our ISCLNet can combine with any CNN-based segmentation backbones. Here, we employ the UNet++ [71], an improved version of UNet, and DeepLab [13], which is widely used in natural image semantic segmentation as alternative backbones. The experimental results (Table IV) show that the segmentation performance of UNet++ does not outperform the UNet on the RETOUCH dataset. One reason is that the RETOUCH dataset is collected from three different devices. The regular UNet or UNet++ is hard to handle the samples from different data sources. In addition, the results of the DeepLab model are worse than a UNet or UNet++ on datasets used in our setting. One reason is that the two OCT fluid segmentation datasets are quite small compared to other natural image segmentation datasets. Therefore, utilizing a model with a huge number of parameters (e.g., ResNet121) as the encoder on a small dataset may easily fall into overfitting, resulting in deteriorating performance. Therefore, to apply our method to a practical dataset, one can choose an appropriate backbone network to adapt to other datasets.

V. Discussion

Recently, the CNN-based segmentation method has achieved great success in several applications. However, the training of CNN-based segmentation models heavily relies on pixel-wise manual segmentation masks. The obtaining of segmentation masks is time-consuming and expensive. This work concentrates on reducing the human labor in training a CNN-based segmentation model for OCT fluid segmentation. Instead of pixel-wise masks, our method only needs a single point of each fluid category as the supervision. The point label only provides limited information about the object, which is insufficient for training a pixel-wise segmentation model. We consider learning inherent features from the OCT image itself through contrastive learning. Contrastive learning discovers the feature pattern by exploiting the similarity and dissimilarity of OCT samples. Specifically, we propose an intra-slice contrastive strategy to learn the fluid-background similarity and retinal layer-background dissimilarity within an OCT slice. Moreover, we propose an inter-slice contrastive architecture to learn the continuity of adjacent OCT slices. Finally, an end-to-end model named Intra- and inter-Slice Contrastive Learning Network (ISCLNet) is built to predict retinal fluid.

Although our ISCLNet achieved good segmentation results validated on two public OCT fluid segmentation datasets, limitations still exist. The intra-slice contrastive learning is based on the observation that the fluid region has the same reflection characters as the background since both are filled with liquid. The inter-slice contrastive learning is based on the similarity of adjacent OCT slices. However, in the clinic, the outpatient ophthalmologist would not save all OCT slices. They only focus on one to three slices. The inter-slice contrastive method may be hard to apply to incomplete OCT data. Given this limitation, one of our future directions is to learn from incomplete 3D OCT data where only a few slices are available.

VI. Conclusion

This paper proposed an Intra- and inter-Slice Contrastive Learning Network (ISCLNet) that learned the inherent similarity and dissimilarity of OCT images to improve the point-supervised OCT fluid segmentation. The ISCLNet learned the intra-slice fluid-background similarity and the fluid-retinal layers dissimilarity within an OCT slice. Furthermore, an inter-slice contrastive learning architecture was built to learn the similarity among adjacent OCT slices. With the intra- and inter-slice contrastive learning, we trained a CNN-based OCT fluid segmentation network with only one point for each fluid type as supervision. Experimental results on two publicly available OCT fluid segmentation datasets demonstrated that our ISCLNet outperformed other well-known point-supervised segmentation methods. Our segmentation method helped ophthalmologists obtain a rapid, accurate prediction of fluid regions that reduced the burden of manual quantitative analysis of several fluid-related retinal diseases. However, our method relied on complete OCT volumes that might be difficult to access in the clinic. In the proposed work, we considered learning from incomplete 3D OCT volume with only a few available OCT slices.

Acknowledgment

The authors gratefully acknowledge the editors and the four anonymous reviewers for their valuable comments and suggestions, which greatly help them to improve the technical quality and presentation of their manuscript.

References

[1] K. A. Goatman, "A reference standard for the measurement of macular oedema," *Brit. J. Ophthalmol.*, vol. 90, no. 9, pp. 1197–1202, Sep. 2006.

[2] G. Coscas, J. Cunha-Vaz, and G. Soubrane, "Macular edema: Definition and basic concepts," in *Macular Edema*, vol. 47. Basel, Switzerland: Karger Publishers, 2010, pp. 1–9.

[3] D. Huang *et al.*, "Optical coherence tomography," *Science*, vol. 254, no. 5035, pp. 1178–1181, 1991.

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

[5] S. Wolf and U. Wolf-Schnurrbusch, "Spectral-domain optical coherence tomography use in macular diseases: A review," *Ophthalmologica*, vol. 224, no. 6, pp. 333–340, 2010.

[6] G. R. Wilkins, O. M. Houghton, and A. L. Oldenburg, "Automated segmentation of intraretinal cystoid fluid in optical coherence tomography," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 4, pp. 1109–1114, Apr. 2012.

[7] T. Wang *et al.*, "Label propagation and higher-order constraint-based segmentation of fluid-associated regions in retinal SD-OCT images," *Inf. Sci.*, vols. 358–359, pp. 92–111, Sep. 2016.

[8] A. Rashno *et al.*, "Fully automated segmentation of fluid/cyst regions in optical coherence tomography images with diabetic macular edema using neutrosophic sets and graph algorithms," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 5, pp. 989–1001, May 2018.

[9] 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. Exp.*, vol. 8, no. 3, pp. 1874–1888, 2017.

[10] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[12] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Oct. 2016.

[13] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[14] A. G. Roy *et al.*, "ReLayNet: Retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks," *Biomed. Opt. Exp.*, vol. 8, no. 8, pp. 3627–3642, 2017.

[15] J. De Fauw *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature Med.*, vol. 24, no. 9, pp. 1342–1350, Sep. 2018.

[16] 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, Jul. 2019.

[17] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1713–1721.

[18] Y. Liu, Y.-H. Wu, P.-S. Wen, Y.-J. Shi, Y. Qiu, and M.-M. Cheng, "Leveraging instance-, image- and dataset-level information for weakly supervised instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Sep. 10, 2020, doi: 10.1109/TPAMI.2020.3023152.

[19] J. Dai, K. He, and J. Sun, "BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1635–1643.

[20] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Simple does it: Weakly supervised instance and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 876–885.

[21] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 549–565.

[22] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3159–3167.

[23] B. Wang *et al.*, "Boundary perception guidance: A scribble-supervised semantic segmentation approach," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3663–3669.

[24] X. Ma, Z. Ji, S. Niu, T. Leng, D. L. Rubin, and Q. Chen, "MS-CAM: Multi-scale class activation maps for weakly-supervised segmentation of geographic atrophy lesions in SD-OCT images," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 12, pp. 3443–3455, Dec. 2020.

[25] W. Drexler *et al.*, "Enhanced visualization of macular pathology with the use of ultrahigh-resolution optical coherence tomography," *Arch. Ophthalmol.*, vol. 121, no. 5, pp. 695–706, 2003.

[26] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, Dec. 2020.

[27] S. Standring, *Gray's Anatomy e-Book: The Anatomical Basis of Clinical Practice*. Amsterdam, The Netherlands: Elsevier, 2015.

[28] 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. Exp.*, vol. 8, no. 7, pp. 3440–3448, Jul. 2017.

[29] T. Schlegl *et al.*, "Fully automated detection and quantification of macular fluid in OCT using deep learning," *Ophthalmology*, vol. 125, no. 4, pp. 549–558, 2018.

[30] A. Rashno, D. D. Koozekanani, and K. K. Parhi, "OCT fluid segmentation using graph shortest path and convolutional neural network," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 3426–3429.

[31] R. Tennakoon, A. K. Gostar, R. Hoseinnezhad, and A. Bab-Hadiashar, "Retinal fluid segmentation in OCT images using adversarial loss based convolutional neural networks," in *Proc. Int. Symp. Biomed. Imag.*, 2018, pp. 1436–1440.

[32] Y. Xu *et al.*, "Dual-stage deep learning framework for pigment epithelium detachment segmentation in polypoidal choroidal vasculopathy," *Biomed. Opt. Exp.*, vol. 8, no. 9, pp. 4061–4076, 2017.

[33] 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, Feb. 2019.

[34] D. Lu *et al.*, "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, May 2019.

[35] D. Mahapatra, B. Bozorgtabar, J.-P. Thiran, and L. Shao, "Pathological retinal region segmentation from OCT images using geometric relation based augmentation," 2020, *arXiv:2003.14119*.

[36] L. B. Sappa *et al.*, "RetFluidNet: Retinal fluid segmentation for SD-OCT images using convolutional neural network," *J. Digit. Imag.*, vol. 34, no. 3, pp. 1–14, 2021.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Jul. 2015.

[38] B. Hassan *et al.*, "Deep learning based joint segmentation and characterization of multi-class retinal fluid lesions on OCT scans for clinical use in anti-VEGF therapy," *Comput. Biol. Med.*, vol. 136, Sep. 2021, Art. no. 104727.

[39] P. Vernaza and M. Chandraker, "Learning random-walk label propagation for weakly-supervised semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7158–7166.

[40] M. Pu, Y. Huang, Q. Guan, and Q. Zou, "GraphNet: Learning image pseudo annotations for weakly-supervised semantic segmentation," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 483–491.

[41] D. Pathak, P. Krahenbuhl, and T. Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1796–1804.

[42] C. Ahlers *et al.*, "Automatic segmentation in three-dimensional analysis of fibrovascular pigmentepithelial detachment using high-definition optical coherence tomography," *Brit. J. Ophthalmol.*, vol. 92, no. 2, pp. 197–203, Feb. 2008.

[43] M. Tang, F. Perazzi, A. Djelouah, I. B. Ayed, C. Schroers, and Y. Boykov, "On regularized losses for weakly-supervised CNN segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 507–522.

[44] D. Marin, M. Tang, I. B. Ayed, and Y. Boykov, "Beyond gradient descent for regularized segmentation losses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10187–10196.

[45] R. Qian, Y. Wei, H. Shi, J. Li, J. Liu, and T. Huang, "Weakly supervised scene parsing with point-based distance metric learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8843–8850.

[46] H. Kervadec, J. Dolz, J. Yuan, C. Desrosiers, E. Granger, and I. B. Ayed, "Constrained deep networks: Lagrangian optimization via log-barrier extensions," 2020, *arXiv:1904.04205*.

[47] O. Veksler, "Regularized loss for weakly supervised single class semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 348–365.

[48] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 69–84.

[49] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101539.

[50] S. Manna, S. Bhattacharya, and U. Pal, "SSLM: Self-supervised learning for medical diagnosis from MR video," 2021, *arXiv:2104.10481*.

[51] F. Navarro *et al.*, "Evaluating the robustness of self-supervised learning in medical imaging," 2021, *arXiv:2105.06986*.

[52] A. Taleb, C. Lippert, T. Klein, and M. Nabi, "Multimodal self-supervised learning for medical image analysis," in *Proc. Int. Conf. Inf. Process. Med. Imag.*, 2019, pp. 661–673.

[53] X. Zhuang, Y. Li, Y. Hu, K. Ma, Y. Yang, and Y. Zheng, "Self-supervised feature learning for 3D medical images by playing a Rubik's cube," in *Medical Image Computing and Computer Assisted Intervention*. Cham, Switzerland: Springer, 2019, pp. 420–428.

[54] J. Zhu, Y. Li, Y. Hu, K. Ma, S. K. Zhou, and Y. Zheng, "Rubik's Cube+: A self-supervised feature learning framework for 3D medical image analysis," *Med. Image Anal.*, vol. 64, Aug. 2020, Art. no. 101746.

[55] X. Tao, Y. Li, W. Zhou, K. Ma, and Y. Zheng, "Revisiting Rubik's cube: Self-supervised learning with volume-wise transformation for 3D medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention*. Cham, Switzerland: Springer, 2020, pp. 238–248.

[56] A. Jamaludin, T. Kadir, and A. Zisserman, "Self-supervised learning for spinal MRIs," in *Proc. Int. Workshop Deep Learn. Med. Image Anal.*, 2017, pp. 294–302.

[57] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," 2020, *arXiv:2010.00747*.

[58] W. Lei, W. Xu, R. Gu, H. Fu, S. Zhang, and G. Wang, "Contrastive learning of relative position regression for one-shot object localization in 3D medical images," 2020, *arXiv:2012.07043*.

[59] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12546–12558.

[60] X.-B. Nguyen, G. S. Lee, S. H. Kim, and H. J. Yang, "Self-supervised learning based on spatial awareness for medical image analysis," *IEEE Access*, vol. 8, pp. 162973–162981, 2020.

[61] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, Nov. 2021.

[62] N. Araslanov and S. Roth, "Single-stage semantic segmentation from image labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4253–4262.

[63] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[64] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised scale equivariant network for weakly supervised semantic segmentation," 2019, *arXiv:1909.03714*.

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

[66] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[67] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 118–126.

[68] H. Bogunović *et al.*, "RETOUCH: The retinal OCT fluid detection and segmentation benchmark and challenge," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1858–1874, Aug. 2019.

[69] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR (Poster)*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

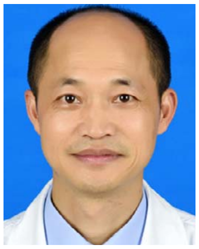
[70] A. Paszke *et al.*, "Automatic differentiation in PyTorch," in *Proc. NIPS Workshop Autodiff Submission*, Long Beach, CA, USA, 2017. [Online]. Available: https://openreview.net/group?id=NIPS.cc/2017/Workshop/Autodiff

[71] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 3–11.

Xingxin He received the M.S. degree in biomedical engineering from Hunan University, Changsha, China, in 2017, where he is currently pursuing the Ph.D. degree with the Laboratory of Vision and Image Processing. His research interests include retinal OCT image analysis and medical image processing.

Leyuan Fang (Senior Member, IEEE) received the Ph.D. degree from the College of Electrical and Information Engineering, Hunan University, Changsha, China, in 2015. From September 2011 to September 2012, he was a Visiting Ph.D. Student with the Department of Ophthalmology, Duke University, Durham, NC, USA, supported by the China Scholarship Council. From August 2016 to September 2017, he was a Post-Doctoral Researcher with the Department of Biomedical Engineering, Duke University. He is currently a Professor with the College of Electrical and Information Engineering, Hunan University, and an Adjunct Researcher with the Peng Cheng Laboratory, Shenzhen, China. His research interests include sparse representation and multi resolution analysis in remote sensing and medical image processing. He was a recipient of one 2nd Grade National Award at the Nature and Science Progress of China in 2019. He is an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and *Neurocomputing*.

Mingkui Tan (Member, IEEE) received the bachelor's degree in environmental science and engineering and the master's degree in control science and engineering from Hunan University, Changsha, China, in 2006 and 2009, respectively, and the Ph.D. degree in computer science from Nanyang Technological University, Singapore, in 2014. From 2014 to 2016, he worked as a Senior Research Associate in computer vision with the School of Computer Science, The University of Adelaide, Australia. He is currently a Professor with the School of Software Engineering, South China University of Technology. His research interests include machine learning, sparse analysis, deep learning, and large-scale optimization.



Xiangdong Chen is currently a Professor with the Department of Ophthalmology, First Affiliated Hospital of Hunan University of Chinese Medicine, Changsha, China. His main research interests include ophthalmology and Chinese medicine.